Cooling the AI Surge

Strategic Infrastructure for Hyperscaler and Colocation Success

AIRSYS
Balance the Environment

AIRSYS is bridging the gap between today's infrastructure and tomorrow's AI-fueled thermal challenges with systems built for resilience, scalability, and performance.

Contents

Introduction

A Wake-Up Call from the Brink of an Outage

Where were you on July 10, 2024?

On that day, Virginia was saved from a major blackout when 60 data centers in the "Data Center Alley" consuming 1,500MW of power dropped off the grid simultaneously, forcing network operators to switch to backup generators. If the blackout had happened, it wouldn't have been only Virginia's problem. 70% of the world's internet traffic flows through that area.

Luckily, the crisis was averted, but it exposed critical vulnerabilities in data center resilience, which becomes increasingly problematic as we step into the AI era.

AI is advancing at a staggering pace, reshaping industries and infrastructure alike, and it's just the beginning. Hyperscalers and colocation facilities that power this revolution face unimaginable pressure to deliver, knowing that data load will exponentially and consistently grow. This growing strain on infrastructure prompts essential conversations about how data centers should respond to escalating pressures tied to AI expansion.

For example, in response to Virginia's almost-blackout incident, some grid operators have proposed forcing data centers to "ride through" voltage dips without disconnecting, effectively asking facilities to risk equipment and cooling systems in the name of grid stability.

This approach treats cooling infrastructure as a passive component expected to endure whatever instability the grid delivers, absorbing voltage fluctuations, frequency variations, and inconsistent power delivery. But in reality, cooling has the potential to be an adaptive, resilient system that not only withstands the demands of high-density data processing, but actively enables it.

This brief explores how cooling infrastructure design can evolve to meet the realities of AI-driven demand. Armed with this knowledge, hyperscalers and colocation operators can upgrade their cooling systems to ensure reliability, sustainability, and performance, no matter what comes next.

Overworked & Overheating

Critical Gaps in Current Cooling Approaches

Trying to cool tomorrow's AI data centers with today's infrastructure is like running a marathon in shoes built for walking; they're simply not engineered for the pace, pressure, or performance required. Likewise, cooling systems that once met traditional needs quickly become unfit for AI-data centers' sophisticated cooling demands.

THERMAL DENSITY SPIKE

The rapid acceleration of AI workloads is pushing data center thermal densities beyond the limits of legacy (and even some modern) cooling systems. Rack densities of 50–100kW are becoming the norm, with individual chips like NVIDIA's Blackwell GPUs drawing up to 1,200 watts.

At the same time, computing power is consolidated into hyperscale environments at an unprecedented speed. Hyperscalers now account for 41% of global data center capacity, a figure expected to exceed 60% by 2029, and many regions are already feeling the strain. In Ireland, for instance, data centers consumed nearly 20% of the national electricity supply in 2024, prompting a halt on new data center builds in Dublin. This trend highlights the compounding impact of thermal density at scale: not only are individual sites running hotter, but their collective energy and cooling demands are outpacing what existing infrastructure can support.

COOLING RESOURCE CONSTRAINTS

AI-driven workloads amplify pressure on two of the most limited resources in data center cooling: water and energy.

Liquid-based cooling systems, while promising higher thermal efficiency, could drive water consumption to as much as 1.7 trillion gallons annually by 2027. Individual hyperscale facilities may require over 50 million gallons per year, much of which is drawn from municipal supplies and lost permanently to evaporation, raising serious long-term sustainability and community impact concerns.

On the other hand, air-based cooling systems can consume up to 40% of a facility's total electricity, creating unsustainable energy burdens as rack densities climb. These patterns present a tricky trade-off between PUE and WUE, as improving one metric often comes at the expense of the other. With both resources under pressure, data center teams are forced to re-evaluate cooling strategies, not just at the facility level, but also within broader environmental, regulatory, and regional infrastructure constraints.

POWER INFRASTRUCTURE RELIABILITY

AI cooling's high power demands collide with power infrastructure that was never designed for this level of throughput. To start, existing grids were designed for predictable usage cycles, with demand surging during work hours, then tapering off. But AI data centers break that rhythm, running at maximum capacity 24/7. Also, according to the Department of Energy, the average age of large power transformers, which handle 90% of the electricity flow in the U.S., is more than 40 years.

The incompatibility between AI infrastructure requirements and grid limitations challenges data center cooling systems. For example, fluctuations from AI training loads can cause cooling equipment to operate outside design parameters, reducing efficiency and accelerating wear.

When standard chillers and cooling pumps experience voltage fluctuations or frequency variations, their thermal transfer capabilities degrade, ultimately jeopardizing the entire facility's operation.

DEPLOYMENT SPEED

Accelerated AI adoption has compressed data center build timelines, creating tension between rapid delivery and long-term reliability. Market expectations now call for hundreds of megawatts to be delivered in 12–18 months, while traditional build-to-suit projects average over two years from conception to commission. Failing to meet these timelines is costly, with lost revenue estimated at $10,000 to $20,000 per minute. This makes speed-to-market a critical priority, but one that also strains every aspect of cooling system deployment, from procurement lead times to commissioning processes that can't be safely rushed.

The scarcity of off-the-shelf capacity in major markets forces operators into high-risk deployment strategies where cooling systems must be designed, procured, and installed in parallel, rather than sequential phases. Recent case studies demonstrate that while 40+ megawatts of capacity can be delivered in 16 months with aggressive project management, this acceleration typically requires accepting higher operational risks and potentially suboptimal cooling efficiency.

Given all of these widening gaps, it's unrealistic to expect traditional cooling systems to step into the AI era confidently.

Meeting tomorrow's demand will require fundamentally rethinking how AI data centers should approach thermal management.

Smarter Cooling for Smarter Data Centers

Principles of AI-Ready Cooling Infrastructure

AI-driven data centers deliver advanced, intelligent capabilities, and their cooling systems should be just as advanced. This means that cooling solutions for AI-driven data centers must be more powerful and precise than those used in traditional environments. Here's how cooling solutions are purpose-built to meet the escalating demands of AI, hyperscale, and colocation data centers.

MODULARITY

Modular cooling is key to maintaining flexibility, speed, and efficiency in hyperscale, colocation, and AI data centers, where demand grows fast and unpredictably. Modularity can be implemented through decentralized cooling units, scalable system blocks, and variable-speed components that respond to changing thermal loads in real time. Liquid-based cooling technologies such as chilled water systems and immersion cooling are inherently scalable, making them strong examples of how modular cooling can support cooling demands and drive operational and environmental benefits. Examples include:

- Up to 50% energy savings for colocation providers using immersion cooling.
- PUE below 1.1 in modular deployments utilizing chilled water systems.
- 15–21% reduction in greenhouse gas emissions from immersion and cold plate liquid cooling systems.

MULTI-MODE READINESS

Despite the industry push to favor liquid over air cooling, neither technology is disappearing. In fact, hybrid cooling often proves more resilient than relying on a single cooling mode, especially in high-density environments and resource-constrained regions. By integrating liquid cooling alongside air-based systems, operators can more effectively manage thermal hotspots, scale with demand, and gain inherent redundancy if one mode fails.

A study by the American Society of Mechanical Engineers (ASME) reported several notable performance improvements in high-density hybrid-cooled data centers:

- Transitioning to a 25% air / 75% liquid configuration resulted in a significant reduction in PUE.
- Facility power consumption was reduced by 27% with the above configuration.
- Overall site energy usage dropped by 15.5% compared to a fully air-cooled data center.

INTELLIGENT CONTROLS

If high-density data centers power AI capabilities, shouldn't they also harness AI to streamline their own operations? Many of the world's largest colocations have already deployed AI in their environments, and the results are already showing. As early as 2016, Google's AI implementation cut its data center cooling costs by 40%. Gartner analysts predict that early AI adoption will be a defining factor separating next-generation facilities from the others. Interestingly, AI is both the source of today's data center cooling challenges and the key to solving them.

AI-driven control systems support critical cooling requirements for AI-ready data centers by:

- Enabling real-time thermal optimization to accommodate fluctuating workloads.
- Improving energy efficiency by dynamically tuning fan speeds, pumps, and chillers.
- Reducing water usage through predictive allocation and flow control.
- Enhancing resilience via early fault detection and failure prediction.

BUILT-IN REDUNDANCY

AI racks drawing 50–100kW can't operate reliably under extreme thermal density and unstable cooling power infrastructure. According to the Uptime Institute's 2024 Outage Analysis, cooling system failures accounted for 19% of all data center outages, second only to power-related issues. These failures aren't only common; they're expensive: 54% of respondents reported losses over $100,000, and 16% faced damages exceeding $1 million. With stakes this high, cooling redundancy is absolutely essential.

Key strategies to enable redundancy in cooling systems include:

- Using N+1, 2N, or 2N+1 architectures.
- Deploying redundant liquid cooling loops with multiple independent sets of piping, pumps, and heat exchangers.
- Implementing hybrid cooling systems that provide passive redundancy.
- Employing modular cooling infrastructure.

SUSTAINABLE OPERATION

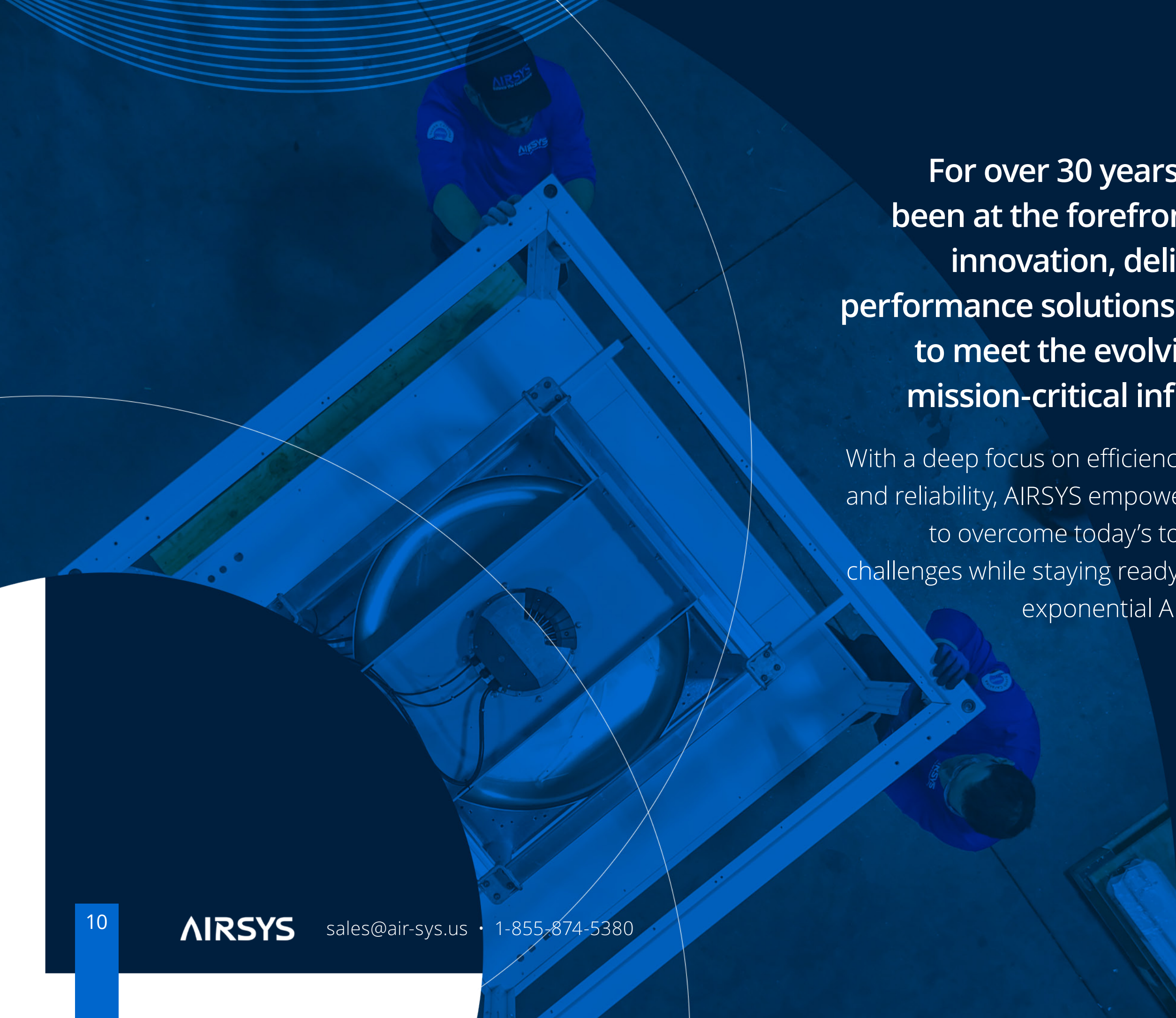
As AI workloads drive up thermal loads, sustainable cooling strategies for high-density data centers offer a dual advantage: supporting environmental goals and reducing operational costs. These strategies focus on two main approaches — optimizing energy use through techniques like liquid cooling, variable-speed components, and intelligent controls; and integrating green features such as heat recovery systems and low-GWP refrigerants to cut emissions and improve efficiency. This balanced use of energy for AI processing also ensures humans are left with enough power for lighting and AC.

Quantifiable benefits of sustainable cooling include:

- 15–20% lower energy demand and 31–52% reduction in water usage with liquid-based cooling approaches.
- 23.45% reduction in electricity expenses and 16.4% overall energy savings through integrated heat recovery systems.
- Up to 70% cooling cost reduction by replacing mechanical chillers with cooling towers for chilled water supply.

AIRSYS

Empowering High-Density Data Centers for the AI Era



For over 30 years AIRSYS has been at the forefront of cooling innovation, delivering high-performance solutions engineered to meet the evolving needs of mission-critical infrastructure.

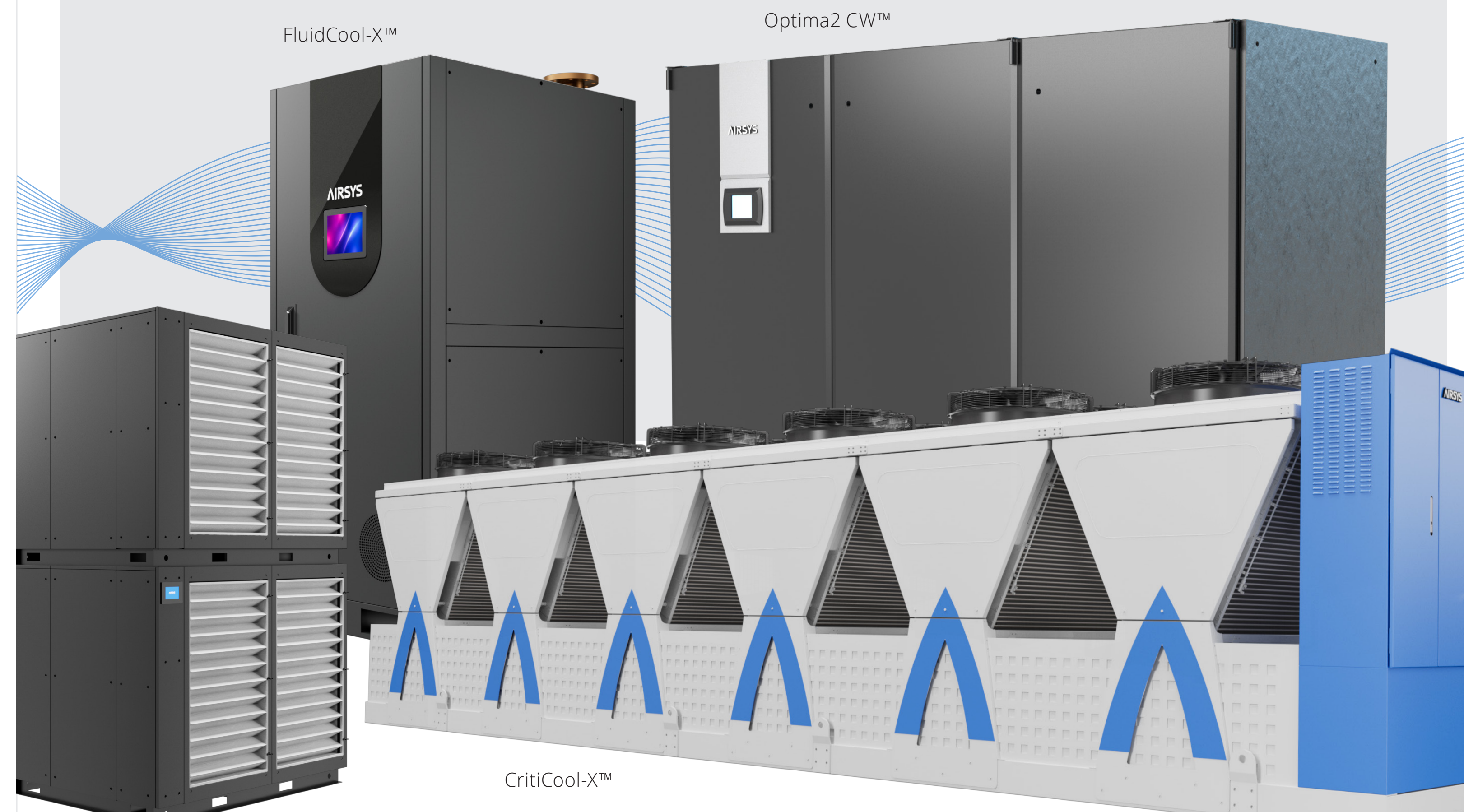
With a deep focus on efficiency, sustainability, and reliability, AIRSYS empowers data centers to overcome today's toughest thermal challenges while staying ready for tomorrow's exponential AI-driven growth.

PowerOne™

Cooling Technology Made to Meet Tomorrow's Demands

PowerOne is AIRSYS' flagship line of intelligent, high-efficiency, and scalable cooling solutions, purpose-built for the intense cooling demands of hyperscale, colocation, and AI-driven facilities. Designed to support ultra-high-density IT loads, each system in the PowerOne portfolio is engineered to close the cooling gaps in modern data center infrastructure, especially those amplified by AI workloads and rapid scaling requirements.

Whether you're expanding hyperscale capacity or retrofitting colocation facilities for high density, PowerOne delivers the resilience, efficiency, and flexibility required to operate reliably in an era of unprecedented compute density.



FluidCool-X™ *for Seamless Scaling*

When uptime is non-negotiable and density is accelerating, FluidCool-X delivers. Engineered for reliability, speed, and scale, from edge AI pods to hyperscale clusters.

- » Available in 500kW/1000kW/2000kW configurations
- » Up to 99.999% uptime with 3-tier redundancy
- » Front-access design for fast maintenance without downtime
- » Ultra-low PUE <1.05
- » VFD pumps and 4K approach temperature technology
- » BPHE and water reuse

Optima2 CW™ *for Precision Cooling*

Built for operators who demand precision, efficiency, and reliability, Optima2 CW is a high-performance CRAH unit designed to meet the thermal challenges of modern data centers. With coil technology, precision flow control, and modular design, it delivers stable, energy-saving cooling that scales with your footprint while keeping maintenance downtime to a minimum.

- » Flexible entering water temperature operation for efficient high-temp designs
- » PUE <1.2
- » Precision flow control for energy savings
- » 100% front-access maintenance
- » Modular cooling capacity between 60kW and 800kW
- » Remote smart control systems

CritiCool-X™ *for Ultra Efficiency*

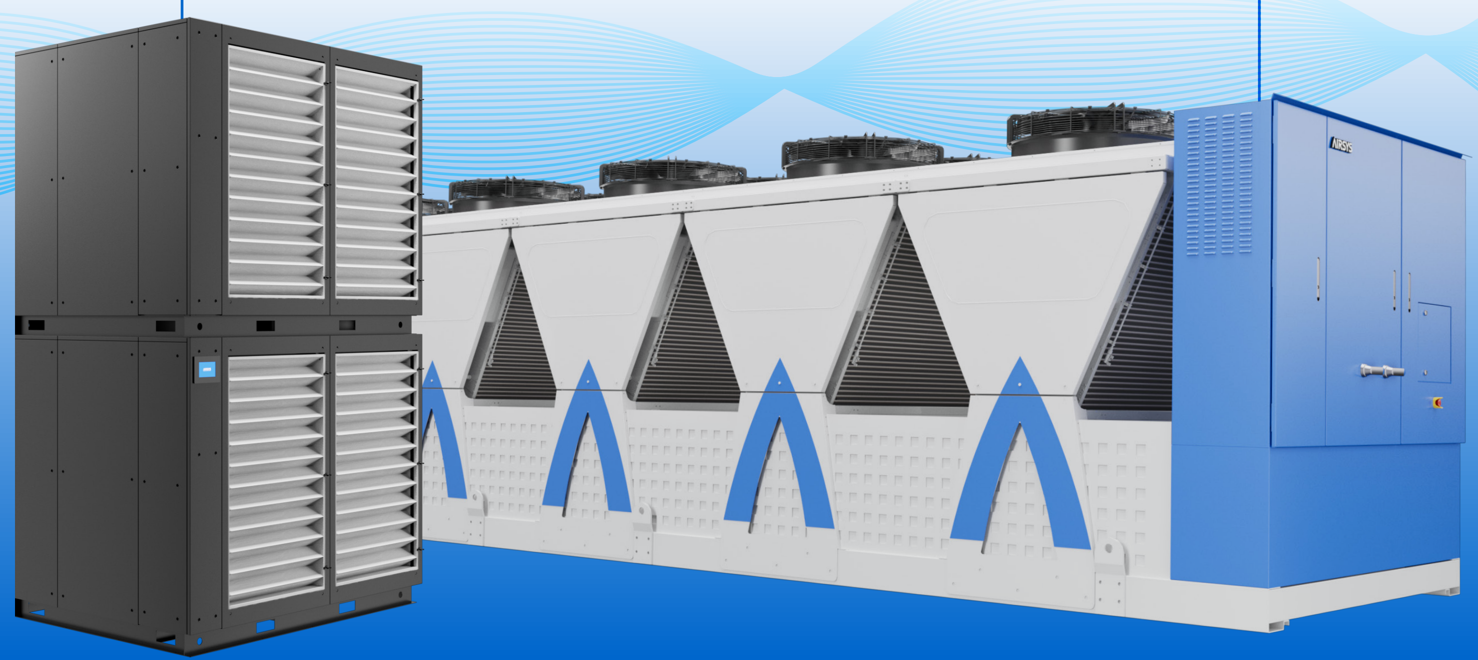
AIRSYS' magnetic bearing centrifugal chiller is built for operators who can't afford inefficiency or downtime. If you're scaling dense AI clusters or facing rising energy bills, CritiCool-X delivers ultra-efficient, oil-free cooling designed to keep up with your performance needs and sustainability goals.

- » Incorporated Turbocor® compressors for 30% less energy consumption
- » Industry-leading PUE <1.15
- » Up to 90% energy savings
- » Scalable cooling capacity up to 2.5MW per unit
- » Stackable or split installation options

MaxAir™ *for Simplified Retrofitting*

Airflow limitations, high energy use, and retrofit constraints slow down high-performance computing data centers, and MaxAir is the solution. This modular CRAH system delivers smart airflow control, high-efficiency EC fans, and a high water temperature design that dramatically improves free cooling and cuts system water demand, all in a footprint that fits virtually anywhere.

- » Flexible scalability, up to 240kW per unit
- » Unified fan wall control minimizes redundant operation
- » Available in quad fan modules
- » Modular coil sizes and fin configurations tailored to application needs
- » Rapid onsite deployment
- » Designed for compact footprints



Conclusions

An Answer to the Call

Certainly, AI data centers should expect exponential growth in thermal demand as AI workloads intensify. But adding cooling capacity alone is an incomplete answer to a complex challenge, and one that will ultimately become insufficient and financially unsustainable.

With that in mind, the future also holds an opportunity for those ready to scale smarter, and not just faster.

To stay competitive and resilient, hyperscaler and colocation operators must adopt a new operational mindset that sees cooling as a dynamic solution rather than a fixed utility. This shift calls for:

- **Awareness:** Recognizing that AI-driven workloads fundamentally reshape thermal demands, and that legacy strategies won't keep pace.
- **Proactive Planning:** Selecting cooling systems with the scalability and adaptability needed to accommodate future density and deployment shifts.
- **Integrated Decision-Making:** Aligning cooling operations with broader infrastructure, sustainability, and capacity planning strategies.
- **Strategic Partnerships:** Collaborating with cooling technology providers who offer long-term vision, flexibility, and ongoing innovation.

In July 2024, Virginia's near-blackout was a warning sign; a glimpse into what happens when infrastructure can't keep up with demand. The data centers that endure and lead in the AI era will be those that treat cooling not as an afterthought, but as a strategic advantage.

The future holds opportunities for those ready to scale smarter.

Adopt an operational mindset that sees cooling as a dynamic solution rather than a fixed utility.

Let's Talk!

Bridge the gap between today's infrastructure and tomorrow's AI-fueled thermal challenges with a system built for resilience, scalability, and performance.

Together, we'll keep your data center ready for what's coming.

Scan to request a FREE consultation or
give us a call at 855-874-5380 to learn more.

AIRSYS
Balance the Environment

7820 Reidville Road, Suite 100
Greer, South Carolina 29651, USA

1-855-874-5380

airsysnorthamerica.com