TechArena®

# Data Center Infrastructure Requirements to Scale AI

Kelley A. Mullick, PhD, Avayla
Mohan Kumar, Perpetual Intelligence
Erica Thomas, Liquid Cooling Coalition
May 2025

**This special report explores the infrastructure innovations required to support AI-scale data centers, highlighting the escalating demands of generative AI on power, cooling, and rack architecture. Learn about the critical role of liquid cooling, optimized rack design, and heat reuse.**

## AUTHORS

**Mohan Kumar - Perpetual Intelligence**
**Hillsboro, OR**
**mohan@perpetual-intelligence.com**

Mohan Kumar is Chief Architect and VP at a stealth-mode AI datacenter company where he drives next-generation AI technology from concept to data center deployment. He is also technical advisor to Perpetual Intelligence Inc., which specializes in AI solutions. Prior to these roles, Mohan was an Intel Fellow in Intel Data Center and AI Group where he led cloud architecture. At Intel, he was the chief architect for many key Intel initiatives such as sustainability, seamless firmware update, rack scale architecture, persistent memory, machine check architecture and trusted execution technology. Mohan has 30+ years of experience in AI, cloud architecture, server architecture, pathfinding of RAS features such as machine check recovery, persistent memory, sustainability, and fleet management solutions and has played key roles in many industry initiatives such as OCP, ACPI, PCI, PCIe, and Redfish. He is also a regular presenter at various industry forums. Mohan holds close to 300 issued patents, and authored a book titled *A Vision for Platform Autonomy.*

**Kelley A. Mullick, PhD - Avayla**
**Hillsboro, OR**
**kelley.mullick@avayla.net**

Kelley Mullick is Founder and CEO of Avayla, a consulting company focused on the design of AI datacenters that incorporate the latest advancements in liquid cooling. Previously, she was VP of Corporate Strategy and Business Development at Iceotope Technologies Ltd, where she drove strategic initiatives and fostered key partnerships within the IT industry. Her work emphasized the importance of liquid cooling for the sustainability of edge and cloud datacenters, demonstrating its ability to deliver superior Total Cost of Ownership (TCO) for end customers. Before joining Iceotope, Dr. Mullick was a leader in Intel's Datacenter and AI group, where she played a pivotal role in shaping the company's liquid cooling strategy and, alongside her team, developed the warranty strategy for immersive cooling solutions. With nearly 20 years of experience, Dr. Mullick's career spans systems engineering, business development, and technical leadership, with a focus on building platforms for sustainable datacenters, optimizing cloud workloads, and creating software-defined infrastructure and innovative business models that generate revenue. A passionate advocate for diversity in technology, Dr. Mullick has championed initiatives to support women and underrepresented minorities in tech, developing programs that have benefited over 3,000 women in advancing their careers. She is actively involved in the Open Compute Project (OCP), and chairs the Industry Liaison team, focused on advancing standards for immersion cooling.

**Erica Thomas - Liquid Cooling Coalition**
**Washington, DC**
**thomas@co2efficient.com**

Erica Thomas is the Executive Director of the Liquid Cooling Coalition (LCC). In this role, she draws on her vast experience in energy, environmental and climate policy, including leadership positions at a premier technology trade association and over two decades as a U.S. diplomat. Before leading the LCC, Erica worked at the Information Technology Industry Council (ITI) where she both led The Green Grid—an industry consortium dedicated to efficient data centers—and ran several industry committees focused on climate and sustainability issues. Prior to her time at ITI, Erica served a career as a U.S. diplomat in the Foreign Service where she served in senior energy and environmental roles in the U.S., China, Belgium, and Taiwan.

AI data center designs differ significantly from traditional data center builds, primarily due to the vastly greater compute and power requirements. While most modern data centers are designed to support free air cooling, the power and cooling demands of AI workloads need specialized approaches.

In this special report, we explore the design considerations for AI clusters compared to traditional workloads, with a focus on the unique infrastructure needed to support AI at scale. We delve into the types of liquid cooling technologies required to meet the thermal demands of AI data centers. And finally, we address the importance of collaboration to drive adoption of these key technologies.

## INTRODUCTION

The rise of generative AI marks the beginning of a new era in computing, representing what may be the most transformative technological advancement of the past 50 years. Throughout 2024, the industry has focused heavily on scaling the computational power and infrastructure needed to support these demanding workloads. However, the architectural and infrastructure challenges associated with these advancements remain a work in progress. Read on to learn how AI workloads interact within the compute stack for optimal performance, the necessary data hall infrastructure modifications to support these innovations, and the critical role of liquid cooling in addressing both current and future thermal demands, thereby fostering continued progress in AI development.

## CHANGES NEEDED IN INFRASTRUCTURE DESIGN

The thermal design power (TDP) of all IT components is rising. Furthermore, the TDP for the latest GPU's is increasing at a logarithmic fashion as compared to other IT components. Figure 1 illustrates this trend when comparing CPU and GPU power over time.

With GPU trends predicted to reach 600kW within the next one-to-two years, it is critical to address the changes needed to retrofit existing data centers as well as how to build out new data centers that are future proofed for increasing thermal load and power requirements, thereby requiring liquid cooling, inclusive of cold plate, immersion, or hybrid, in these designs.

Furthermore, in the next three-to-five years, data center designs will need to consider air, cold plate, immersion, and hybrid technologies co-existing in the same data hall. Most cold plate designs are focused on addressing hot spots in the CPU and GPU while air is left to cool the rest of the IT. While an effective solution today, cold plate options can produce hot spots within the chassis resulting in higher instances of failures. However, as GPU TDP advances

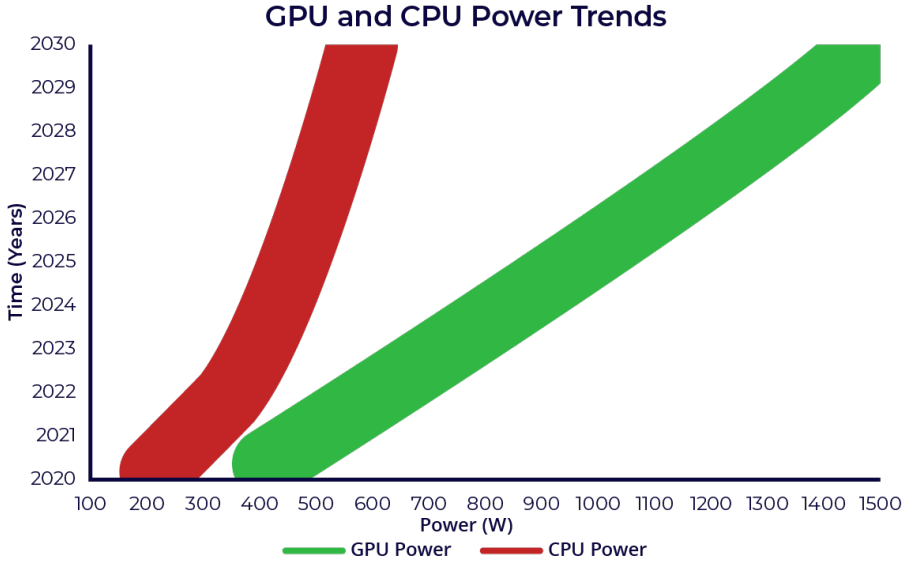


Fig. 1: GPU and CPU Power Trends[1]

beyond the limits of air and cold plate designs alone, single phase immersion, hybrid immersion/cold plate, and two phase technologies will need to be considered for maximum chip cooling requirements. Figure 2 illustrates the inflection points of CPU and GPU TDP that would trigger an industry transition from one cooling technology to another based on projected TDP growth of both CPUs and GPUs. In addition to obstacles with hot spots in cold plate technology, immersion cooling is challenged due to material compatibility, signal integrity, and reliability issues. Two phase technologies must overcome controversial issues with their chemistries. All these challenges will need to be addressed while also considering sustainability and safety standards for AI clusters to scale now and into the future.

## AI CLUSTER OPTIMIZATION AND RACK REDESIGN

Many data centers today use traditional air-cooled architectures, typically housed in 19-inch vertical racks. Over 90% of existing data center infrastructure utilizes free-air cooling, with hot and cold aisles designed to optimize airflow management. IT requirements vary based on workload types. To address these evolving needs, newer rack architectures have been proposed, some of which can accommodate up to 36 CPUs and 72 GPUs per rack.

AI workloads are typically divided into two main categories: training and inference. During the training phase, large volumes of data are fed into the model, enabling it to learn and identify patterns and correlations within the dataset. This iterative process allows the model to continually refine its predictions as it processes more data. Once training is complete, AI inference involves the model making predictions based on new, unseen data. Common applications of AI inference include large language models and predictive analytics. The reason AI workloads are so compute-intensive lies in the complex computations, simulations, and renderings they demand, which drives the need for a high density of CPUs and GPUs per rack.

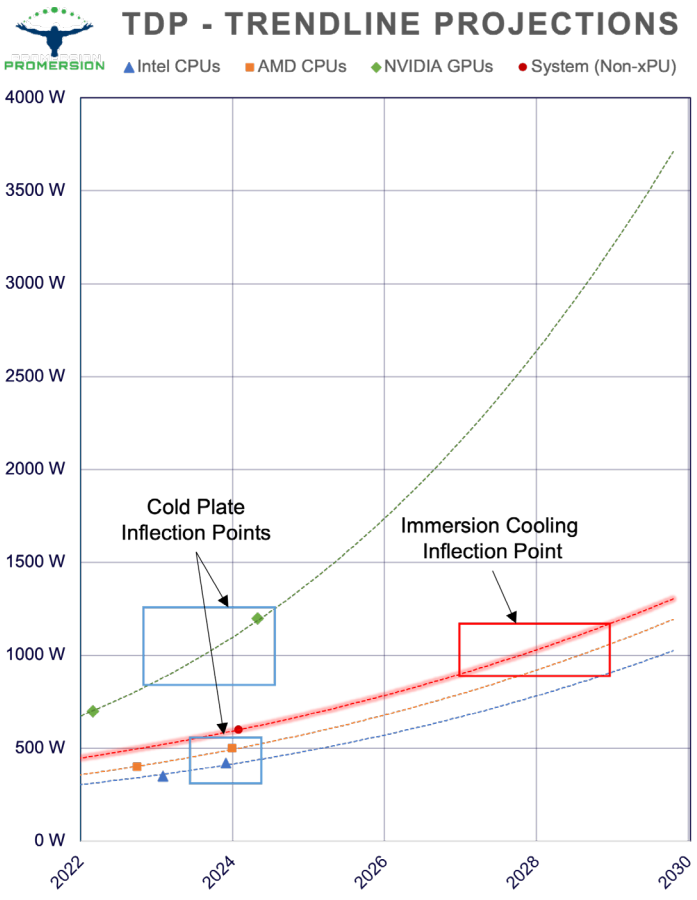


Fig. 2: Timeline for scaling cold plate and immersion technology[2]

Optimizing performance in an AI rack differs significantly from traditional workloads due to the importance of latency in AI processing. Latency refers to the time it takes for a system to respond to a request, and minimizing it is crucial for achieving optimal AI performance. In AI workloads, lower latency leads to faster processing and better overall system responsiveness, which is essential for maximizing computational efficiency.

In contrast, each rack in traditional data centers function independently, and the location of one rack within the data center has little-to-no impact on the performance of others. For instance, a storage-focused rack can be placed next to a networking rack without affecting their individual performance. Most modern hyperscale data centers are designed to allow racks to operate in parallel, providing flexibility for varied tasks. This setup is illustrated in Figure 3, which shows how such racks are arranged for optimal flexibility in general-purpose workloads.

However, this design is vastly different when it comes to AI clusters where the spatial relationship between racks becomes more critical because the system's performance is tightly coupled with latency and the speed of communication between nodes. In an AI cluster, optimizing performance requires a different rack layout compared to traditional workloads. Given the highly compute-intensive nature of AI tasks, racks must be placed in close proximity and operate in a serial fashion to ensure the required performance within a reasonable timeframe. This configuration allows the pooled compute resources across all racks to function cohesively, with the performance of each rack being interdependent on the others. As such, the failure of a single node in an AI rack can jeopardize the entire cluster, potentially resulting in significant financial losses. Figure 3 illustrates this key distinction, highlighting the difference in how traditional and AI workloads are managed.


Fig. 3:Data hall layout operating in parallel mode (left) and serial mode (right)[3]

To effectively scale AI workloads, data halls will need to be redesigned to accommodate AI-specific rack layouts. Current designs, such as the one shown in Figure 3, prioritize flexibility within the data hall. This flexibility allows for easy repositioning of racks when power usage approaches its maximum capacity, without affecting the performance of traditional workloads. However, in an AI data center, where racks must operate in a serial configuration for optimal performance, both power distribution and cooling infrastructures will need to be re-engineered to handle the high demand. This adjustment is crucial for supporting the unique requirements of AI workloads and ensuring that the system operates efficiently at scale as shown in Figure 3.

AI data centers must be designed with redundancy in mind to minimize the impact of failures, particularly given the high cost of downtime. A single AI rack equipped with multiple GPUs can be worth several hundred thousand dollars, meaning even minutes of downtime can result in millions of dollars in losses within a hyperscale data center. Therefore, ensuring that critical components are in close proximity to one another is essential for maintaining the performance and reliability of the cluster.

## INTEGRATING HEAT REUSE INTO DATA CENTER ARCHITECTURE

Heat reuse is another key example of how data center infrastructure must evolve in the age of AI. As research advances, the reuse of waste heat from liquid-cooled AI and accelerated compute data centers is poised to play a critical role in sustainable and efficient facility design. AI-driven data centers generate significant thermal energy, making heat reuse increasingly important. Liquid cooling is a game changer, as it enables the efficient capture and repurposing of waste heat. This not only improves overall efficiency and reduces power demands for cooling but also provides valuable benefits to local communities. Recovered heat can lower energy costs or support sustainable applications such as heating homes, pools, or greenhouses for agricultural production—turning a byproduct of AI workloads into a resource for broader societal and environmental gains.

This is particularly critical as communities worldwide push back against data center developments due to concerns over energy and resource consumption. In Europe, where district heating infrastructure is already in place, some countries have implemented regulations mandating the reuse of data center waste heat. In several regions, this has led to waste heat being repurposed to warm apartment buildings. These policies have also spurred a broader global consideration of heat reuse in new data center designs—including in the U.S., where some hyperscale facilities are now incorporating immersion cooling with plans to use excess heat for greenhouse crop production.

## FUTURE-PROOFING DATA CENTERS

AI demands fundamental design changes to future proof data centers. This will only happen through ecosystem enabling efforts between government, industry, and nonprofit groups to support the buildout of sustainable infrastructure at scale. Organizations like the Open Compute Project Foundation (OCP) and ASHRAE are actively addressing the technical challenges of AI by developing best practices for rack layout design and new enabling technologies such as liquid cooling and heat reuse. Once technical bodies identify best practices and research gaps, industry advocacy groups such as the Liquid Cooling Coalition work to secure governmental support for research and policies to incentivize adoption of critical new technologies. Continued industry collaboration through both these technical and advocacy bodies is vital to enable the scalable deployment of evolving AI-driven infrastructure.

## REFERENCES

1. Data reproduced from “Coolant Temperature for Next Gen IT and Durable DC designs, OCP Regional Conference, 2023.

2. Title - Timeline for Cold Plate and Immersion Technology
   Reference: Brink, Rolf. Cooling Environments Insights.
   OCP EMEA Summit, 30 Apr. 2025, Promersion. Reproduced with permission.

3. Images created from Meta AI, https://www.meta.ai/, Date Accessed February 7, 2025

TechArena 2025