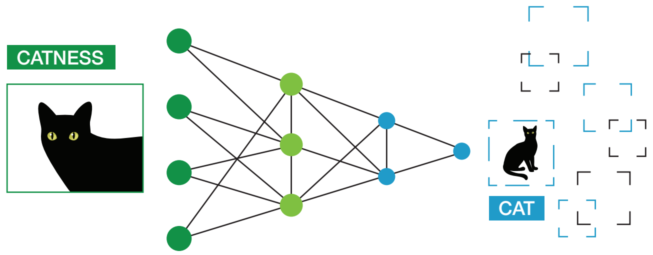
Direct liquid cooling of AI infrastructures enables greater efficiency and reliability

Why AI is suddenly such a hot topic – and how to cool it down

There's no doubt that the demand for AI has blown sky high. However, the density of chips that power AI compute infrastructures has also exploded. Running increasingly complex and data-intensive AI algorithms requires a number of GPUs and accelerators, which can incorporate hundreds of billions or even trillions of transistors. And this trend is not likely to abate any time soon, which is partly because the AI industry is reaching an inflection point.



Training vs. Inference—AI developers train neural networks on massive data sets (with as many as billions or trillions of parameters) to create 'foundational models' that understand general topics, which they further fine-tune to perform specific useful tasks. For example, a model trained on 'catness'—triangle ears, fuzzy tails, round eyes, etc.—could determine (infer) that a specific image is, in fact, a cat. Both training and inference require compute power—a lot of it.

In the short period following the debut of ChatGPT and similar generative artificial intelligence (GenAI) applications, most of the processing for GenAI has involved training large language models (LLMs), not in applying those models to solving real-world problems. But now, many models that were busy learning general patterns from massive data sets have been fine-tuned to perform useful tasks.

This means GenAI workloads are graduating from training to what is known as inference, or the execution phase where a fully trained AI model evaluates real-world data—new, unseen data, not the data on which it was trained prior to release—and provides valuable benefits for humans. For example, during a chat with an LLM, you may ask the AI to write a one-paragraph story about your dog based on details in a prompt and in the style of Kurt Vonnegut. While the model may have been trained on broad topics—such as the general characteristics of dogs vs. cats or English literature or English grammar—Inference puts that general training to a specific use.

In other words, inference is where the rubber meets the road in GenAI applications, from generating a summary of NVIDIA's history in manufacturing GPUs in just seconds to driving decision-making for growing a grocery business.

Original promo poster for Kurt Vonnegut's award-winning novel, *Slaughterhouse-Five*

> “With cooling becoming a paramount concern in AI data centers, suddenly running AI workloads has morphed from a computer science problem to a chemical and mechanical engineering problem. That’s because the massive complexity of the latest AI chips translates into unprecedented heat, which requires deploying new solutions and careful data center planning and implementation to keep server racks cool.”

Ken Howard
CEO and Founder, CommetCo

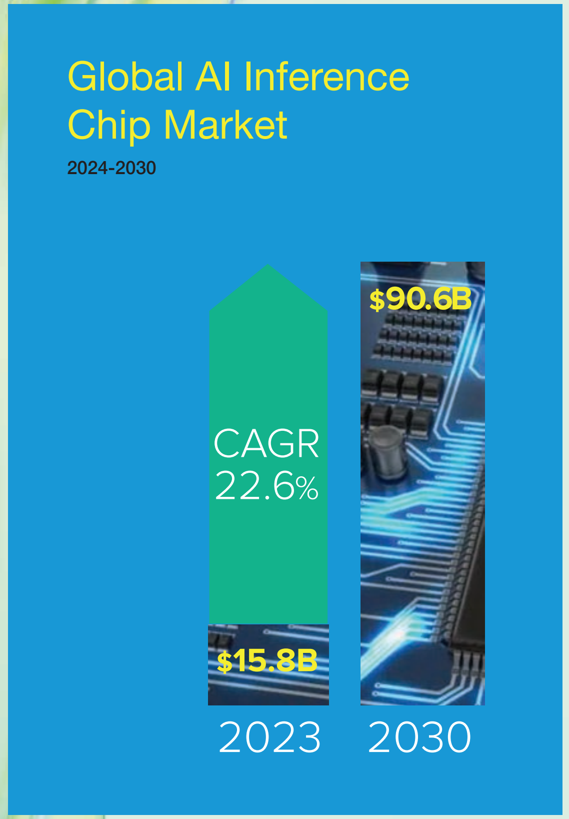


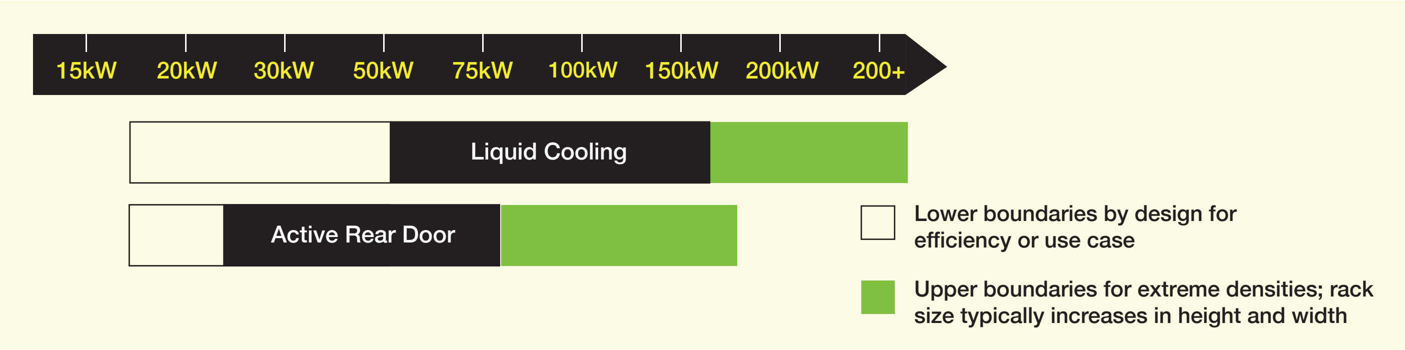
Figure 1, The AI Inference Chip Market is projected to reach USD 90.6 Billion by 2030, representing a CAGR of 22.6% during the period 2024-2030.

All these amazing capabilities promise to yield explosive growth in the AI chip and compute infrastructure market (Fig.1). In fact, industry research suggests the AI Inference Chip Market will reach USD 90.6 Billion by 2030, representing a 22.6% CAGR.[1] Factors driving this trend include the growing use of artificial intelligence across a range of industries—from healthcare and life sciences to financial services to manufacturing.

Meanwhile, handling the massive size and growing complexity of datasets available for training GenAI models—combined with demand for inference applications in almost every industry—drives software developers to create more sophisticated models. Similarly, as deep learning methods advance and become more complicated there is an increasing need for powerful compute infrastructures capable of managing intricate neural network topologies. These models require faster hardware that can process and send data across complicated neural networks quickly and efficiently.

Cooling off period

Now, whether the demand for more powerful AI compute infrastructure stems from training models or using the models for valuable inference, or both, it has hit a wall in terms of cooling.



Since the inception of data centers, keeping server racks cool has relied on air cooling. Traditionally, this involves circulating cool air through server racks and then passing the hot exhaust air through mechanical chillers to absorb heat—and then sending it back to the servers—helping the chips maintain optimal reliability and efficiency. But as AI chips become hotter and faster, maintaining those optimal conditions and preventing failures not only becomes more difficult but also more expensive and more energy-intensive.

Not long ago, 20 kilowatts per rack was not uncommon as peak energy consumption in a data center. However, AI hardware is typically deployed in tightly packed server racks, resulting in higher power densities than in traditional data centers. These high-density racks can sometimes exceed 400 kilowatts per rack, making it orders of magnitude harder to dissipate the heat they generate. by high-density racks. This means that as AI workloads grow, traditional air-cooling methods will no longer be enough.

Data center cooling must also become more energy-efficient so that AI can meet sustainability goals. Recently, the International Energy Agency (IEA) reported that data centers globally, used 2% of all electricity in 2022 and the IEA predicts that could more than double by 2026. Therefore, data centers will need to run AI workloads more efficiently, and today’s facilities are not equipped to support the energy-intensive cooling demands of growing HPC processor power.

A cool 208 billion

Keeping AI and other compute-intensive HPC applications running efficiently and reliably requires new approaches to cooling. AI applications typically employ massively complex GPUs and CPUs to run at acceptable speeds. For example, the latest NVIDIA GPUs incorporate 208 billion transistors. Super-high-powered semiconductor devices like these generate significant heat, which if not managed properly can push chip operating temperatures beyond safe limits and lead to reduced MTBF (mean time between failure) as well as other system issues. This makes it crucial to keep AI compute infrastructure cool by deploying the latest data center cooling methods.

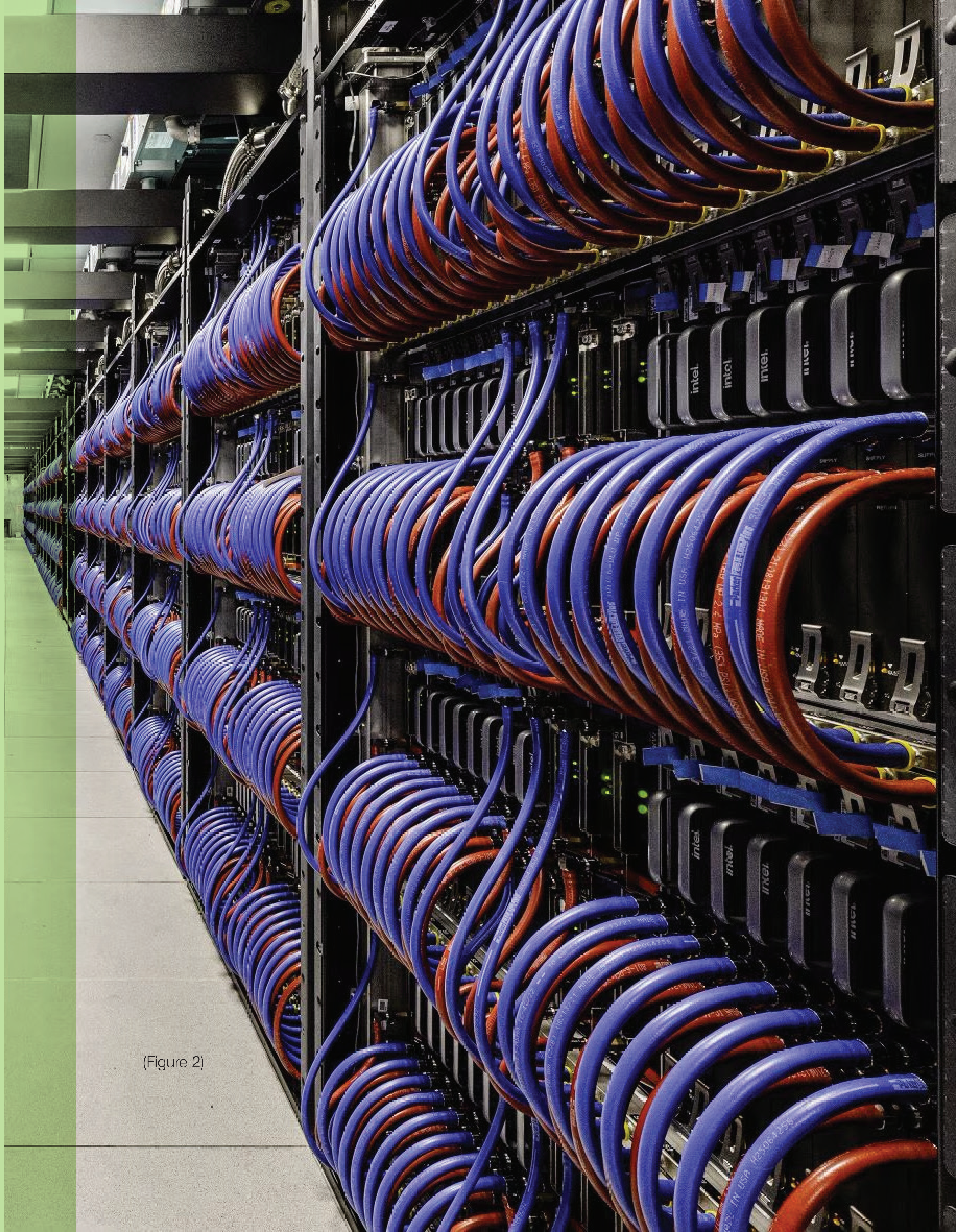
Thus, we are entering the era of Direct Liquid Cooling (DLC) — the most effective way to cool next-generation AI systems. While not every AI project needs the power of a supercomputer, Hewlett Packard Enterprise (HPE) has pioneered advanced cooling solutions in parallel with exascale computing, deploying the three top systems in Top500 supercomputing list and the only three exascale systems on the planet. For example, HPE recently launched Aurora, the world's most powerful supercomputer purpose-built for AI. With its 21,248 Intel processors and 63,744 Intel GPUs, the HPE Cray EX machine can run AI workloads at speeds that not long ago were considered impossible.

Exascale - Unthinkable artificial intelligence speed

Let's try to put the ‘mind-boggling speed of exascale computing in perspective. The human brain can solve one simple arithmetic problem, say for example, 2+2 = 4, in approximately one second, which in computer terms translates to one FLOP, or floating point operation per second. “A modern personal computer processor can operate in the gigaFLOP range, at about 150,000,000,000 flops, or 150 gigaFLOPS,” explains the US Department of Energy. “Tera” means 12 zeros. Computers hit the terascale milestone in 1996 with the Department of Energy's Intel ASCI Red supercomputer. ASCI Red's peak performance was 1,340,000,000,000 FLOPS, or 1.34 teraflops. Exascale computing is unimaginably faster than that. ‘Exa means 18 zeros.”

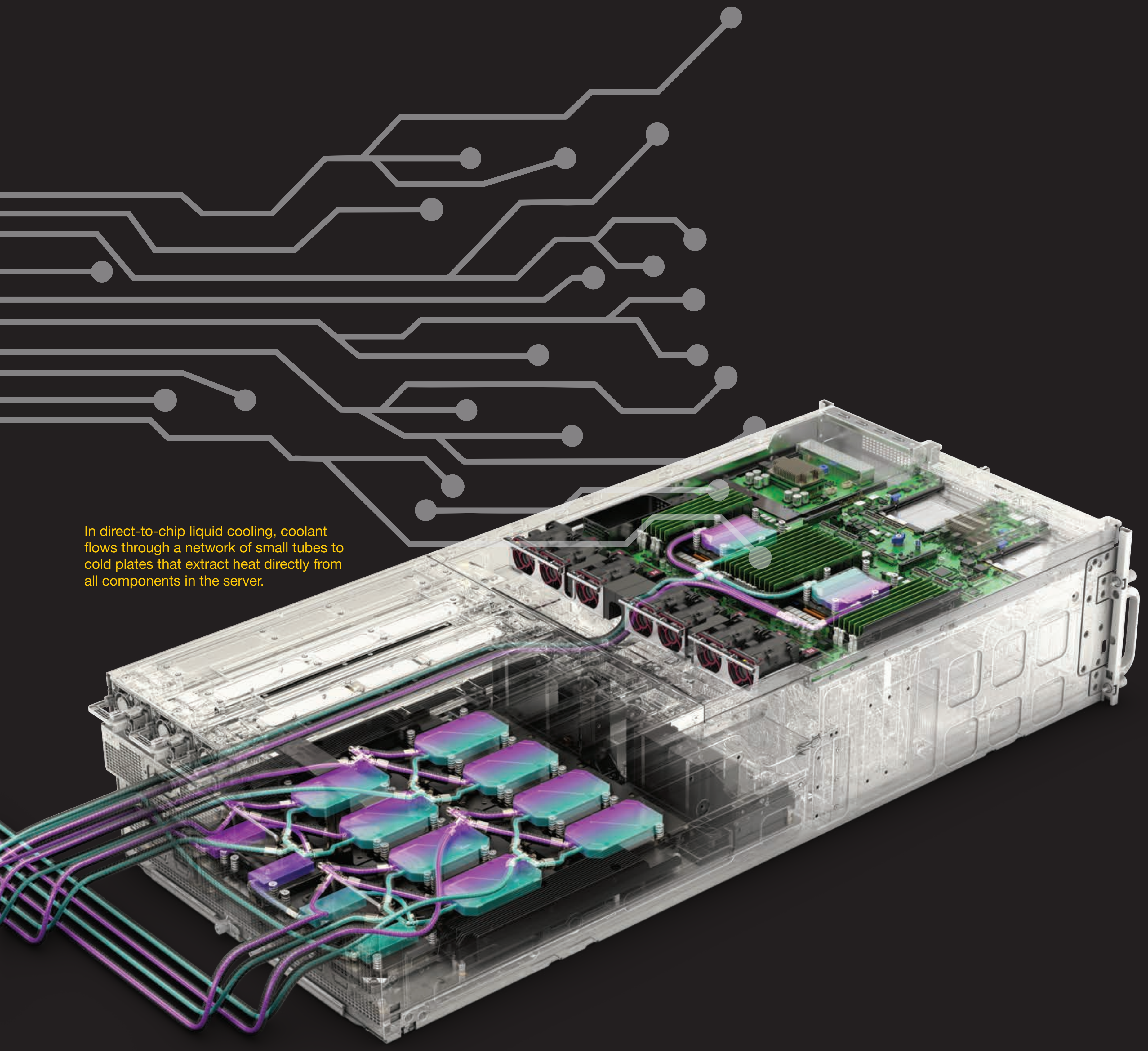
Aurora, for instance, can perform 1,012,000,000,000,000,000 FLOPS, or just over one exaflop That is more than one million times faster than ASCI Red's peak performance in 1996. Planned as an AI-capable system from inception, that kind of processing power enables Aurora to run generative AI models to accelerate scientific discovery. Early AI-driven research that scientists have run on Aurora include brain mapping to better understand the human brain's 80 billion neurons, high energy particle physics enhanced by deep learning, and machine-learning accelerated drug design and discovery, among others.

However, all that compute power generates so much heat that, for example, Aurora requires a specially engineered DLC system.



(Figure 2)

Figure 2, The Intel/HPE Cray EX ‘Aurora’ supercomputer at the U.S. Department of Energy's Argonne National Laboratory is capable of exascale computing. Tens of thousands of Intel chips generate so much heat that HPE engineered a direct liquid cooling system to keep the machine operating efficiently and reliably. The blue tubes deliver chilled liquid into the server blades and directly onto the chips, while the red tubes carry away the heated water.



In direct-to-chip liquid cooling, coolant flows through a network of small tubes to cold plates that extract heat directly from all components in the server.

HPE ProLiant Compute XD685 with NVIDIA

D2C — A more direct approach to keeping AI running efficiently and reliably

In short, we've come a long way from running chilled air through server racks. Which is good because data center cooling systems, including fans, CRAC (Computer Room Air Conditioner) units, and other components of air-cooling systems, can account for a substantial portion of a data center's total energy usage—sometimes as much as 40%.

One new method that can substantially cut energy usage, while offering other benefits, is direct-to-chip liquid cooling, or D2C, which runs coolant through tubes and cools GPUs and CPUs directly. This approach offers several advantages, including:

Enhanced Efficiency: Removes heat more quickly and efficiently, requiring less energy to cool the AI compute infrastructure.

Higher Rack Densities: Compared to air-cooling, D2C allows data centers to pack more servers into a given space, increasing compute capacity.

Extended MTBF: By maintaining chips within safe operating temperatures, direct-to-chip cooling helps prevent failures and extends the lifespan of the chips.

The exploding demand for compute power to train AI foundational models (which can now contain as many as a trillion parameters) and run inference using those trained models, has further driven innovations in cooling. One new method for example, immersion cooling, submerges the entire server in liquid. Direct liquid cooling (DLC) and direct-to-chip liquid cooling (D2C) approaches, on the other hand, drive coolant flows through a network of tubes and cold plates to extract heat directly from critical components on the server, which is often augmented with cooling fans.

Unlike in traditional air cooling, in direct to chip (D2C) systems, chilled coolant runs through small channels and to cold plates that actually touch the chips. This direct contact allows the coolant to absorb heat more efficiently. Cooling Distribution Units (CDUs) then pump away the warm liquid away to a dry cooler or an external radiator where it is cooled and recirculated.

Larger AI data centers, which may have 400 kilowatts and greater racks that produce extreme heat, may circulate warm liquid from the data center to a cooling tower, where exposure to air causes evaporative cooling. The cooled water is then recirculated back to the data center where chillers further reduce the temperature. This approach is particularly advantageous in AI data centers and high-performance computing environments where managing excessive heat is crucial for maintaining system stability and optimizing performance.

To ensure the efficiency and performance of the most powerful AI compute infrastructure, HPE recently introduced the first 100% fanless DLC system architecture—that is, an all-direct-liquid-cooling solution that dispenses with cooling fans entirely. This latest advance in DLC yields 90% reduction in cooling power consumption and costs as compared to traditional air-cooled systems.

100% fanless DLC yields 90% reduction in cooling power consumptio and costs as compared to traditional air-cooled systems.

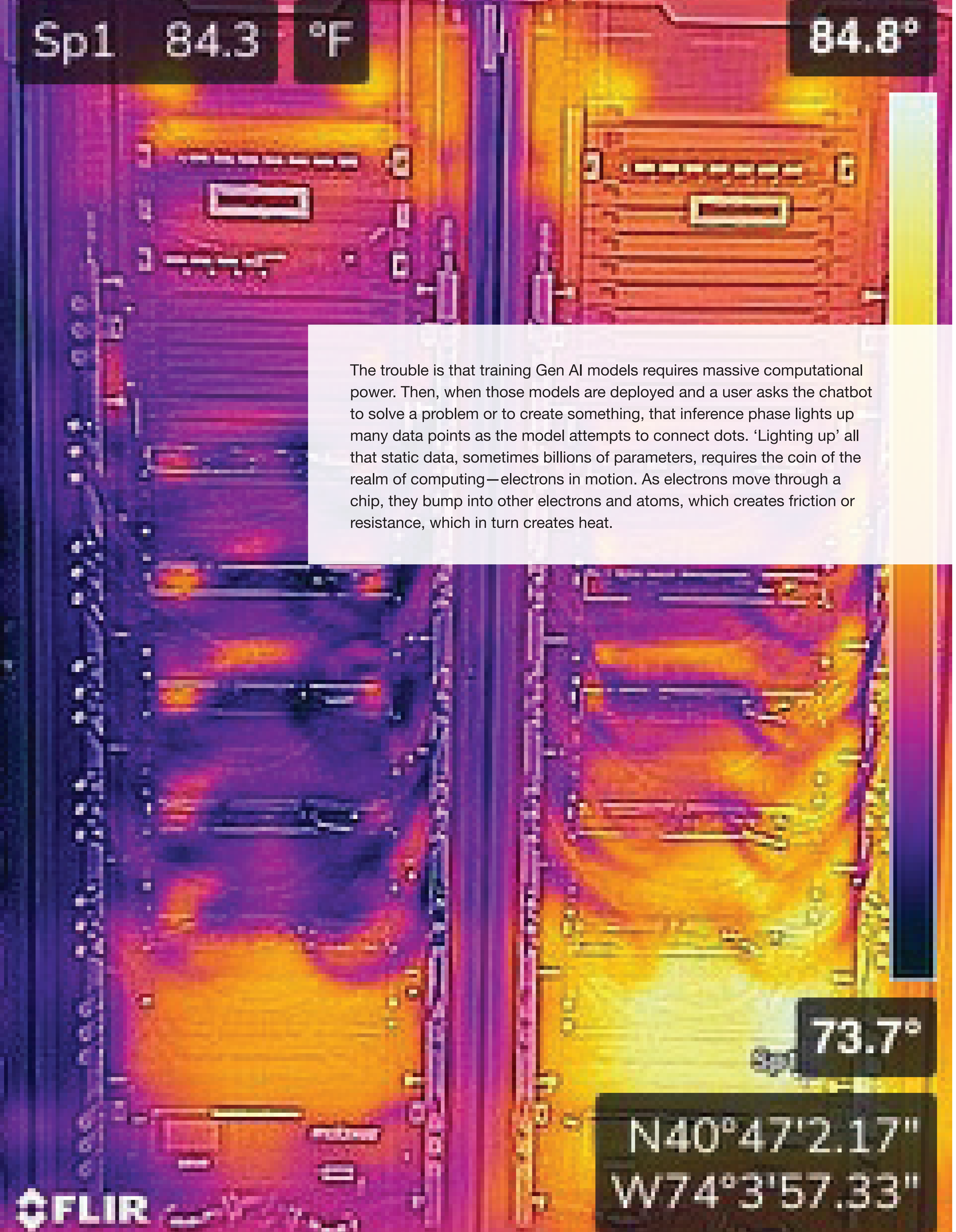
Mix it up

Despite its significant benefits, 100% DLC is not appropriate for every AI factory. A more typical method involves a blended approach. This hybrid method mixes direct liquid cooling with rear door heat exchangers (RDHX). RDHX systems use fans to draw in hot air from servers. This approach allows for targeted cooling of high-density racks while leveraging the efficiency of liquid cooling and potentially reducing the need for traditional data center air conditioning units.

All these approaches face several challenges for teams deploying and maintaining liquid cooling systems. In DLC systems, for example, after the coolant is expelled from the server it then needs to be pumped away by Cooling Distribution Units (CDUs) to a dry cooler or an external radiator to be cooled and recirculated. For one, in a hybrid system, you must still cool the water coming from the RDHX. A typical AI data center, which can consume as much electricity as 100,000 households, may circulate warm water from the data center to a cooling tower, where exposure to air causes evaporative cooling. The cooled water is then recirculated back to the data center where chillers further reduce the temperature.

Liquid Cooling
sustainability and cost savings

20% More performance per kilowatt(kW)
87% Carbon reduction
86% Cost savings



The trouble is that training Gen AI models requires massive computational power. Then, when those models are deployed and a user asks the chatbot to solve a problem or to create something, that inference phase lights up many data points as the model attempts to connect dots. 'Lighting up' all that static data, sometimes billions of parameters, requires the coin of the realm of computing—electrons in motion. As electrons move through a chip, they bump into other electrons and atoms, which creates friction or resistance, which in turn creates heat.

Cool Tools – In a hybrid DLC deployment, experts use tech tools like anemometer and forward looking infrared cameras (FLIRs) to measure heat and airflows.

For the air cooling portion of a hybrid approach, a DLC deployment expert must still calculate the cooling capacity of the data center, which includes measuring air temperatures at different levels (Figure 3), air flow from floor (cool air) to ceiling (where the hot air is exhausted). Or engineers may need to analyze heat maps created using forward-looking infrared, FLIR, cameras. Then experts also need to consider factors such as the size of the grates in the floor or use an anemometer to measure air flows. All this requires years of expertise coupled with careful design, planning, and onsite system tuning.

Another factor is the coolant itself. First the water coming into the data center for RDHX cooling typically comes from a municipal water supply. And great care must be taken to ensure that iron-related bacterial growth does cause clogging of systems. DLC systems present different challenges. They use coolants, like propylene glycol (PG) or ethylene glycol (EG) mixed with water, which contain additives like corrosion inhibitors and biocides to enhance performance and longevity. Nonetheless, ensuring that this green slime does not degrade the DLC equipment comprises tricky chemistry.

"With cooling becoming a paramount concern in AI data centers, suddenly, running AI workloads has morphed from a computer science problem to a chemical and mechanical engineering problem," says Ken Howard, ComnetCo Founder and CEO. "That's because the massive complexity of the latest AI chips translates into unprecedented heat, which requires deploying new solutions and careful data center planning and implementation to keep server racks cool."

To learn more about how HPE and ComnetCo can help you solve your AI compute infrastructure cooling challenges, explore HPE.com, and comnetco.com.

ComnetCo

About ComnetCo

Two decades of experience in the evolution of HPC helps ComnetCo configure powerful compute and storage systems. This virtually unequalled track record includes delivering some of the world's fastest supercomputers. Together with its primary partner, HPE, ComnetCo helps optimize systems for the unique needs of researchers in Higher Education, Research Institutes, Global Enterprises, and Federal Government Agencies. These solutions—which include purpose-built platforms for AI—help scientists and engineers speed time to discovery in fields ranging from pharmaceutical research like new vaccines to industrial companies creating new materials to supporting deep space exploration. For more information, visit: **www.comnetco.com**

HPE

About Hewlett Packard Enterprise

Hewlett Packard Enterprise (NYSE: HPE) is the global edge-to-cloud company that helps organizations accelerate outcomes by unlocking value from all of their data, everywhere. Built on decades of reimagining the future and innovating to advance the way people live and work, HPE delivers unique, open and intelligent technology solutions delivered as a service – spanning Compute, Storage, Software, Intelligent Edge, High Performance Computing and Mission Critical Solutions – with a consistent experience across all clouds and edges, designed to help customers develop new business models, engage in new ways, and increase operational performance. For more information, visit: **www.hpe.com**