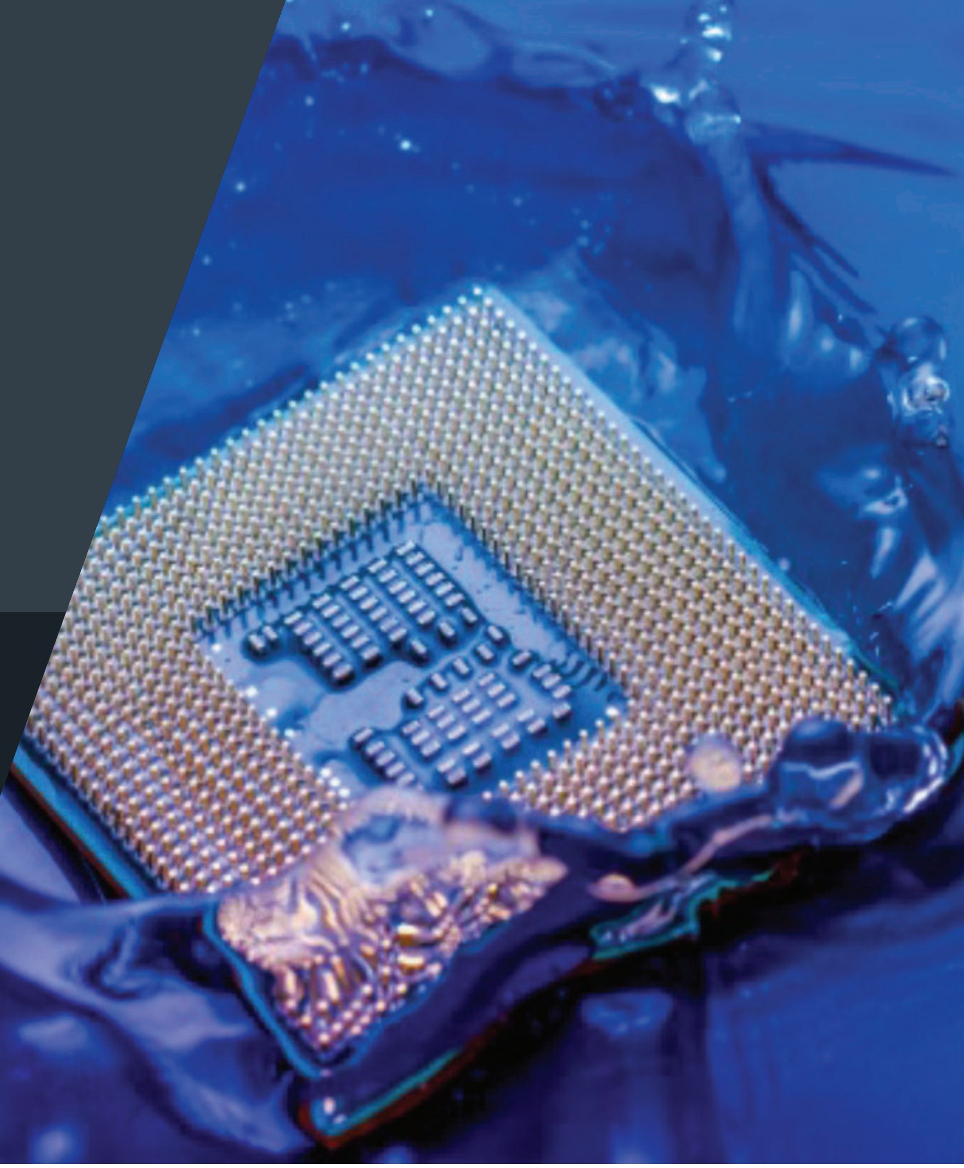
WHITEPAPER

COOLING THE FUTURE

Hybrid Cooling Strategies for Energy Efficient HPC Data Centers

By Gordon Johnson



This white paper shows how hybrid cooling, combining air and liquid cooling methods, enables efficient, scalable, and sustainable data center operations for AI-driven HPC. It outlines key technologies, deployment strategies, and Subzero Engineering's eco-friendly solutions for future-proof designs.

INTRODUCTION

The future of data center cooling has arrived, and it's a hybrid combination of air and liquid cooling technologies. With the arrival of High Performance Computing (HPC) to support Artificial Intelligence (AI) and Machine Learning (ML) workloads comes a rapid increase in chip, server, and rack densities, along with greater power demands and escalating heat generation with rack power approaching 100 kW and higher. Traditional air cooling methods alone are no longer sufficient to manage the thermal load of this increasingly high-density infrastructure, resulting in the move towards hybrid cooling.

Hybrid cooling is the future of cooling because it allows data centers to scale up performance without a full infrastructure overhaul, making it ideal for brownfield sites (existing) looking to retrofit for a gradual transition toward more advanced cooling methods. This has also become the cooling technology method of choice for greenfield (new) data centers being built that are prioritizing long-term growth and sustainable goals.

As the data center industry embarks on an unprecedented future when it comes to cooling ITE, various questions often arise. What technology options are there today for Direct Liquid Cooling (DLC), and where does the industry appear to be going and not going with these technologies? For both legacy retrofits and new greenfield builds, how do designers and managers meet escalating cooling demands while minimizing their overall Total Cost of Ownership (TCO)?

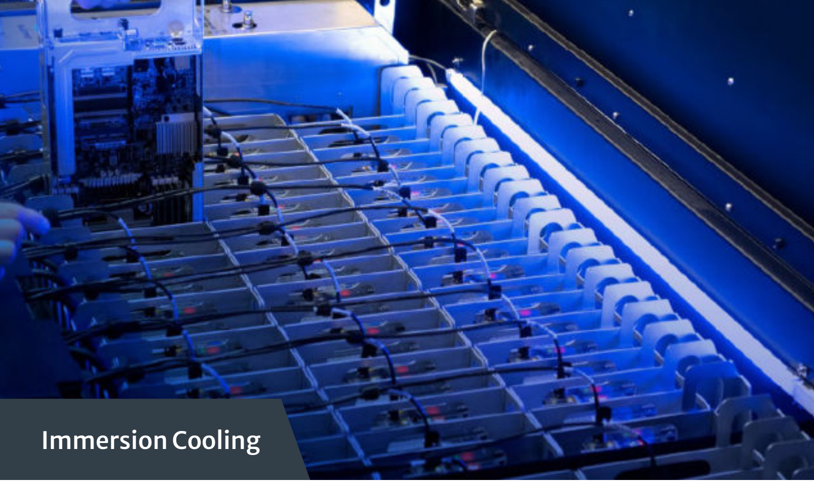
This paper serves to answer these questions while showing how to implement hybrid cooling strategies that are flexible, future-proof, and sustainable with the following topics:

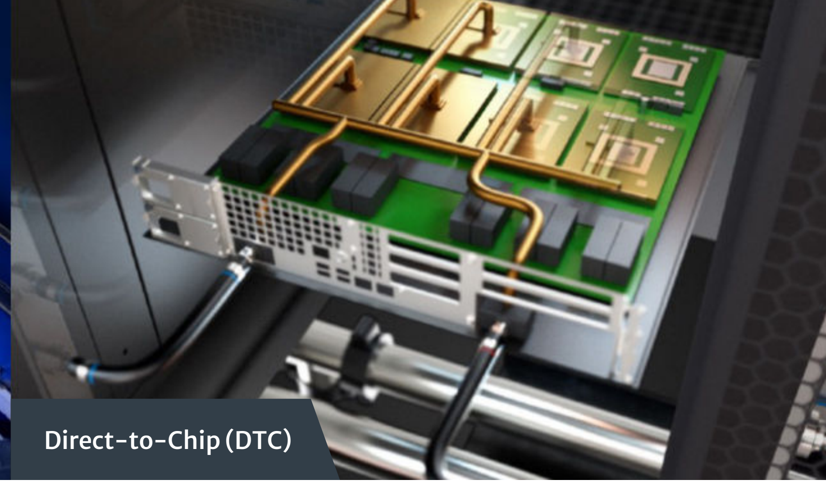
How Hybrid Cooling Works

Hybrid cooling refers to the combined use of air and liquid cooling technologies to manage the heat generated by both low-density and high-density computing equipment. It's a flexible, transitional approach that allows data centers to optimize cooling efficiency, performance, and cost as workloads evolve, especially with the rise of HPC systems due to AI.

How does hybrid cooling work? Air cooling handles heat removal for the lower-density or legacy ITE, while DLC, such as Direct-to-Chip (DTC) targets high-heat components like Central Processor Units (CPUs) and Graphics Processor Units (GPUs). Air cooling is also required to remove approximately 25% or more of residual heat that DTC does not remove from HPC racks. Both air and liquid cooling operate simultaneously, along with either cold or hot aisle containment to separate the cold supply air from the hot exhaust air for optimum thermal management in the data center.

Hybrid cooling also refers to the strategic approach of deploying DLC solutions for high-density HPC while using existing air cooling for low-density ITE, separating the location of the equipment into different low-density and high-density zones within the data center. This approach helps maximize efficiency while reducing the carbon footprint. In addition, this also can have financial benefits by reducing the upfront cost of adding DLC since existing infrastructure with air cooling can continue to operate and DLC can be selectively deployed as needed. Since this approach allows for a gradual transition, it also means data centers don't have to choose just one type of DLC technology over another until needed.


Immersion Cooling


Direct-to-Chip (DTC)

Since DLC is a rapidly changing environment void of the standards that exist for air cooling, hybrid cooling is a great approach to future-proofing data centers so that they continue to be operationally useful and cost effective into the future.

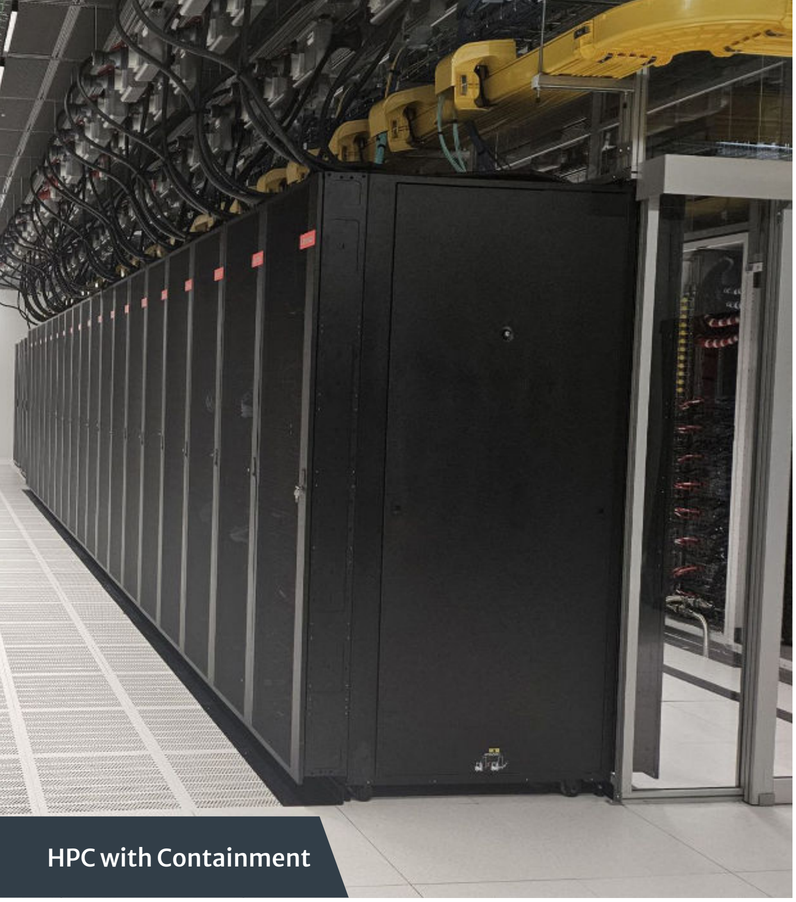
DLC Technologies, Who's Leading and Who's Lagging

When it comes to heat transfer, water is always going to be more efficient than air because it's denser, has a higher specific heat capacity and lower thermal resistance. Since water can hold over 3,000 times more heat by volume than air, this means that liquid cooling can effectively transfer heat from the CPU/GPU using a significantly smaller volume of liquid coolant compared to the air volume required for equivalent air cooling. Using liquid to cool data centers also increases economization hours per year, improves CPU/GPU efficiency, and makes possible waste heat reuse. It's also a quieter data center since ambient noise is reduced, making the data center a more human friendly environment.

What are some different DLC technologies being used, and who's currently leading and lagging in this market? In 2025, DTC (single-phase) is by far the method of choice, making it an excellent choice for HPC and high-density zones. With this method, liquid coolant is circulated and directly cools the components generating the most heat (CPUs/GPUs) via cold plates attached to these chips. Coolant flowing through the cold plates transports the heat energy out of the server, thus removing heat at the source. The heated liquid is then transported or returned to the Cooling Distribution Unit (CDU) which acts as a heat exchanger to eventually reject the heat from the building. This is often referred to as an open-loop liquid cooling system since it requires an external water source like a cooling tower to reject heat from the facility.

DTC (two-phase) is somewhat similar except the heated liquid is turned into a gas and then compressed back to liquid in order to repeat the cooling process. This is often referred to as a closed-loop liquid cooling system that efficiently removes the high amount of generated heat without the use of an external water source.

Immersion cooling is even more efficient at removing heat and offers higher cooling capacity by submerging the entire server vertically in dielectric fluid. With single-phase immersion cooling, the coolant is continuously circulated to dissipate the heat using hydrocarbon dielectric fluids due to their high boiling points. Heated coolant exits the top of



HPC with Containment

the servers to a CDU/heat exchanger and returns to the servers at user specified temperatures. With two-phase immersion cooling, fluorocarbon dielectric fluids are used due to their low boiling points, resulting in the dielectric fluid boiling from the server heat and thus transforms from liquid to vapor. Next, the vapor passively condenses back to liquid when it comes in contact with water-cooled condenser coils. This technology is extremely efficient but difficult to maintain due to fluid types and operating temperatures. No additional air cooling or containment is needed for either type of immersion cooling.

Despite immersion cooling's exceptional efficiency, it requires specialized equipment and training and is lagging in terms of its acceptance in the overall DLC market today. Major hyperscalers and smaller data centers find DTC, especially single-phase, easier to integrate into existing and future rack designs, and thus the remainder of the paper will focus on single-phase DTC as the industry preferred form of DLC. **(see Immersion Cooling and DTC photos on previous page)**

Pros and Cons of Brownfield Retrofit Versus Greenfield Build

There are several differences, including pros and cons to both updating brownfield (existing) and building greenfield (new) data centers. Since brownfield requires retrofitting existing buildings, some pros are they're typically more cost-effective, faster to deploy and can be anywhere that already has existing utilities and connectivity. Some cons are the data center and/or retrofit will be limited by the existing footprint and infrastructure, an existing building may struggle with energy inefficiency, and retrofitting new DLC solutions can be challenging.

For the greenfield data center, since you're designing and building from scratch you're basically working with a blank canvas for implementing efficiency and sustainability goals. The pros are easier design layout for HPC ITE, scalability for future expansion, and the ability to choose the optimal site location for power, water and network access. The cons are higher upfront costs including land purchase and construction, and longer timelines to have a fully operational data center.

Deploying hybrid cooling will also be different in both types of data centers. For example, since a brownfield site already has air cooling, adding HPC workloads will require a form of DLC like DTC. This may simply mean keeping the existing ITE in a low-density zone, while selectively adding HPC workloads into a high-density zone. One pro to this deployment is that it can significantly reduce both TCO and

Brownfield vs. Greenfield

disruption often caused by trying to upgrade the entire data center's cooling infrastructure.

Since the greenfield site is being built from scratch, designers can more easily choose and implement the latest cooling infrastructures meaning DLC can be implemented immediately. Similar to the brownfield site, the HPC ITE should still be separated from the low-density ITE via zones, allowing for a more efficient and cost-effective hybrid cooling system.

When evaluating a site for HPC, typically a greenfield site will offer the overall flexibility needed for the optimal design, but if speed and cost are top priorities, brownfield could still be the better choice. **(see Brownfield vs. Greenfield photo)**

CDUs – The Brains Behind DLC

CDUs have earned earn the nickname "the brains" behind DLC because they do far more than just move fluid for the liquid cooling system. CDUs help orchestrate the entire thermal management system with precision, intelligence, and adaptability. They don't just circulate liquid, but they think, adapt and protect the liquid cooling portion of the hybrid cooling system.

With single-phase DTC the CDUs act as the bridge between the Facility Water System (FWS) and the dedicated Technology Cooling System (TCS) that provides filtered liquid coolant directly to the chips via cold plates. Unlike two-phase DTC described in an earlier section, single-phase DTC is often referred to as an open system because of needing either external chilled water via the FWS or even conditioned air to remove heat from the TCS via the CDU. Additionally, the CDU is connected via inbound and outbound pipes to the chips with cold plates.

CDUs also act to isolate the TCS from external water particles and variating pressure, ensuring a safe and stable environment. Further, CDUs act as "the brains" by regulating liquid coolant temperature to stay above dew point, thus preventing condensation at the chip level. They also maintain enhanced filtration to remove particulates that protect the integrity of the server cold plates and maintain their performance. CDUs also are equipped with variable speed pumps, automated valves, and sensors to dynamically adjust the flow rate and pressure of the TCS, ensuring optimal cooling even when HPC workloads change.

CDUs can be installed to support one rack (in-rack where the TCS is inside the rack), in-row, or in a perimeter layout at the end-of-rows with the option of placing CDUs in separate mechanical spaces. CDUs

can also strategically be placed as needed in order to cool new high-density zones, making them a perfect solution for retrofitting portions of existing data centers needing HPC. CDUs typically have some type of built-in pump redundancy to ensure constant flow but depending on the size and layout of the data center, redundant CDUs and centralized control may be needed to ensure 24/7 operation.

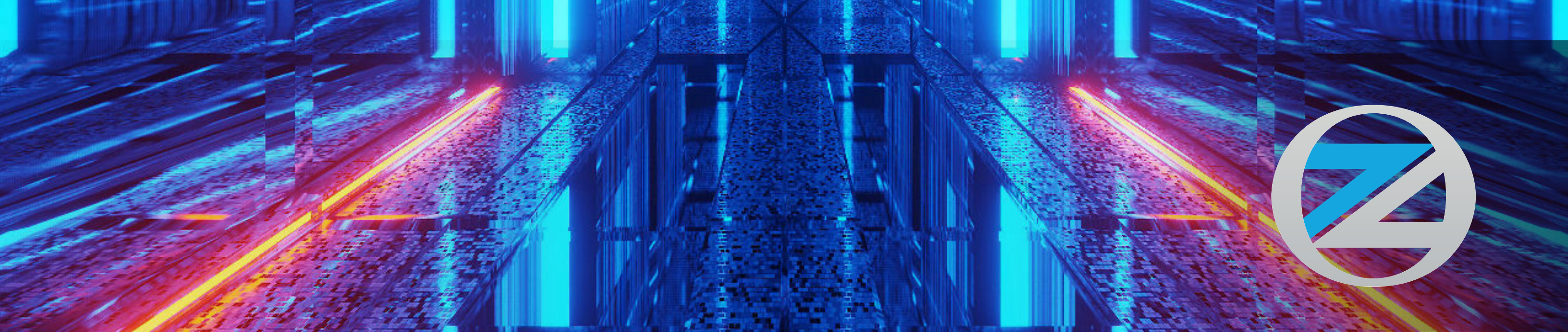
CDUs typically fall into two categories, Liquid to Liquid (L2L) and Liquid to Air (L2A). L2L CDUs are ideal for large scale HPC data centers due to their higher cooling capacity but can be complex since they need to be connected to the facility's chilled water supply or some other heat rejection source to continually provide liquid to the cold plates for cooling. L2L CDUs have heat exchangers which receive the CPUs/GPUs heat and transfer that heat into the isolated chilled water supply or FWS, completing the liquid-to-liquid cooling process.

L2A CDUs don't require a chilled water supply or FWS but instead transfer heat from the returning liquid coolant from the cold plates to the surrounding data center air using an internal air-cooled heat exchanger, that heat is next removed via traditional air cooling and heat rejecting equipment such as rooftop condensing units. L2A CDUs are best for smaller deployments due to limited cooling capacity compared to L2L CDUs but are more cost effective for smaller DTC deployments. **(see L2L and L2A photos)**

DLC Best Practices

The shift towards DLC expresses the industry's desire to be more energy efficient and sustainable. What are some best practices to keep in mind when deploying DLC with a hybrid cooling infrastructure?

During the design phase, you'll want a trusted partner to perform a CFD study for both legacy and new greenfield designs. CFD modeling will ensure that sufficient air cooling equipment is available to be leveraged in the new hybrid cooling infrastructure. You'll also need FNM (Flow



Networking Modeling) to select the correct CDUs, size piping for flow rates, choose manifolds, and evaluate the ability of the cold plates to support HPC workloads.

As discussed earlier, containment continues to be an integral part of any data center and DLC deployments are no different. While the majority of ITE generated heat is removed via the cold plate heat transfer, residual heat (up to 25%) will still need to be removed by traditional air cooling, meaning containment (CAC or HAC) should always be installed in either retrofitted or new DLC designs.

Additional DLC best practices include ensuring power distribution provides for higher amperage required by HPC. Most data centers still operate with 200-250A power panels, designed and specified well before HPC workloads existed, so proper planning is needed to ensure that enough power is available for these high-density racks.

Choosing the right rack dimensions is also important to support both the ITE and the additional pipe spacing needed. This includes choosing rack heights between 42U and 52U, rack widths @ 800 mm (31.5”) minimum, rack depths @ 1200 mm (48”) minimum, and racks that are able to support the increased weight of HPC servers, with some racks already crossing the threshold into the range of 6,000 lbs. and higher.

Consideration also needs to be given to extra ceiling height due to increased rack height, and other infrastructure choices such as pipe placement will need to be determined. For example, when piping is deployed under raised floors to avoid the potential for overhead leakage, the piping needs to adequately fit in the space under the floor and the raised floor needs to support the racks and infrastructure above. For slab floors, piping will need to be supported in some manner such as threaded rods hung from ceiling joists, a high capacity load-bearing structural ceiling grid, or with a frame-based ground supported infrastructure. Power and data cable also need to be properly supported, so a trusted containment partner should always be used to ensure that the containment system can support power, cable, and liquid piping as needed.

The selection of materials for the TCS is also vital to prevent material incompatibility which can result in expensive damage and downtime. For example, the liquid coolant circulating through the TCS comes in contact with the CDUs, various connectors, seals, piping, valves, cold plates, etc. If any of these materials and metals are incompatible with each other in the presence of liquid, it starts a process

> **FUTURE-PROOF APPROACH**
> ***Hybrid cooling is the future-proof approach that balances efficiency, cost, and sustainability in the AI-driven data center era.***

known as galvanic corrosion, which creates debris within the liquid coolant and not only threatens the CPUs/GPUs via the connected cold plates, but also increases risks of leaks and other damage throughout the entire TCS loop.

The best practices mentioned in this section are not inclusive but provide just a few examples for DLC deployments, however it's always recommended to work with trusted vendors from start to finish for all hybrid cooling infrastructure projects.

Introducing Composite AisleFrame (CAF)

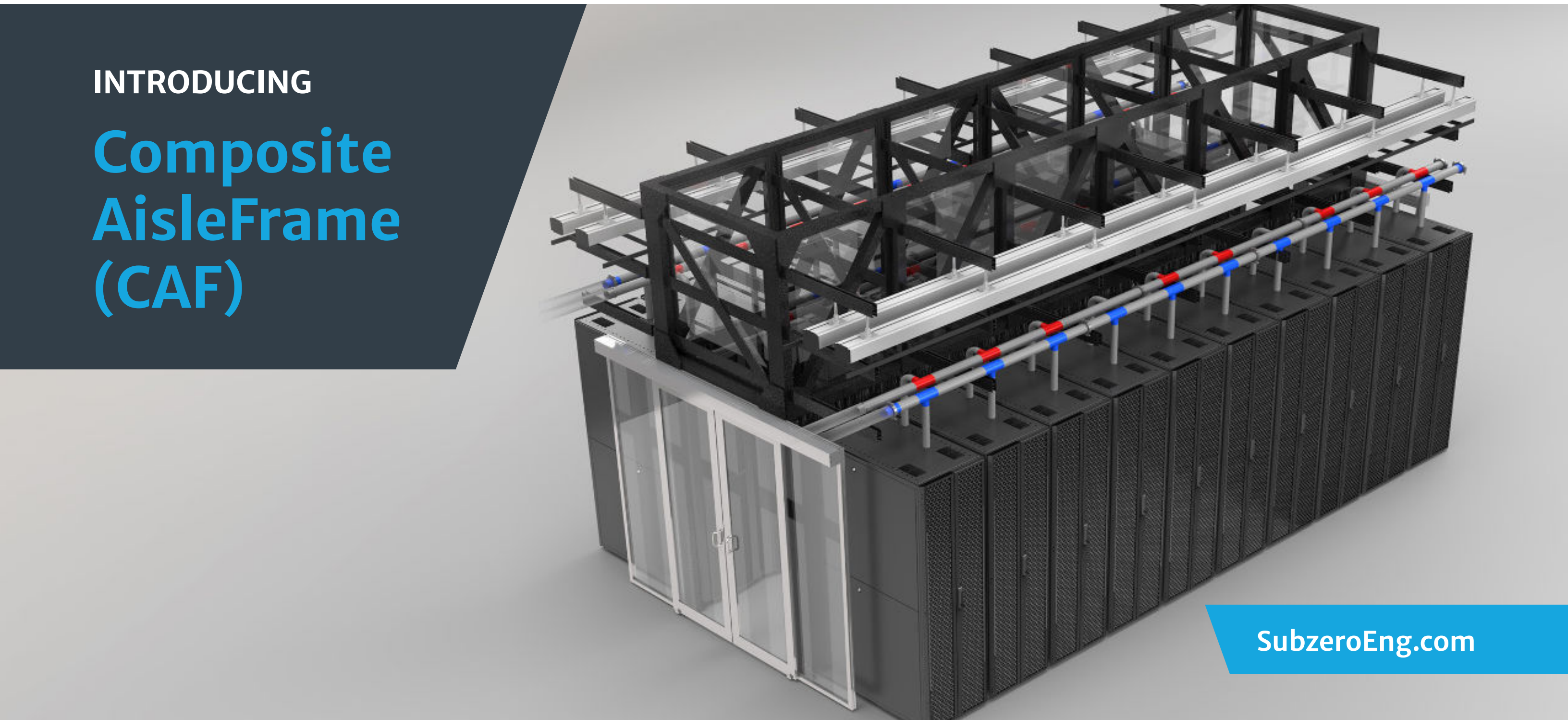
To support DLC, data centers need a specialized infrastructure that goes beyond traditional setups. Subzero Engineering's new Composite AisleFrame (CAF) system provides a frame-based support structure for HPC deployments requiring DLC for environmentally conscious data centers. Made of composite materials, CAF helps reduce environmental impact and promotes a greener, more eco-friendly future, from production throughout its lifecycle.

The CAF framework is every bit as durable and robust as Subzero Engineering's Steel AisleFrame (SAF) system, but due to its lightweight materials (50% lighter than steel) it's easier to transport, install, and configure/reconfigure in data centers. CAF is an excellent green alternative to steel systems with an overall lower Global Warming Potential (GWP), allowing data centers to now take environmental efforts to the actual building infrastructure. CAF also offers greater flexibility, easier scalability, and significant cost savings which result in lower TCO.

DID YOU KNOW?

Subzero Engineering's Composite AisleFrame (CAF) reduces environmental impact with a lighter, greener alternative to steel while supporting scalable HPC deployments.

CAF is the perfect ground supported infrastructure for both brownfield and greenfield sites that can be easily scaled up as HPC deployment increases. CAF combines the essential HPC deployment elements such as cabinet docking, containment, and mounted power delivery including cable trays and fiber runners in an all-in-one framework system.



Subzero Engineering's CAF eco-friendly attributes make it an ideal choice for environmentally conscious data centers looking to reduce their environmental footprint. As the data center industry continues to embrace the principles of sustainability, Subzero Engineering leads the way by playing a vital role in shaping a greener and more sustainable future by protecting our planet and preserving its resources. **(See CAF AisleFrame photo)**

Follow the Path to Sustainability with Subzero Engineering

The future of data center cooling has arrived, and its hybrid cooling. Hybrid cooling combines the strengths and benefits of both air cooling and liquid cooling, helping data centers and organizations achieve a balanced approach to efficiency, sustainability, and TCO. It's no longer a "one size fits all approach" when it comes to cooling. Just as HPC to support AI is no longer a trend but is becoming the new normal, we can expect hybrid cooling moving forward to also be the new normal when it comes to how we cool data centers.

This rapid change has also resulted in a lack of DLC guidelines and standards. These standards, while emerging, are not as mature and complete when compared to air cooling. While this will eventually change (hopefully sooner than later), progress at times with new technology gets held back by lack of transparency, with some in the industry fearing that sharing will give their competitors an edge.

This highlights the importance of working closely with an experienced partner and technology vendor, and one that prides itself on transparency such as Subzero Engineering, for current and future data center designs. Whether it's cold or hot aisle containment (including a complete frame-based ground supported infrastructure for DLC), CFD modeling, embracing innovative solutions by using greener materials in the white space, or 20 years of data center design experience, Subzero Engineering can help future-proof any data center while lowering overall TCO and achieving both sustainability goals and lowering its carbon footprint.

Subzero Engineering can lead you down the road to creating a net zero carbon data center. Scan the QR code to learn more.

About the Author

Gordon Johnson is the Senior CFD Engineer at Subzero Engineering and is responsible for planning and managing all CFD related jobs in the U.S. and worldwide. He has over 30 years of experience in the data center industry which includes data center efficiency assessments, both air and liquid cooling data center design, CFD modeling, and disaster recovery. He is a certified U.S. Department of Energy Data Center Energy Practitioner (DCEP), a certified Data Centre Design Professional (CDCDP), and holds a Bachelor of Science in Electrical Engineering from New Jersey Institute of Technology. Gordon also brings his knowledge and ability to teach the fundamentals of data center energy efficiency to numerous public speaking events annually, white papers, and industry leading research articles globally.

V 01—09.25.2025