DATA CENTER Frontier Special Report

Sustainably Meeting High Density Cooling Challenges

When, Where, and How

Written by Julius Neudorfer, DCEP

Photo: Timofeev Vladimir/Shutterstock.com

brought to you by

NAUTILUS® DATA TECHNOLOGIES

Contents

Introduction

Data centers have been seeing a continuous increase in IT equipment power densities over the past 20 years. The past five years have brought a near exponential rise in power requirements of CPU, GPU, other processors, and other ITE components such as memory. Managing this intensified heat load has become more challenging for IT equipment manufacturers, as well as data center designers and operators.

The power density for mainstream off-the-shelf 1U servers with multi-processors now typically range from 300 to 500 watts (some models can reach 1000 watts). When stacked 40 per cabinet, they can demand 12-20 kW. The same is true for racks loaded with multiple bladeservers.

Cooling at this power density has already proven nearly impossible for older facilities and is challenging to some data centers designed and built only five years ago. Even many newer data centers can only accommodate some cabinets at this density level by various workarounds, and they have realized that this impacts their cooling energy efficiency. Moreover, the demand for more powerful computing for artificial intelligence (AI) and machine learning (ML) will continue to drive power and density levels higher. Processor manufacturers have product roadmaps for CPUs and GPUs expected to exceed 500 watts per processor in the next few years.

The world is trying to mitigate climate change by addressing core sustainability issues. For data centers, energy efficiency is an important element of sustainability; however, energy usage is not the only factor. Today, many data centers use a significant amount of water for cooling.

This paper will examine the issues and potential solutions to efficiently and sustainably supporting High Density Cooling while reducing energy usage and minimizing or eliminating water consumption.

IT Equipment Heat Removal Challenges

IT Equipment Thermal Management

What is thermal management and how is it different than cooling (free or otherwise)? While it may seem like semantics, there is an important difference between a design approach and a technical level. Generally speaking, we have traditionally "cooled" the data center by means of so-called "mechanical" cooling. This process requires energy to drive a motor for the mechanical compressor which drives the system (in reality, it is a "heat pump" since it transfers the heat from one side of the system to another). Getting the heat from the chip to the external heat rejection is the key to end-to-end thermal management effectiveness and energy efficiency.

Traditional mainstream data centers use air-cooled IT equipment (ITE). However, the power density of IT equipment has risen so significantly that it has become more difficult to effectively and efficiently cool IT equipment beyond 20 kW per cabinet using traditional perimeter cooling systems. While improved airflow management, such as cold or hot aisle containment systems, has helped improve the effectiveness of the IT thermal management within the whitespace, it still requires a significant amount of fan energy for the facility cooling units and the ITE internal fans. There are also close coupled cooling systems, such as rear-door heat exchangers and row-based cooling units, which can support higher power densities more effectively.

Air Cooling of Whitespace

Although IT equipment has continuously improved its overall energy efficiency (i.e., power consumed vs. performance), the total power draw has increased tremendously. This has resulted in a rise in average watts per square foot in the whitespace, going from under 100 watts per square foot to 200-300 W/Sf or even higher, for mainstream data centers being designed and built today. While this average power density can be cooled using conditional methods, such as raised floor with perimeter cooling units, it becomes a greater challenge every year.

The bigger challenge starts at the processor level and moves through the heat transfer process within the IT equipment and eventually impacts the rack power density. The thermal design power (TDP) of processors (CPUs, GPUs, TPUs, and other upcoming devices) and many other devices such as memory have increased significantly over the past decade. Today even the CPUs using air-cooled heat-sinks in low profile commodity and mid-level servers can range 100-150 watts each, but most have difficulty moving up to the 200 W per processor level. This has resulted in significant and unceasing increase of power density of the individual IT equipment, as well as the power density per rack resulting in an overall rise in watts per square foot in the whitespace.

As noted in the introduction, the power density for mainstream off-the-shelf 1U servers with multi-processors now typically ranges from 300 to 500 watts (some models can reach 1000 watts). When stacked 40 per cabinet, they can demand 12-20 kW. The same is true for racks loaded with multiple bladeservers.

Understanding Airflow Physics

The nature of the challenge begins with the basic physics of using air as the medium of heat removal. The traditional cooling unit is designed to operate based on approximately 20˚F differential of the air entering the unit and leaving the cooling unit (i.e., delta-t or ΔT). However, modern IT equipment has a highly variable delta-t dependent on its operating conditions as well as its computing load. This means that ΔT may vary from 10˚F to 40˚F during normal operations. This in itself creates airflow management issues resulting in hotspots for many data centers that are not designed to accommodate this wide range of varying temperature differentials. It also limits the power density per rack. There are solutions using various forms of containment that have been applied to try to minimize or mitigate this issue. Ideally, this is accomplished by providing closer coupling between the IT equipment and the cooling units.

The most common disconnect is not just the delta-t but its companion; the rate of airflow required per kilowatt of heat. This is expressed by the basic formula for airflow (BTU=CFM x 1.08 x ΔT ˚F), which in effect defines the inverse relationship between ΔT and required airflow for a given unit of heat. For example, it takes 158 CFM at a ΔT of 20˚F to transfer one kilowatt of heat. Conversely, it takes twice that amount of airflow (316 CFM) at 10˚F ΔT. This is

considered a relatively low ΔT, which increases the overall facility fan energy required to cool the rack (increasing PUE). It also increases the IT internal energy (which increases the IT load without any computing work—thus artificially "improving" facility PUE). This also limits the power density per rack.

Airflow per kilowatt of heat transferred*

Delta-T (°F)	Required CFM
10	316
15	210
20	158
25	126
30	105
40	79

**Note: For purposes of these examples, we have simplified the issues related to dry-bulb vs. wet-bulb temperatures and latent vs. sensible cooling loads.*

To overcome these issues ITE manufacturers of higher density servers, such as bladeservers designed to operate at a higher ΔT whenever possible. This allows them to save IT fan energy, but it also can create higher return temperatures for the cooling units. For most cooling units (CRAH) fed from chilled water, this is not an issue; in fact, it is beneficial since it improves the heat transfer to the cooling coil for a given airflow. However, for other types of cooling units, such as a Direct Expansion (DX) Computer Room Air Conditioner (CRAC), which use internal refrigerant compressors, these higher return temperatures can become a problem and stress the compressor beyond the specified maximum return temperatures.

Liquid Cooling

Liquid Cooling (LC) has become a bit of a catchall term that encompasses many different technologies and methodologies to transfer heat from the IT equipment to the environment. While this is a highly complex subject, for purposes of this discussion it is simplified into several major categories, described in the table below.

For additional details and references for liquid cooling see: The Green Grid whitepaper WP#70 Liquid-Cooling-Technology-Update.

An ASHRAE 2021 whitepaper, *Emergence and Expansion of Liquid Cooling in Mainstream Data Centers*, focuses on the need for data centers to accommodate and adapt to rising power densities.

It also notes that while higher power processors are the primary thermal management challenge, higher density power levels due to an increase of memory DIMMs in new servers add to the load. The whitepaper discusses a CPU roadmap moving from 120-200 watts to 600 watts per processor (CPU, GPU, etc.). This whitepaper reports these limitations and notes that "A fan power percentage of 10% to 20% is not uncommon for some of the denser air-cooled servers."

The Green Grid Metric – Power Utilization Effectiveness

The Green Grid (TGG) created the Power Utilization Effectiveness (PUE) metric in 2007. Since then, it has the most globally recognized facility efficiency

Liquid Cooling

Type	Description	Examples	Power Densities (kW per cabinet)*
Liquid cooling systems for air-cooled ITE	Close-coupled cooling for standard air-cooled ITE (also called supplemental cooling). The fluids going through the cooling equipment can be water, glycol mixture, or refrigerants.	Inrow Cooling, Rear Door Cooling, Overhead Cooling Liquid Cooled Cabinets	15-50 kW Standard 42U IT Cabinet 24" x 42"
Liquid cooled IT equipment	The IT equipment has been modified or manufactured to use liquid entering and leaving the IT equipment. Some or most of the heat is transferred to the liquid; the rest of the heat is rejected to the air.	Standard ITE that have been modified or manufactured with LC cold plates which replace heat sinks on CPUs, and in some cases, cold plates are also added to memory	20-100 kW Standard 42U IT Cabinet 24" x 42"
Immersion cooling	The IT equipment has been modified or manufactured to be fully submersed in a dielectric fluid. Virtually 100% of the heat is transferred to the liquid.	Various form factors Immersion cooling cabinet is a typically horizontal tub style enclosure.	100-250 kW Note: Typical horizontal tub cabinets approx. 30" x 80",

**Note: the power densities ranges per cabinet are not absolute but are representative of currently available solutions.*

metric that helped drive down PUE for "traditional" enterprise and colocation MTDC data centers (from approximately 2.0 in 2010 to 1.4 or less in 2020). In 2016 the International Standards Organization (ISO) adopted PUE and other TGG metrics as Key Performance Indicators (listed ISO/IEC 30134 KPI series). When examining PUE as an energy efficiency comparison yardstick, it is one-dimensional. Nonetheless, its underlying simplicity allowed data center operators to easily calculate (or estimate) a facility's PUE, which drove its widespread adoption.

Moving beyond PUE

For many years, the data center ecosphere quoted and referenced PUE as the primary element of their "green" mantra. As important as PUE was in raising energy awareness, it is only a part of more in-depth conversations about data center sustainability. The current focus is now on decarbonizing and moving to 100% renewable energy sources. While renewable energy is clearly an important factor, many data centers use water as part of the external heat rejection process, regardless of the energy source.

External Heat Rejection Systems

We have been discussing the differences between air cooling and liquid cooling of IT equipment within the data center, which has recently gotten the most attention due to rising power densities. However, the various types of heat rejection systems also have a significant impact on the energy efficiency as well as the sustainability of the data center facility.

Generally speaking, the most common methods of mainstream data center external heat rejection systems fall into categories listed in the table below.

Water is a critical global sustainability resource. According to the US EPA, at least 40 states are anticipating water shortages by 2024, and consider the need to conserve water is critical.

Water Usage Effectiveness

The Green Grid developed the Water Usage Effectiveness (WUE) metric in 2011. While many data center operators measure and report their PUEs publicly, very few disclose their water usage. For data centers, the consumption of water is not taken into consideration in the PUE metric. When discussing water consumption and WUE, it is important to note two types of WUE calculations: WUE site and WUE source.

WUE Site – Based on the water used by the data center; expressed as the annualized liters of water used per kWh of IT Energy (liters/IT kWh). This is the most commonly used reference and is easily measured.

WUE Source – Based on the annualized water used by the energy source (power generation), plus the site water usage, expressed as liters of water used per kWh of IT Energy. (liters/IT kWh).

Many people are unaware of the two types of WUE, and in most cases, data centers that cite their WUE are only referencing their site.

Type	Description	Energy Efficiency	Water Usage	Performance Impact by Climate Zones
Direct evaporative, Open loop	Condenser water from a chiller is sent directly to an evaporative cooling tower.	High	High	Performance reduced by high humidity.
Indirect evaporative, Closed loop	Glycol from closed-loop chiller condenser is sent to a coil in the bottom of cooling tower. Separately, water is evaporated in the cooling tower to cool glycol loop.	High (but slightly lower than direct evaporative)	High	Performance reduced by high humidity similar to direct evaporative.
Fluid cooler	Closed loop Water-Glycol mixture from fluid cooled DX CRAC	Lower due to higher fan energy	None	Performance and capacity impacted as temperatures increase (max temperature limits). Not impacted by humidity.
Condenser DX refrigerant	Closed loop refrigerant from DX CRAC.	Lower due to higher fan energy	None	Performance and capacity impacted as temperatures increase (max temperature limits). Not impacted by humidity.
Adiabatic cooling added to fluid cooler or DX condenser	Water misting system is added to fluid cooler or condenser. Operates during hotter weather.	Improves efficiency	Low to moderate	Improves performance and capacity as temperatures increase. High humidity reduces improvement.
Water-Cooled	Naturally cold water runs through an open and closed loop system. Operates near any body of water.	High	None	Performance not impacted by ambient temperatures. Not impacted by humidity.

Generally speaking, evaporating more water during heat rejection will normally reduce the energy used by mechanical cooling systems. However, in order to put this into perspective, the type of heat rejection used is a design decision, which is based on various factors. Some of those decisions are based on a business perspective, based on the cost of energy vs. the availability and cost of water. From a technical perspective, climate has a significant impact on water usage.

Moreover, large-scale and hyperscale data centers often use evaporative cooling solutions, and a significant number of sites are located in water-stressed areas. Multiple publications and sources have noted this.

According to a Virginia Tech June 2021 white paper, *The environmental footprint of data centers in the United States*, "one-fifth of data center servers direct water footprint comes from moderately to highly water-stressed watersheds, while nearly half of servers are fully or partially powered by power plants located within water-stressed regions."

According to a NASA Earth Observatory March 2021 post: "Almost half of the United States is currently experiencing some level of drought, and it is expected to worsen in upcoming months."

Climate Factors on Energy Efficiency and Water Usage

The effectiveness of evaporative cooling is based on climatic conditions. While this is well known, it was not taken into account in the WUE metric. However, this has been partially addressed in the ASHRAE 90.4 Energy Standard for Data Centers, which in the US has a table of mechanical energy adjustment factors based on the climate zone where the data center is located. Nevertheless, ASHRAE 90.4 does not account for water usage in its energy calculations.

While it is relatively easy to measure site water usage, the WUE source metric is more complex but helps provide a more holistic view of overall sustainability.

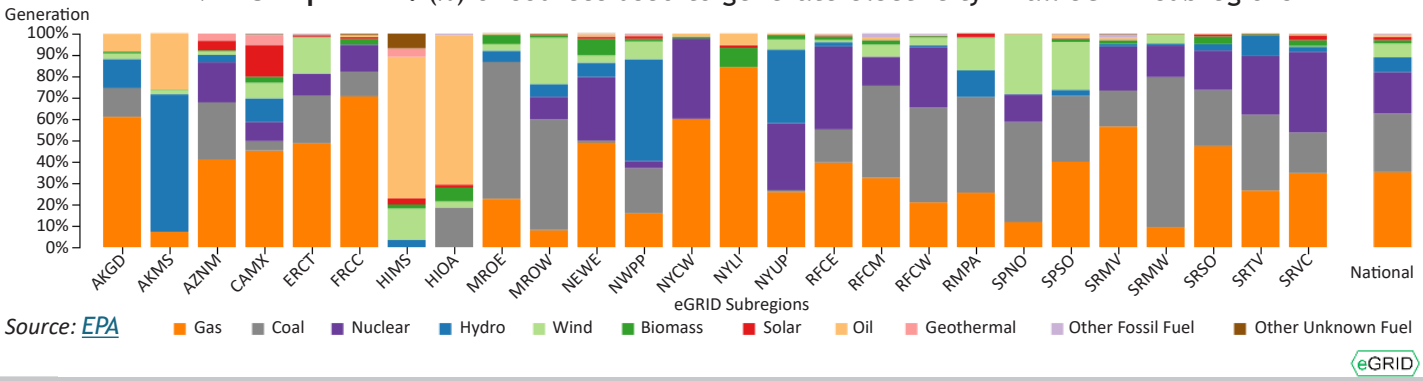
However, power production water usage varies widely based on the fuel source. It, therefore, requires a clear understanding of the type of energy source that is actually supplied to a data center, which can differ by location. While more and more data centers are committed to using renewable energy, calculating the WUE source requires knowing the water consumption of the actual source used to power the data center. This information can sometimes be obtained from their energy provider; however, the validity of this information can be muddied when the energy is acquired via a power purchase agreement (PPA) or especially via a Renewable PPA (RPPA). In most cases, a RPPA is simply a paper transaction rather than the actual renewable energy source. PPAs and Renewable Energy Certificates (RECs) are used by many organizations and some data centers.

For reference, the U.S. Energy Information Administration (EIA) recognizes and tracks water usage by power plant generation.

Water Treatment Chemicals Impact Sustainability

It is important to note that in addition to the water used by evaporation, cooling towers must be properly maintained and cleaned for operational efficiency, as well as for health safety to prevent legionella growth. This requires water treatment chemicals that have a negative impact on environmental sustainability. The amount of chemicals used increases with water usage and the energy consumption of the data center. The use of water treatment chemicals has a double impact on sustainability, both by the amount used by the facility and the chemical waste and water used in the manufacturing processes of the water treatment chemicals.

Fuel Mix Comparison: (%) of sources used to generate electricity in all eGRID subregions



Source: EPA

eGRID

Carbon Usage Effectiveness

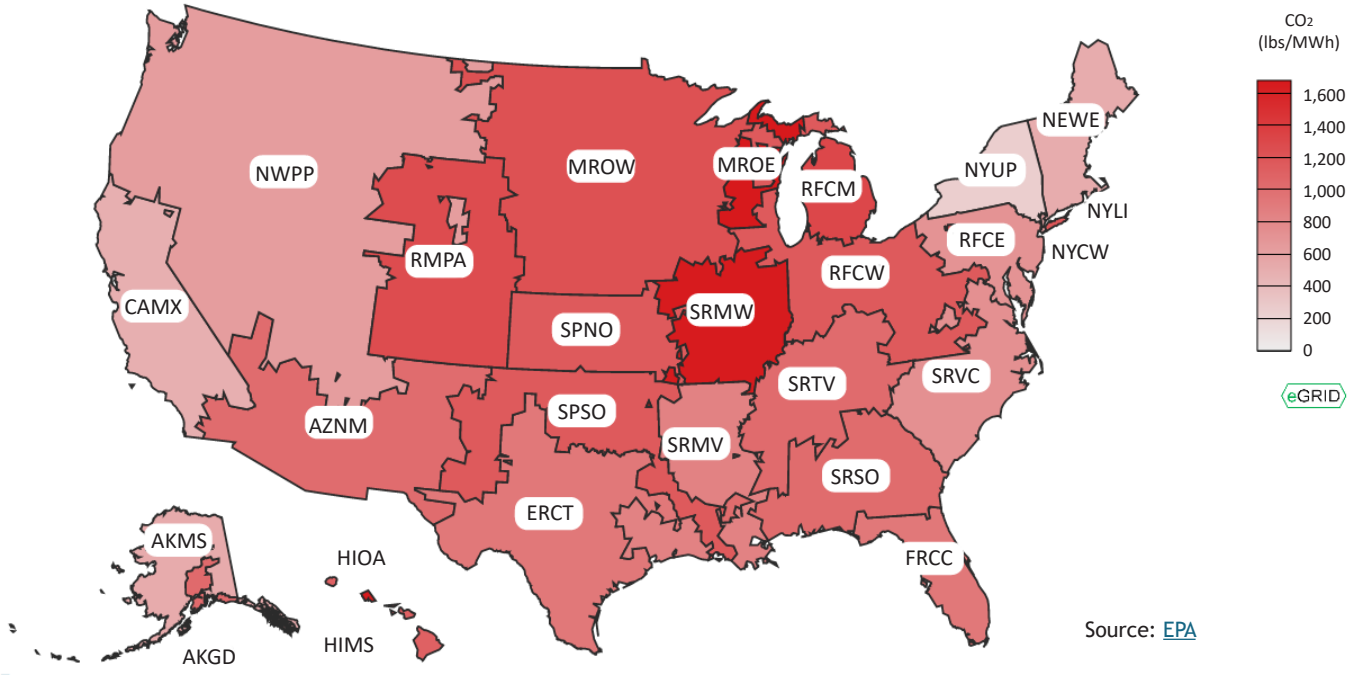
Carbon Usage Effectiveness (CUE) is another essential sustainability metric directly tied to the source of energy generation.

It is expressed as CUE Source – Based on the annualized carbon emitted by the energy source (power generation), expressed as kilograms of carbon emitted per kWh of IT Energy generated. (kg/IT kWh).

The EPA's Emissions & Generation Resource Integrated Database (eGRID) website has a tool that allows users to enter their zip codes (or select a region) to view their power profiles.

Emission Rate Map

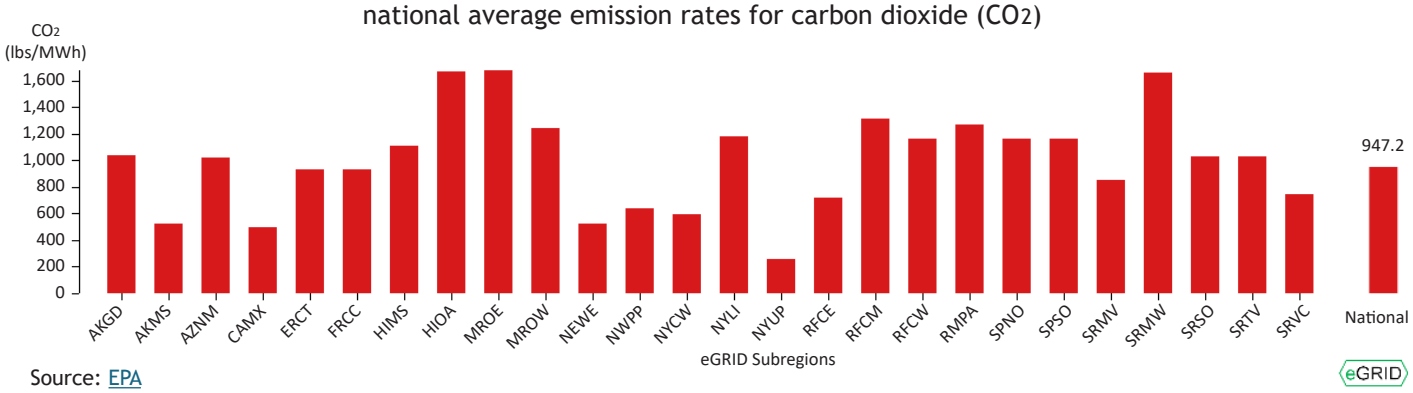
Provides eGRID subregion average emission rates in pounds per MWh



Source: EPA

Emission Rates Map

Provides the average emission rates in pounds per MWh in all eGRID subregions to the national average emission rates for carbon dioxide (CO_2)



Source: EPA

Free Cooling

Free cooling is typically referred to as any type of cooling system which reduces or eliminates the need for mechanical cooling (i.e., compressor based). This can include direct or indirect air-side economizers, as well as water-side economizers. The water-side economizer is generally based on a heat exchanger which allows cooling tower water to reduce the load on water-cooled chillers during cooler weather. All of these systems generally save a percentage of mechanical cooling energy during cooler weather. Overall, free cooling is more effective in colder climates, for both air-side and water-side economizers. There are many variations and combinations of methodologies that can be combined to maximize energy effectiveness and optimize water usage over the seasons.

While water in itself is not energy, it requires energy to process and deliver clean water. This is often overlooked or ignored. However, as water shortages increase, this is being addressed in the California Energy Code: Codes and Standards Enhancement (CASE) Initiative 2022 – "Title 24 2022 Nonresidential Computer Room Efficiency" (CA-Title 24), as "Embedded Electricity in Water."

Alternately, waste heat can also be rejected into bodies of water, such as lakes and rivers or even the ocean. In many cases, the temperature ranges of those bodies of water are such that they can be used all year round to cool data centers without the need for mechanical cooling, a significant energy savings. These savings can also reduce or eliminate consumption of source water, as well as reducing fossil fuels used in non-renewable power generation.

Energy Reuse Effectiveness

The total waste heat generated by an air-cooled or a liquid-cooled data center is virtually the same for a given amount of total energy consumed. Clearly, a lower PUE will reduce the total energy consumed for a given IT load. While end-to-end thermal management typically refers to "chip-to-atmosphere" heat rejection, it is still just waste heat. One of the long-term sustainability goals is being able to recover and reuse waste heat.

The Green Grid introduced the energy reuse effectiveness (ERE) metric in 2011, energy reuse factor (ERF), with relatively little impact. This was primarily because waste heat from air-cooled ITE being rejected is more difficult to recover effectively.

Liquid Cooling provides a pathway to more effective opportunities for energy reuse and recovery. Its higher fluid operating and return temperatures improve the ability to recover a portion of the waste heat energy. However the challenges are multi-dimensional, but as time progresses the interest, technology improvements and cost effectiveness will continue to drive this initiative. While higher power densities can create cooling challenges, it also offers multiple benefits and opportunities to improve overall energy efficiency, IT performance, and sustainability. However, the challenges are multi-dimensional, but the interest, technology improvements, and cost-effectiveness will continue to drive this initiative as time progresses.

Higher Density Supports Space Reduction Benefits

The ability to increase power density from 5-10 kW per rack to 25-100 kW per rack offers the opportunity to substantially reduce the facility's size. This in turn results in far less material being used by the facility and reduced waste material from construction. This also reduces the number of racks and power distribution equipment, further lowering the material consumption.

Higher Density Supports Higher IT Performance

Moreover, high-density computing also benefits the overall performance of computing systems. Reducing the distance between IT and network equipment results in lower latency, which is attributable both to shorter intersystem connectivity distances within the rack and improved overall multisystem throughput due to reduced distances for rack-to-rack and rack-to-core connectivity. This is especially true for High Performance Computing (HPC) and Artificial Intelligence (AI) workloads.

The Bottom Line

We are at a pivotal moment in the sustainability roles that data centers play as part of the critical infrastructure and the digital economy. While there is no question that mainstream data centers will not suddenly become functionally obsolete in the next few years, clearly the recognition that ignoring all the elements of sustainability is no longer an option for new designs.

The Nautilus water-cooled data center provides an embodiment of several thermal management technologies which can be applied in whole or in part on water or land. It is more than just a proof concept; it is a fully operational colocation facility capable of effectively and efficiently delivering high-density cooling at over 800 watts per square foot and 50-100 kW per cabinet without consuming any water.

It took several years after the advent of the PUE metric in 2007 before improving energy efficiency became part of the priorities for data center operation and new designs. And almost ten years before it became a primary concern for many organizations with social responsibility ethos. This decade has shown that the data center industry and its users have moved to the next level of awareness to face the responsibility of long-term sustainable design and operation in a holistic manner to maximize the use of resources and performance advantages for a given geography and climatic conditions.

When deciding when and where to consider investing in building a new facility, data center operators make decisions based on business and customer market demands. In addition, the designers and engineers have to balance the factors related to location climate zones, cost, and availability of sufficient energy and other resources such as water, before committing to building it.

Major data center operators and hyperscalers, as well as the industry, have sustainability initiatives. Nonetheless, as a business, they may need to build in areas that may not be ideal from a sustainability viewpoint, but offers a market opportunity and near-term return on investment. However, investors and financial professionals worldwide are increasingly factoring in environmental, social, and governance (ESG) risks and opportunities to their investment decision-making process, reflecting a view that companies with sustainable practices may generate stronger returns in the long term.

The Nautilus water-cooled data center... is a fully operational colocation facility capable of effectively and efficiently delivering high-density cooling at over 800 watts per square foot and 50-100 kW per cabinet without consuming any water.

Nautilus will begin using its technology to take advantage of the opportunity to reuse a portion of an abandoned paper mill in a sustainable manner. In June 2021, Nautilus announced its land-based facility that takes advantage of the topography in Maine that taps a reservoir to create a gravity-fed cooling system that not only avoids the cost of energy use for mechanical cooling and it also reduces the pumping energy.

Moreover, it is also able to use energy from an existing hydroelectric power plant operated on the site. The combination of zero-emission supply and the data center's energy efficient high density cooling substantially enhances the GHG benefit of the zero-emissions power procurement. In addition, future prospects of beneficial uses of its warmed discharge water will produce further improvements of the overall sustainability profile for the project's utilization of the repurposed site.

The Nautilus data center in Stockton confirmed the technology processes and performance. Conceivably, this first land-based project will become a reference design model. Moreover, it will hopefully incentivize future high-density data centers, which can strategically be located to efficiently and sustainably optimize natural resources with minimal impact on the surrounding ecosystems.