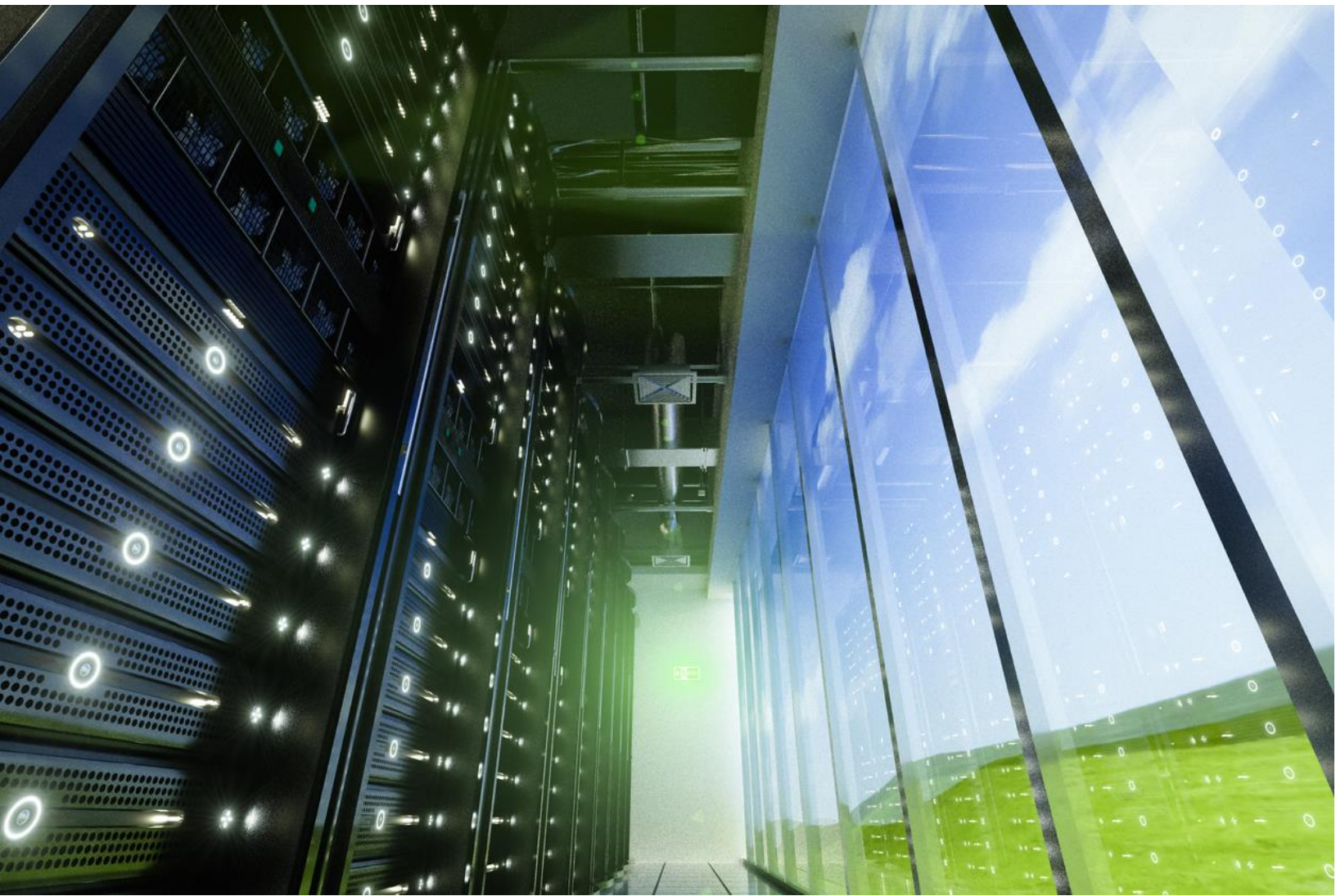
DELL'ORO GROUP

WHITEPAPER | **October 2022**

ACHIEVING SUSTAINABLE DATA CENTER GROWTH

Authors:
Baron Fung, Research Director for Data Center IT Capex
Lucas Beran, Principal Analyst for Data Center Physical Infrastructure



Abstract

The global COVID-19 pandemic has had a significant impact on the data center industry, accelerating years of digital transformation into a matter of months. This has been driven by behavioral shifts in businesses and consumers, including a shift to remote and hybrid work, automation of business processes and increased reliance on the digital economy. The pace and success of this digital transformation has garnered significant attention for the data center industry, as a broader audience began to understand the mission critical nature and the value that data centers brought to their lives.

It has also, however, called attention to the growing energy and water use of data centers and the resulting impact on environmental sustainability. For this reason, sustainability quickly became the buzzword in the data center industry, as industry stakeholders began educating customers, regulators, and investors on the opportunity and challenges associated with achieving sustainable data center growth.

This paper discusses the key trends and the five-year outlook for the data center industry. We explore the drivers for sustainable data center growth, the technologies to support it, and the current metrics being used to evaluate and track sustainability progress in the data center ecosystem. Finally, we share some of the current best practices and use cases for realizing end-to-end sustainable data center planning and design, construction, management, and optimization throughout a data center's entire life cycle.

Contents

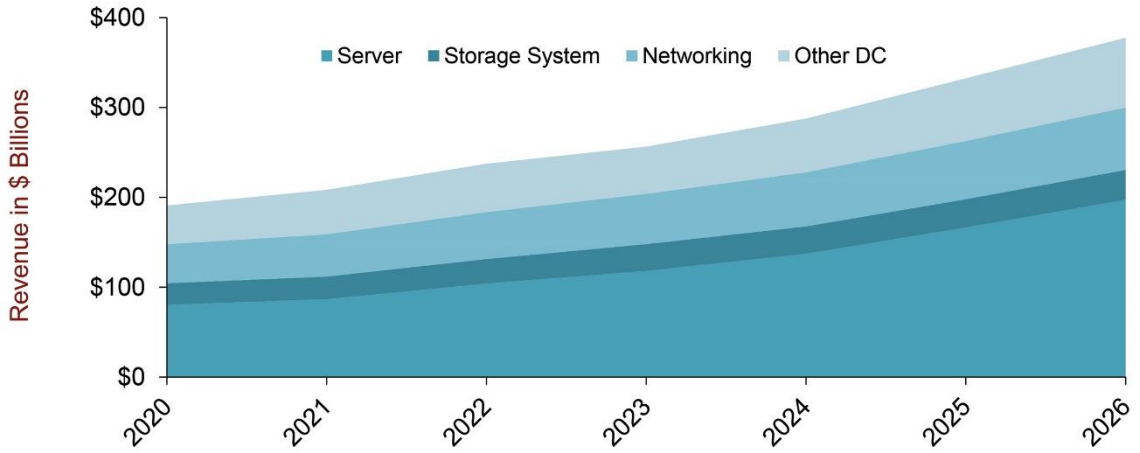
Introduction

Data centers today are estimated to constitute nearly 2% of global energy consumption. The decade-old fear of runaway growth in data center energy consumption hasn't occurred. Hyperscale cloud service providers (CSPs) have largely managed that concern, with the help of industry vendors, through IT virtualization and higher utilization of power and thermal management infrastructure. At the same time, enterprises, usually less efficient than CSPs at managing data centers, have transitioned many of their workloads and applications to the cloud. Yet, as the global COVID-19 pandemic accelerated digital transformation, data center demand has accelerated. This has rekindled fears as to where future data center energy consumption is headed, with customers of data center service providers, regulators, and investors demanding environmentally sustainable growth from the data center industry.

Data Center CapEx: Sustainability Requirements Must Evolve with IT Infrastructure

Worldwide capital expenditure (CapEx) on data center infrastructure is forecast to increase at a 13% compound annual growth rate (CAGR) over the forecast period of 2021–2026, to $377 B (Figure 1).

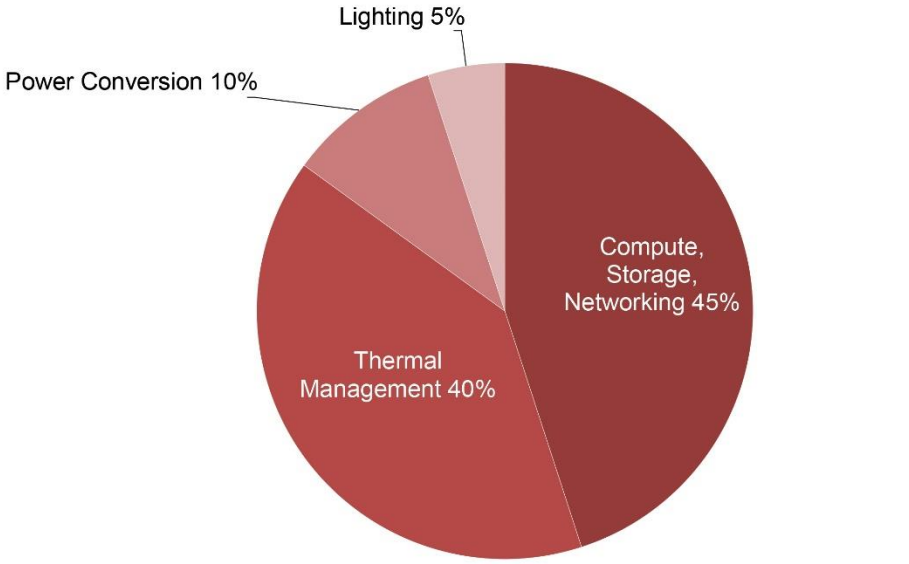
Figure 1: Data Center Capex Forecast



Source: Dell'Oro Group

Data center CapEx includes IT infrastructure (such as servers, storage systems, and networking infrastructure) and physical infrastructure (such as backup power, power distribution, and thermal management). We forecast that this growth will be driven by the ongoing deployments of new data centers by cloud, colocation and telco service providers, and enterprise data center modernization. To embed data center sustainability in this industry growth, it's important to understand how data centers consume energy. While energy use can vary based on a data center's location and the specific infrastructure deployed, a typical data center's energy consumption is driven by the IT infrastructure (compute, networking, and storage) and software, which accounts for 45% of total energy consumption (Figure 2). Data center physical infrastructure accounts for the remaining 55% of total energy consumption, with 40% consumed by thermal management systems, 10% by power conversions, and the remaining 5% by lighting.

Figure 2: Data Center Infrastructure Energy Consumption (2021)

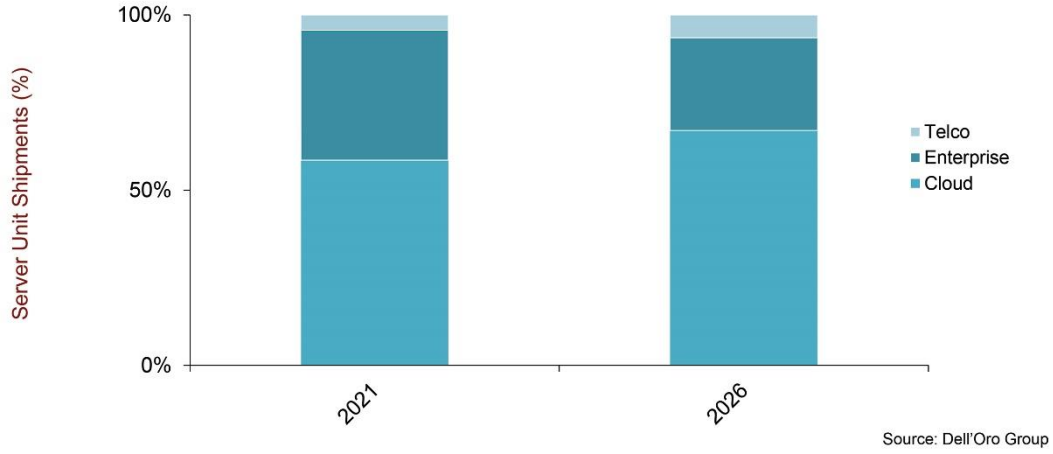


Source: Dell'Oro Group

There are opportunities to increase efficiencies and improve sustainability in both IT infrastructure and physical infrastructure based on today's best practices and technology. Additionally, with fast-paced innovation in the data center industry, sustainability must evolve with the technology transitions that will inevitably occur. For IT infrastructure, the continued transition from enterprise-owned data centers to cloud services, the proliferation of edge computing, and the increased adoption of accelerated computing all have significant implications for data center sustainability, as explained below.

First, cloud momentum has been accelerating. We estimate that CSPs accounted for 59% of global server shipments in 2021, a figure that we expect to increase to 67% by 2026 (Figure 3).

Figure 3: Server Shipments by Customer Segment



The economics of cloud computing may be appealing, especially as enterprises are increasingly preferring an operating expense (OPEX), consumption-oriented usage model—as opposed to a CapEx model—to satisfy growing needs for IT equipment and resources, while preserving capital. As workloads continue to shift to the cloud, we expect servers and its infrastructure to be consolidated in fewer mega cloud data centers that could provide greater capacity and operational efficiency than the same number of servers spread out across thousands of enterprise data centers. In particular, we predict that the cloud service providers, with their ability to scale complex and costly infrastructure, will be well-positioned to host AI and ML applications for enterprises. Furthermore, these CSPs must contend with stringent environmental regulations and adhere to corporate social responsibility requirements. This gives them strong incentives to deploy infrastructure that is more energy efficient and sustainable when operating these power-hungry platforms.

Next, edge computing applications—such as cloud gaming, autonomous driving, and industrial automation—are latency-sensitive, requiring edge data centers to be situated at the network edge, where sensors are located.

While these edge applications have yet to see widespread adoption, we project that server shipments deployed at the Telco and Enterprise edge will grow at a 74% CAGR during our forecast period, comprising 8% of the total server volume in 2026. As a result, we anticipate that large, centralized service provider data centers will be increasingly augmented with growing number of small data centers at the network edge. Sustainability becomes harder with more assets, in many locations, with new and unique environmental challenges. To achieve sustainable operations in edge computing deployments, monitoring and management software is the name of the game for ensuring uptime and optimized performance.

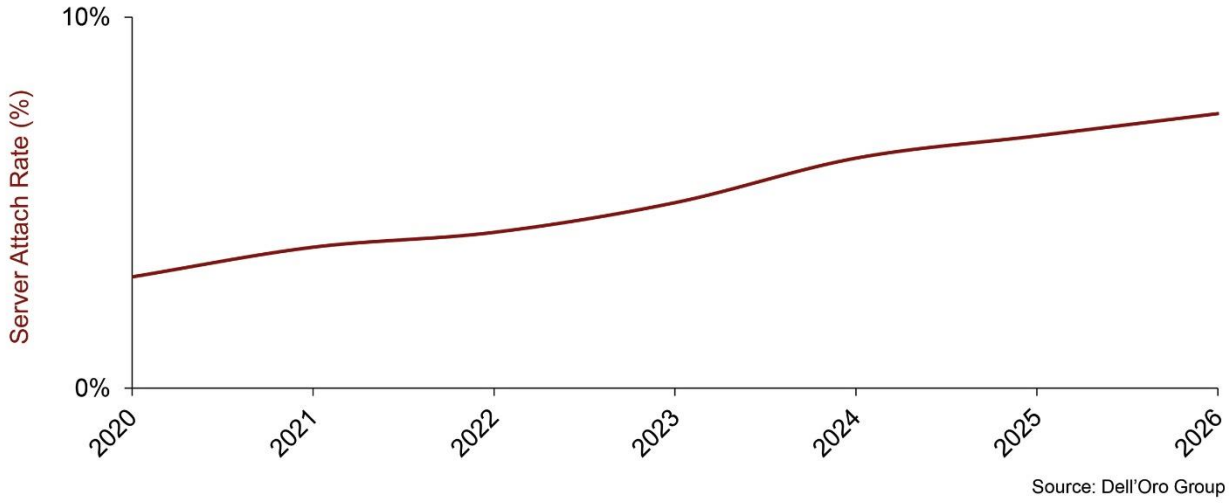
Lastly, IT demand and spending will also be affected by data center architecture changes, most notably from accelerated computing.

The proliferation of artificial intelligence (AI) and machine learning (ML) workloads has emerged as a disruptive force, enhancing applications such as image recognition, security, real-time text translation, autonomous driving, predictive analytics, and more recently, the metaverse. Accelerated servers typically have AI accelerator chips, such as GPUs, FPGAs, and custom processors that are tailored to accommodate high-end workloads. Some Cloud SPs, such as Alibaba, Amazon, Google, and Huawei, have deployed accelerated servers using internally developed AI chips, while other Cloud SPs and enterprises have commonly deployed accelerators from AMD, Intel, and NVIDIA. We estimate by 2026, 12% of the server unit shipments will be attached with these AI accelerator chips (Figure 4).

These accelerated servers consume significantly more power than a general-purpose server. For example, a NVIDIA DGX system with 8 A100 GPUs can consume up to 6.5kW, roughly a factor of 10 compared to a general-purpose server. However, because these AI accelerators are more efficient at processing select workloads, they consume less power in total than if those workloads were running on CPUs. Yet, with the higher power consumption of these accelerators, more heat is generated, requiring new thermal management, power distribution, and rack designs to support their deployment and sustainable management.

While these data center IT infrastructure trends will have an impact on data center sustainability, they can only move the needle so far on their own. There are more significant opportunities in data center physical infrastructure to support sustainable data center growth. This is because of the fact that data center physical infrastructure consumes just as much energy, if not more, than IT infrastructure.

Figure 4: Accelerated Server Adoption



Data Center Physical Infrastructure to Play a Significant Role in Data Center Sustainability

Data Center Physical Infrastructure (DCPI), sometimes referred to as facilities infrastructure, is often called the "backbone" of the data center, supporting the deployment and reliable operation of IT infrastructure (compute, networking, and storage). As tracked by Dell'Oro Group, DCPI includes critical power systems such as uninterruptible power supplies (UPSs), cabinet power distribution units (PDUs) and busway, and rack power distribution units (rPDUs), in addition to thermal management systems, IT racks and containment, and the software and service to manage and maintain DCPI hardware.

Data center physical infrastructure accounted for an estimated $22 billion in vendor revenues in 2021, forecast to grow to $32 billion by the end of 2026. This growth is forecast to be driven by expanding cloud, colocation, and telecommunication service provider data center footprints, with a particular emphasis on energy efficiency and sustainably minded infrastructure. While opportunities abound for improving sustainability related to DCPI, two specific trends—the utilization of liquid cooling in data center thermal management and grid-connected three-phase UPSs—represent significant opportunities. Best practices for the implementation of these technologies are still being determined for widespread data center use. Regardless, it remains critical to develop an understanding of these technologies to achieve long-term sustainability goals, as modern facilities may need to be retrofitted in order to support these technologies over the next 3–5 years.

Data Center Liquid Cooling Is the Thermal Management Technology at the Center of the Sustainability Conversation

Data center thermal management consumes 30–40% of a data center's annual energy consumption, second only to compute, making it the logical starting place. Today, data center thermal management technologies are primarily air-based, in the form of computer room air conditioners (CRACs) and computer room air handlers (CRAHs). Improvements in efficiency and adoption of free cooling and evaporative technologies have continued to help reduce power consumption associated with air-based thermal management technologies. In some cases, this has come at the expense of increasing water usage to reduce electricity consumption. However, with increasing power consumption of CPUs, GPUs, and an increasing attachment of accelerators to servers, rack power densities are on the rise. This is leading to the need to manage higher heat loads, where air has begun to reach the limits of what can be physically cooled.

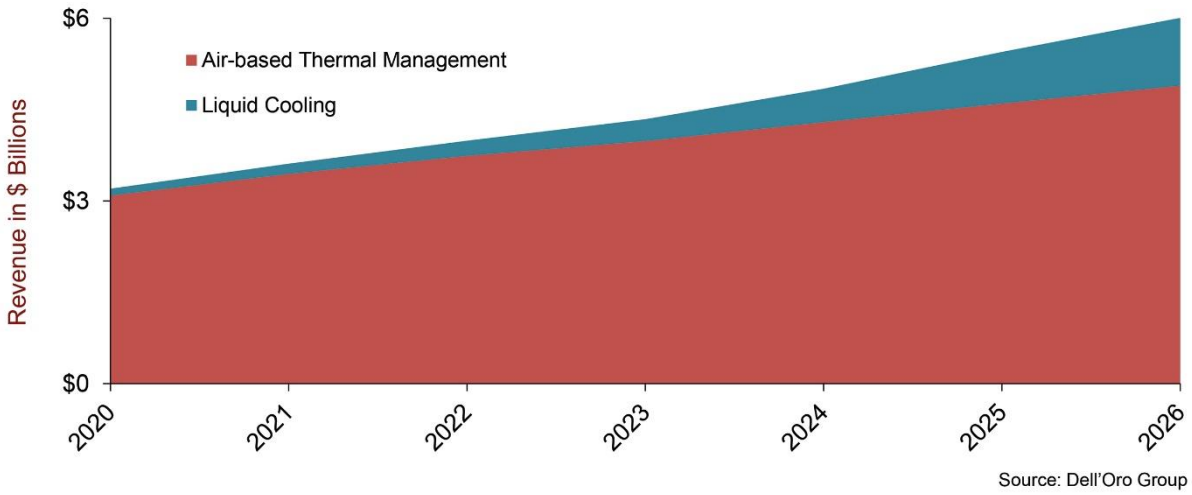
For this reason, the industry has begun to look at bringing liquids into the white space. This is materializing in the data center industry in two primary liquid-based cooling technologies. The first is direct liquid cooling (DLC), which involves piping a liquid to a cold plate directly attached to a server's CPU, GPU, and/or memory. DLC captures an estimated 80% of the heat generated by a server, which usually requires a DLC deployment to be augmented with air-based thermal management as well. The second form of liquid cooling used in data centers is immersion cooling. This is the process of submerging a specially designed server, or a slightly modified off-the-shelf server, in a tank of a specific fluid, to capture nearly 100% of the generated heat. Both of these liquid thermal management technologies operate at significantly improved energy efficiencies when compared with air-based thermal management systems, helping drive PUEs to as low as 1.03. However, the perceived added complexities of introducing liquid into the data center's white space has so far limited adoption in mission-critical data center environments.

Despite this limited adoption, both DLC and immersion cooling technologies have been in use for about a decade in high-performance computing (HPC) and, in some instances, computing deployments with unique requirements. Liquid cooling's track record in the HPC space and the increasing focus on sustainability constitute the driving force behind today's data center decision-making, with both DLC and immersion cooling viewed as potential paths to more sustainable data center operations in the future. Data center ecosystem partners from semiconductor manufacturers to server original equipment manufacturers (OEMs) are acknowledging the important role that liquid cooling will play in future data center growth and the designing of next-generation semiconductor components and servers with liquid thermal management in mind.

At the end of 2021, DLC and immersion cooling accounted for approximately 5% ($168 million) of total data center thermal management revenues. Dell'Oro Group forecasts the adoption of liquid cooling to continue and accelerate through 2026, reaching a combined $1.1 billion (19%) of thermal management revenues, in 2026 (Figure 5). We expect growth to be driven by the proliferation of high-density computing applications among enterprises and service providers,

while supporting improved energy efficiency in thermal management, achieving sustainability targets such as reduced water usage and enabling heat reuse.

Figure 5: Liquid Cooling Adoption



Grid-interactive UPS support Data Center Renewable Energy Integration and Grid Decarbonization

Another significant opportunity in DCPI related to sustainability is the centralized three-phase UPS system. For data centers, UPS systems provide basic power conditioning but more importantly, short-term backup power in case of mains (utility) power failure. This backup power is usually in the form of lead-acid and lithium-ion batteries. Today, these energy storage assets sit idle, waiting to be used in case of mains power failure. However, with the increasing adoption of lithium-ion batteries and the benefits that come with the battery chemistry (Figure 6), opportunities to proactively use these energy storage assets to support sustainable data center operations are developing.

Figure 6: VRLA vs. Lithium-ion Battery Comparison

	VRLA	Lithium-ion
Cost (CapEx and TCO)	Lower CapEx Higher TCO	1.5 – 2x VRLA CapEx Lower TCO
Battery lifespan (Battery replacements over UPS life)	3-5 years (1-2 replacements)	10-15 years (no replacements)
Operations	Stricter temperature requirements, slower recharge time, fewer battery cycles	Less sensitive temperature requirements, faster recharge time, more battery cycles
Sustainability	Higher Environment Impact (Scope 3 GHG Emissions)	Lower Environmental Impact (Scope 3 GHG Emissions)

UPSs are evolving, to be bi-directional and grid-interactive. A traditional UPS is only focused on distributing power downstream to IT appliances but a grid-interactive UPS, which usually only requires a firmware update, enables the flow of energy upstream back to the grid as well. The UPS continues to serve its primary function of providing backup power in case of mains power failure but enables new modes of operation. The simplest addition is peak shaving, by which batteries are charged by the mains power during non-peak hours and discharged during peak hours, to reduce demand on the grid and to lower peak electricity costs. Another application is providing energy services back to the grid, such as frequency regulation or demand response. These services actually turn the UPS into a revenue-generating asset. When large numbers of data centers participate in utility programs together, they limit the need for utilities to fire up demand-response generators during peak usage, lowering carbon emissions from the grid operator.

Another aspect of grid-interactive UPS that is not to be overlooked is enabling easier integration of renewable energy resources. The intermittent nature of renewable energy resources requires energy storage solutions (ESSs) to be deployed alongside renewable power generation in many cases. These ESSs then help smooth out the intermittent generation and predictable consumption of renewable energy. As large energy consumers, data centers already have a large energy storage footprint that is substantially unutilized. Taking advantage of those data center assets, interfacing through a grid-interactive UPS, enables the data center to become a "buffer" for the grid, absorbing excess renewable energy capacity when it's available and using it for peak shaving or providing grid services when there is a shortfall. As the data center industry grows at the same time as increasing penetration of renewable energy assets into power grids, grid-interactive UPSs can play a critical role in ensuring appropriate use and storage of renewable energy resources.

The Data Center Industry Must Grow Sustainably or Prepare to Be Regulated

Following the onset of the pandemic and the resulting shift in business and consumer behavior, the data center industry received broad acknowledgement for keeping the internet—and the global economy—from falling into despair. And while the industry received accolades for what it has been able to accomplish during the pandemic, many began to realize the growth trajectory that the data center industry was on and questioned what that meant for environmental sustainability.

This concern has since been voiced by regulatory bodies, investors, and customers alike, placing increased scrutiny and pressure on the future of the data center industry as it relates to sustainability. Examples of regulatory efforts on data center sustainability, in the context of energy use and rising carbon emissions, have already started to materialize. Data center moratoriums have been put in place in a number of countries and cities with a concentrated number of data centers. Dublin (Ireland), Amsterdam (the Netherlands), and Singapore, for example, all currently have or recently have been under data center moratoriums, preventing or limiting new data center construction, while governments develop further regulations regarding how to approve new construction projects and manage existing data center operations.

Likewise, there are emerging regulatory actions around equipment efficiency, refrigerants used, and eWaste. Notably, in 2021 the Chinese government developed a three-year data center plan (Ministry of Industry and Information Technology Communications [2021] No. 76) to achieve average utilization of data center space and power assets of 60%, while also lowering power usage effectiveness (PUE) of newly built data centers to below 1.3 by 2023. Additionally, this plan also promotes the adoption of renewable energy, particularly distributed photovoltaic (PV) power generation for new data centers.

Lastly, the investor community has been growing in terms of its activism for companies to do more, factoring in sustainability as a risk or growth factor for companies. One way this is happening is the issuance of "green bonds" from data center colocation providers. These green bonds are used to increase capital to finance new projects and innovations adhering to specific sustainability standards. In 2020 and 2021, Dell'Oro Group observed the issuance of more than $21 billion in green bonds specific to data center deployments, such as deploying water-saving thermal management systems. A significantly larger number of green bonds was issued related to the data center industry, to fund projects focused on renewable energy production and reduce construction waste to avoid generation of millions of metric tons of CO2.

As these regulations—together with investor and customer requirements—materialize, this truly is resulting in sustainability becoming a competitive advantage. And it's becoming clearer what those competitive advantages are. They include, but are not limited to:

- The opportunity to reduce CapEx and/or OPEX, ultimately lowering the total cost of ownership (TCO) of data center assets.
- The ability to attract new customers based on sustainability requirements and generate new revenue streams.
- The generating of investment through green bonds and company evaluations, as a result of Environmental, Social, and Governance (ESG) reporting.

As data center owners and operators continue to increase awareness about the importance of building and operating more sustainable data centers, enterprises are looking for transparent partners to which to outsource cloud and colocation services to support their own sustainability agenda. For enterprises to evaluate vendors to support their on-premises data center needs and cloud and colocation providers for their outsourced data center requirements, they need to understand how sustainability metrics are tracked.

Scope 1, 2, and 3 GHG Emission for Data Centers

When it comes to quantifying sustainability, the primary metric used is greenhouse gas (GHG) emissions, measured in metric tons of CO2 emitted. To better understand GHG emissions, the tracking of which is complex, they are categorized into scope 1, 2 and 3 emissions. The different scopes of emissions help decipher who is responsible for their generation and how they are reported. These are the basic requirements for developing a strategy to track and improve the sustainability impact of a business.

Scope 1

Emissions are a reporting company's direct emissions. These emissions are generated from directly owned assets. For a data center, this is largely associated with on-site power generation from generators, potential refrigerant leaks from thermal management systems, and company owned vehicles. For data centers, scope 1 emissions are a relatively small source of GHG emissions. However, data center owners and operators are looking at ways to reduce scope 1 emissions. Eliminating diesel generators in favor of natural gas generators, deploying UPSs with longer energy storage runtimes, and exploring opportunities in fuel cells are a few examples. Performing routine maintenance and servicing of thermal management infrastructure to limit potential refrigerant leaks and adoption of electric vehicles (EVs) for company uses represent two other examples.

Scope 2

Emissions are indirect emissions generated from the purchase of electricity for the reporting company. This is a significant source of GHG emissions for data centers, as they consume a substantial amount of electricity. Today, this is where the majority of action on data center sustainability is taking place. Cloud and colocation service providers are utilizing purchase power agreements (PPAs) to secure renewable electricity capacity and renewable energy certificates (RECs) to offset fossil fuel energy use. While those are positive developments, these RECs and PPAs have developed in such a way that renewable power generation is not necessarily on the same grid or even in the same region that a data center is operating. This leads to a slight imbalance with regard to the benefits and burdens that data centers have on various grids.

Another way that data center owners and operators are actively reducing scope 2 emissions is through increasing data center efficiency, to limit electrical losses. Modern-day data center infrastructure has higher energy efficiency, even relative to products developed several years ago. Power conversion in UPS systems primarily operates in the range of 97–99%, while thermal management infrastructure systems more commonly utilize free cooling and evaporative technology to lower electricity use. When using evaporative cooling, however, it's important to consider the relative increase in water consumption to achieve lower electricity use.

Lastly, scope 3 emissions are any other indirect emissions (excluding scope 2 emissions) in the upstream and downstream supply chain of the reporting company. For a data center owner or operator, examples include purchase of goods and services such as construction materials, embodied carbon in data center IT and physical infrastructure, transportation and distribution of materials and products within your supply chain, and generated waste. For an enterprise, outsourcing an on-premises data center in favor of a public cloud or colocation service provider would constitute scope 3 emissions, as opposed to scope 1 and 2 emissions in the context of an on-premises data center.

Scope 3

Emissions represent the significant majority of GHG emissions in the data center ecosystem, and yet, are the most difficult to measure. That's because they are not under the direct control of the company that is reporting GHG emissions. Measuring or estimating scope 3 emissions requires a much more sophisticated process, which involves collaborative efforts from suppliers and customers—and is still being defined. That's why, when considering how the data center industry can increase its focus on sustainability in the context of accelerated digital transformation, we are turning our attention to scope 3 GHG emissions. But in order to do that, we need to discuss standardized sustainability reporting.

Standardized Sustainability Metrics: You Can't Manage What You Can't Measure

The difficulty of tracking data center emissions has prompted an industry review of data center specific metrics. Today, Power Usage Effectiveness (PUE) is the industry standard for measuring data center efficiency, which loosely serves as a measuring stick for a data center's sustainability. PUE is calculated by considering the total power consumption of the data center divided by the power usage of IT equipment:

$$PUE = \frac{total\ data\ center\ power}{IT\ equipment\ power}$$

The simplicity of calculating PUE has helped make it the industry standard. Because of that simplicity, however, data center architectures and product designs have helped improve PUE measurements without addressing greater sustainability concerns. For example, adding more internal server fans for thermal management, instead of centralized thermal management systems, would improve the PUE despite consuming more energy. Additionally, PUE does not account for consumption of water. As evaporative heat rejection technologies have been adopted, data center water usage has risen significantly, while PUE measurements declined.

This led to the development of another data center sustainability related metric, water usage effectiveness (WUE). This metric was designed as a standalone metric, to measure how effectively water is used, relative to IT equipment energy use. It is calculated in the following manner:

$$WUE = \frac{annual\ water\ usage\ (L)}{IT\ equipment\ energy\ use\ (kWh)}$$

This gives data center owners and operators the ability to not only measure the efficiency of the data center, but the specific role that water plays in reducing electricity use. This also provides data center customers with a standardized metric for making comparisons among data center service providers to influence their decision making, depending on their criteria.

Looking to the future, the idea of data center sustainability is expanding beyond data center efficiency to reusing data center byproducts, namely the heat generated from IT appliances. As data center owners and operators increasingly become willing to invest in infrastructure and architectures that enable such heat reuse, preventing the need for electricity use elsewhere to generate that heat, it only makes sense to quantify those beneficial efforts. This led to the development of the metric of Energy Reuse Effectiveness (ERE). ERE is similar to PUE in that it measures the ratio of total data center power relative to IT equipment power, while also providing a "credit" for energy reuse.

$$ERE = \frac{total\ data\ center\ power - energy\ reuse}{IT\ equipment\ power}$$

But these metrics of PUE, WUE, and ERE alone won't be sufficient to measure, report on, and drive sustainability decision making in the data center industry. Additional metrics—covering waste, equipment reuse, and IT efficiency—are still being investigated. Future evaluation, development and standardization of data center sustainability metrics are still called for. Without that, data center owners and operators are likely to focus on the metrics that present them in the best light, while leaving gaps in their reporting where they fall short. For enterprises, this means potential uncertainty on the holistic impact on sustainability from their data center footprint or service provider, with guesswork and unverified claims driving sustainability decision making.

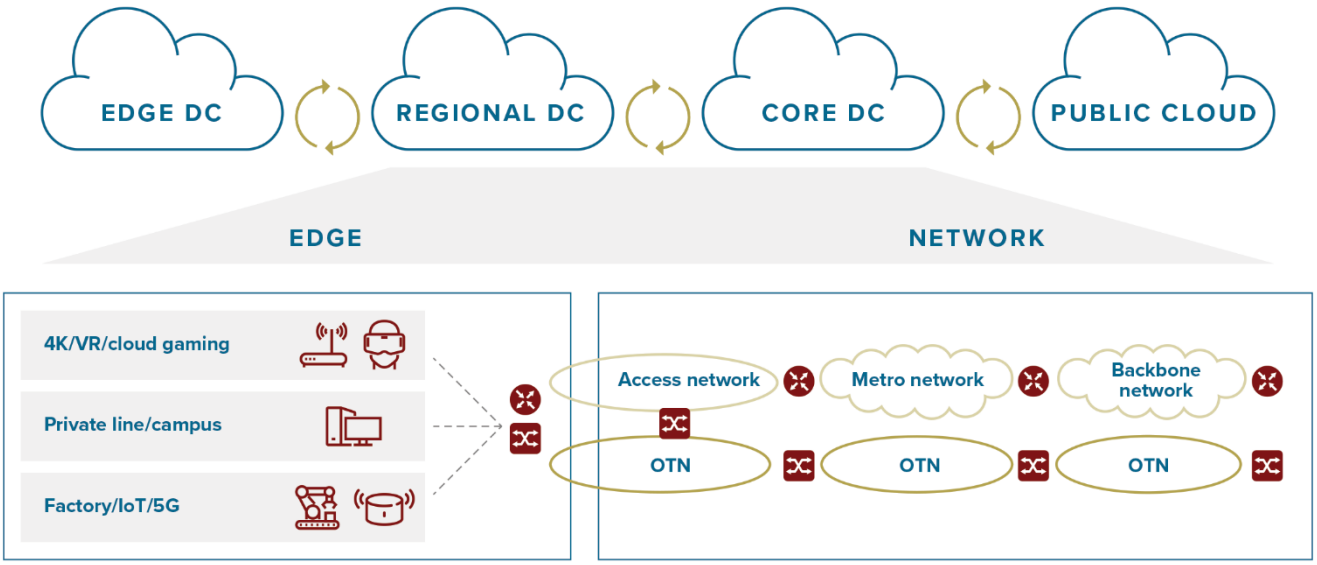
End-to-End Sustainability Best Practices

Data center planning, construction, and life-cycle management are complex. From right-sizing data center needs for the business to choosing which technologies to deploy—and whether that's on-premises or outsourced in the cloud or a colocation facility—many factors call for consideration. Adding sustainability criteria to that process, without increasing costs or impacting reliability, only increases the complexity. However, with today's modern tools and technologies, best practices can enable an enterprise or service provider to plan and design, construct, operate, and maintain a data center, to optimize its sustainability impact.

Planning & Design: Intelligent Simulation and Site Selection Optimizes Design PUE

As enterprises and telecom operators embark on the next generation of digital transformation, future-proof sustainable digital infrastructure becomes mission critical for meeting the challenges and ensuring successful digital strategies. Figure 7 depicts a typical data center architecture, where a layered data center structure of cloud, on-premises (core) and regional data centers work together with communication networks to offer highly scalable and agile infrastructure, to create a high performance foundation for digital services moving forward. Sustainability needs to be considered and fulfilled throughout all layers in the architecture.

Figure 7: Data Center Target Infrastructure



Source: Dell'Oro Group

Excellent planning and design form the foundation for achieving efficient and energy-saving goals for data centers. Selections of efficient infrastructure, coordination of control logic, and seamless collaboration among various power and thermal management systems are all critical for efficient system operations. Traditional design methods rely on human design experience by utilizing accumulated knowledge to manually calculate and predict PUE outcomes. This oversimplifies the complex working mechanisms, system coordination, and control process of the thermal management system. Therefore, it becomes difficult to ensure the accuracy of the PUE calculation, with results sometimes deviating from the initial design. The utilization of intelligent digital models to simulate PUE design based on design state modeling of building information modeling (BIM) technology can effectively resolve this problem.

An intelligent digital model, sometimes referred to as a "digital twin," is built to simulate the operation and PUE of the data center based on environmental and operational factors such as local climate and IT load. Based on the one-dimensional thermal fluid simulation, three-dimensional air distribution simulation, and temperature field simulation technology, a data center spatial model, a thermal management system model, and a power distribution model are built for physical modeling. Then, the operational and control process of the system is simulated to obtain the dynamic and steady-state performance data of the chiller, the cooling tower, the water pump, and other equipment. This yields a prediction of the complex heat flow system. As a result, it becomes possible to simulate the system running status with high precision by the hour, to verify and optimize the design scheme, and to ensure that the PUE can be accurately implemented.

While intelligent simulation drives optimization in a data center design, site selection also plays a critical role in determining scope 2 GHG emissions. The local climate and available electricity

source in particular become defining factors. For some data center deployments, the site or location will be predetermined based on business needs, but others may have the luxury of evaluating different locations. The climate will help dictate what type of thermal management can be used in maximizing PUE. In regions with cool, dry climates, the option to use free cooling, where ambient air can be used to directly or indirectly cool data center operations, may be possible. This enables a data center to limit thermal management energy consumption and be designed with a lower PUE. In regions with hotter and more humid climates, chilled water or direct expansion thermal management systems are required, which while necessary, result in higher design PUEs.

The source of electricity also plays a critical role in site selection. With data centers consuming a significant amount of electricity, how that electricity is generated—together with its availability—may determine a data center site's viability. As not all grids are created equal, it is the power generation mix of a particular country or region that will determine scope 2 GHG emissions for a data center. Renewable power generation (such as solar, wind, or hydropower) provides zero-emissions electricity, but with lower availability. When renewable energy is not available, fossil fuels—which carry considerable scope 2 GHG emissions—are used. To ensure renewable energy availability, some data center service providers carry out power purchase agreements (PPAs) to secure long-term capacity.

When renewable power generation isn't an option, it's possible to offset fossil fuel GHG emissions with renewable energy credits (RECs), an approach commonly used by hyperscaler data center service providers. However, it is important to note that sometimes renewables are not on the same grid or even in the same region in which a data center is operating. This leads to a slight imbalance of the benefits and burdens of a data center on various grids. Aligning renewable power generation on the same grid as data center operations not only increases the sustainability of the data center's operations but opens the door to the utilization of grid-connected UPSs, supporting decarbonization of the grid.

Construction: Reducing Scope 3 GHG Emissions with Prefabricated Construction and Digital Delivery

The effective implementation of the PUE design depends on the high-quality construction and deployment of a data center. Construction is often overlooked in its sustainability impact, primarily because it contributes to the scope 3 GHG emissions of a data center, which are largely untracked and unreported. However, sustainability best practices are already being established in data center construction. The sustainable practice of "digital delivery" is emerging through the adoption of prefabricated construction, digital operation capabilities, and visualized management of construction processes.

Prefabrication plays a critical role in digital delivery. In the data center industry, this construction practice has led to the development of prefabricated modular data centers (PMDCs) as a solution.

PMDC solutions integrate data center infrastructure in ISO container-like modules in a factory-controlled environment. This construction practice can be used to build out a data center's physical infrastructure such as power or thermal management, white space, or an entire data center from just a few racks to multi-megawatt (MW) deployments.

The construction of a PMDC in a factory-controlled environment allows for significant sustainability improvements in the process. PMDC construction relies more heavily on steel than concrete, in contrast to brick-and-mortar or "stick-built" facilities. Steel carries a lower amount of embodied carbon than concrete, reducing scope 3 GHG emissions through the use of this material. Welding components such as pipes in a factory-controlled environment can significantly improve the construction environment relative to a traditional construction site.

The controlled environment allows for more efficient use of materials, resulting in less than half the water consumption and generating less than half the waste. The controlled construction environment and concurrent construction on site also shorten the time-to-market (TTM) of the data center, from the 18–24 months generally required for a stick-built facility to the 3–6 months needed for a PMDC solution.

Digital delivery also relies on digital operation capabilities and visualized management of construction processes. Current brick-and-mortar data center construction involves a wide range of construction crews with various specialties and technologies, which is difficult to manage. Digital delivery resolves many problems and pain points during data center delivery through integrated and visualized methods to avoid quality problems during deployment, increasing reliability and uptime in future operations. Digital delivery helps realize standardization and automation, embedding intelligence in project planning tasks and quality, problem, and monitoring management.

These benefits are realized in a variety of ways. Monitoring and management solutions can reduce onsite inspection time, while environmental health and safety (EHS) intelligent warning systems can automatically identify problems and record them in the system. Data collected from monitoring is transferred to the digital delivery platform through cloud services, enabling data center monitoring remotely and at any time. Recordings can be replayed to ensure that the quality of projects can be monitored and tracked. 3D modeling is used to integrate data in various stages of digital delivery for documentation such as guidebooks, records, and reports, enabling efficient queries to optimize operation quality. During acceptance, augmented reality (AR) technology can be used to compare the virtual environment with the physical deployment. The difference between design and construction can be found quickly and conveniently, enhancing acceptance efficiency.

These digital assets are needed to build higher-standard data centers with sustainable construction, operation, maintenance, and management, at a lower cost and with greater efficiency. Data center vendors are relied upon in this process and need to help enterprises obtain

data in the construction phase from the source, improve project execution efficiency, support open data platforms for future data center operations and maintenance (O&M) optimization, and streamline upstream and downstream systems to build digital delivery and AI energy-saving capabilities.

Operations & Maintenance: Reconstructing DCIM for Sustainable Data Center Operation and Maintenance

After construction is complete, the data center is ready to go live. But even if sustainability decision making was ingrained in every element of planning, design, and construction, sustainability still needs to be kept at top of mind. In order to optimize sustainability during the operation and maintenance of a data center's lifecycle, data center infrastructure management (DCIM) has been reconstructed.

Over a decade ago, DCIM software was introduced to the data center industry, with the promise of collecting and aggregating data to optimize data center efficiency. However, that vision never materialized. This was primarily a result of the fact that at the time, DCIM was largely tailored for individual data centers and not always interoperable with multiple vendors' products. DCIM collected data that still needed to be analyzed to create action items and was passive in nature, detecting issues but still requiring manual intervention. Failing to live up to the promised vision, DCIM was never widely adopted.

Today, DCIM is being revisited and reconstructed. Based on the proliferation of sensors, utilization of centralized cloud services, and AI-driven automation, DCIM now has the tools to live up to the hype. Today's versions of DCIM are interoperable with different vendors' products, with optimization based on billions of data points in the cloud. They are proactive, not only in data aggregation but analysis, forecasting future performance and failures. Lastly, today's DCIM automates efficiency improvements, acting without requiring human intervention.

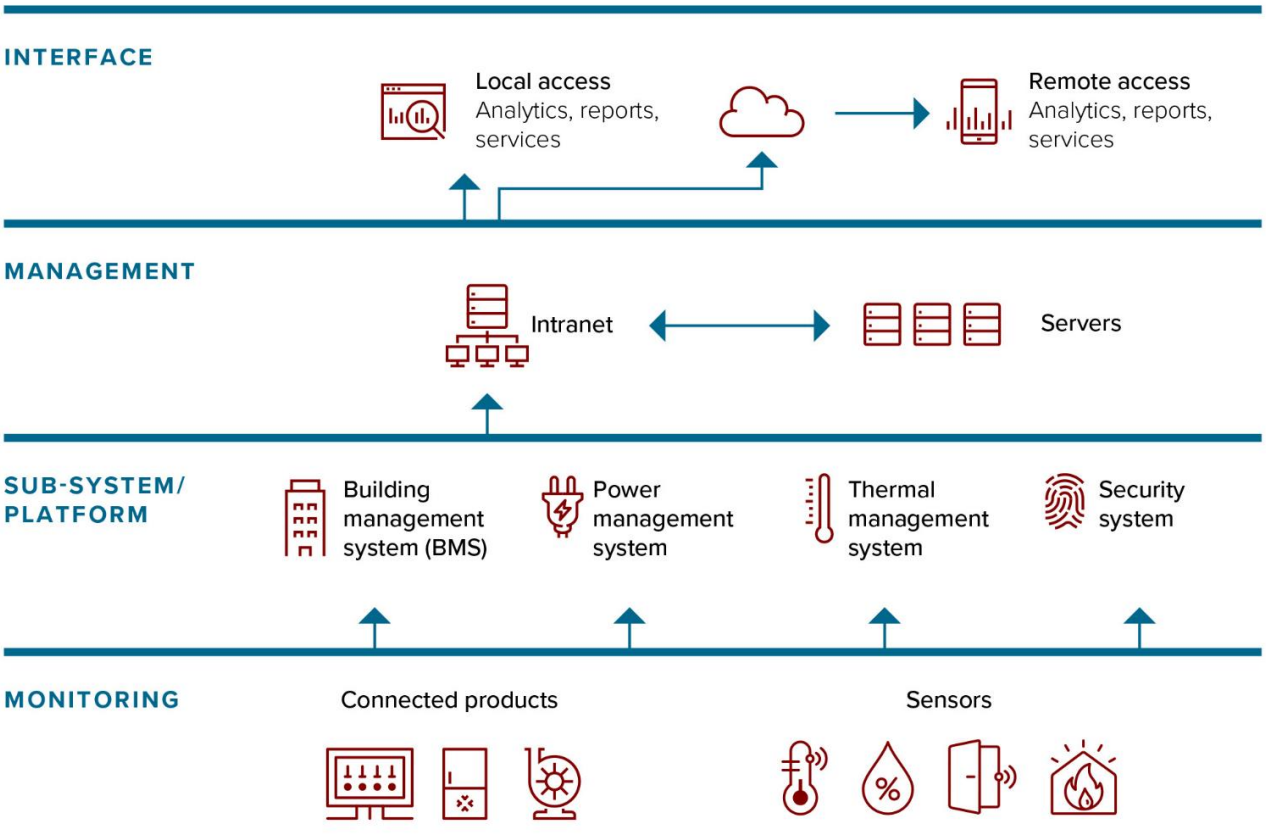
DCIM system architecture starts with the monitoring layer. Sensors and connected devices collect various data points that reflect operating conditions in order to build the digital foundation of a DCIM system. A sub-system monitoring layer aggregates those data through various platforms or systems, designed with the appropriate communication protocols for each specific system. The management layer supports secure communication of the data to a local intranet or centralized cloud, subject to a data center user's preference. Users interact with the DCIM system through the interface layer to access data sets and run analyses, visualize system performance, and execute services.

Examples of DCIM software include Huawei's NetEco6000, Schneider Electric's EcoStruxure, and Eaton's VCOM from Brightlayer. Figure 8 depicts implementation of the DCIM software with

autonomous maintenance and other features, which can help data centers reap a variety of benefits, including the following:

- AI energy optimization is able to perform energy efficiency analyses to optimize PUE, reducing data center energy consumption by 5–8%.
- Maintenance of equipment is simplified, with AI platforms predicting failures, automating maintenance practices, and increasing labor efficiency to reduce maintenance labor costs by up to 35%.
- Operational resource management is improved, automating asset discovery and increasing capacity usage by up to 20%, through aligning available power, thermal management, and rack unit (U) space.

Figure 8: Data Center Infrastructure Management (DCMI) System Architecture



Source: Dell'Oro Group

These customer benefits can significantly impact data center sustainability, while also increasing reliability and lowering a data center's TCO. By optimizing PUE and reducing energy consumption, electricity costs are lowered and scope 2 GHG emissions are reduced. AI supported maintenance can reduce scope 1 GHG emissions by quickly detecting refrigerant leaks and supporting expedited resolution. Additionally, utilizing predictive maintenance limits unnecessary maintenance visits, leading to OPEX savings and increasing reliability, while also reducing scope 3 GHG emissions. Lastly, AI-supported improvements in operational resource management maximize utilization of resources, limiting the need to grow a data center footprint prematurely, also contributing to a reduction in scope 3 GHG emissions. These wide-ranging benefits demonstrate the critical importance of today's reconstructed DCIM software and its significant potential impact on data center sustainability.

Continuous Optimization Can Further Reduce PUE

A data center can experience significant change over its lifecycle, ranging from different types of infrastructure deployments to changes in the local environment. This impacts the many devices and unique relationships they share in relation to energy consumption. Product-level energy-saving optimizations are approaching their limits. Traditional control methods operate too independently between these different devices, even when combined with experienced management. To achieve new levels of energy savings, continuous system-level optimization is required throughout the entire lifecycle of the data center. This relies on streamlined thermal and power management control logic, to adjust parameters in real time to match the operating environment. Collection of data from facility sensors and AI training can be used to unlock intelligent data center management for sustainable operations.

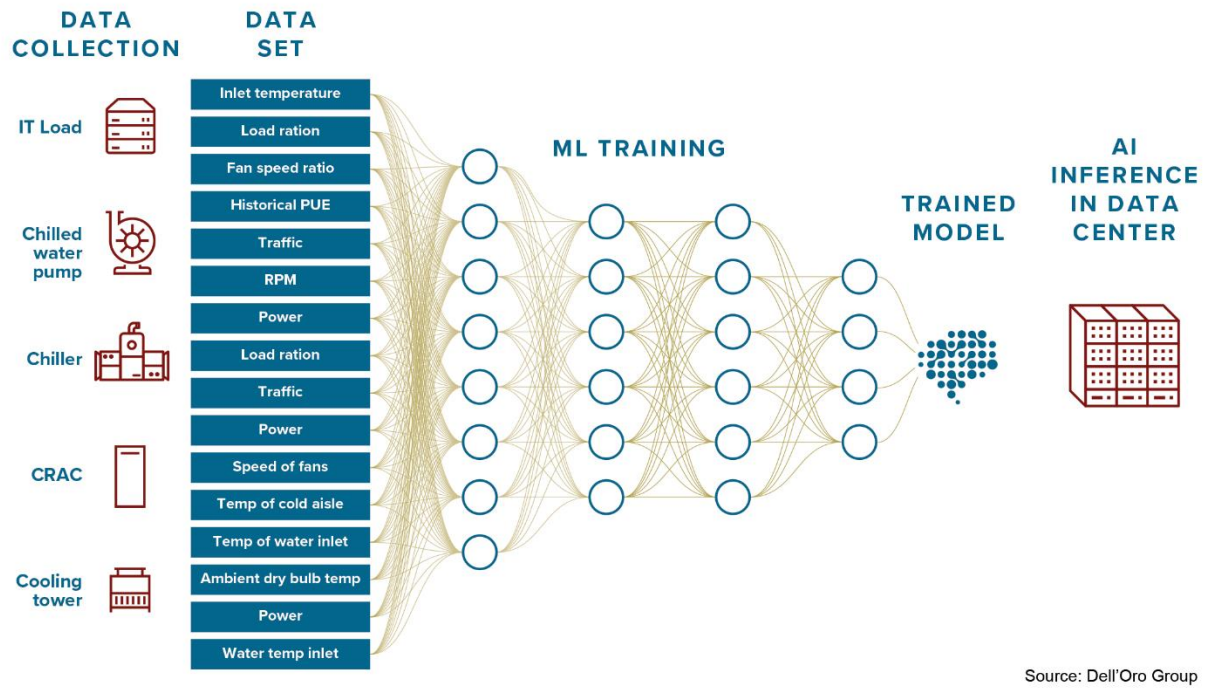
In order to achieve an optimized PUE, continuous simulations and analyses are needed to train the AI management platform. In the early stages of a data center, the design PUE will be modeled from a general knowledgebase of existing data center operations with similar characteristics. As the data center begins operations, inputs from on-site sensors are collected, delivering more precise operational impact from changes in parameters, both nonadjustable (temperature, IT power consumption, etc.) and adjustable (thermal and power management infrastructure). The incorporation of those expanded data can effectively improve the inference effect of AI algorithms, thereby further optimizing PUE. The PUE optimization increases from the original 5–8% design benefit to 9–12% as the AI algorithm matures.

The key to realizing such benefits is the simulation of different scenarios to which the data center may be exposed. For example, when a data center is faced with an abnormally hot day, the AI platform may start with overcooling the facility, to ensure reliable operations. However, the AI

platform can simulate different scenarios to achieve different outcomes. For instance, it may test the difference in PUE impact when increasing all thermal management units by 10% cooling capacity—or it might do so by only increasing half the thermal management units by 20% capacity. The cooling capacity would be the same but depending on the efficiency curve of the thermal management units, this may affect PUE differently. The AI platform will learn from these different simulation outcomes and apply what has been learned to optimize energy savings in future scenarios. With time, the AI platform will have gathered enough data to optimize the PUE of the data center through the end of its lifecycle.

Figure 9 shows an example of the AI optimization process, which consists of two phases: model training and inference. During the training phase, the system collects historical raw data from related devices and sub-systems, including IT load, chilled water pump, chiller, CRAC, cooling tower, and others. After the data cleansing and correlation process, the system generates useful data sets such as inlet temperature and fan speed ratio from IT load, load ratio and power usage from chillers, water temperature and ambient dry bulb temperature from cooling tower, etc. The system utilizes machine learning algorithms to generate a model and continuously train the model until it passes the test and becomes the trained model for use. The inference phase utilizes real-time data from production and fits it into the trained model to get the best configuration combination of the various parameters, automatically controlling the devices and get the optimized energy efficiency.

Figure 9: AI Optimization Model and Interfence



Data Center Sustainability Use Cases

Figure 10 below summarizes some best practices in data center sustainability, along with the impact discussed above. These best practices pertain to data centers transitioning to renewable energy, reducing water usage and carbon footprint, improving utilization of servers and IT equipment, and optimizing data center equipment life-cycle management. These actions taken in combination have helped the cloud service providers and other data center operators to meet their sustainability targets.

Figure 10: Best Practices in Data Center Sustainability

Sustainability Impact	Data Center Sustainability Action	Details
Reduce Scope 2 & 3 GHG Emissions	Transition to Renewable Energy	Enter into long-term purchase agreements with independent power producers to supply renewable electricity.
		Invest in self-owned renewable energy projects on site.
	Reduce Waste and Carbon Footprint	Achieve TRUE certification (Total Resource Use and Efficiency, formerly known as Zero Waste certification), with more than 90% of the data center waste generated sent to recycling or composting.
		Divert waste from global data center operations away from landfills.
		Utilize building strategies such as prefabrication to reduce waste during the construction of the data center and enable faster time to market.
		Match data center electricity use with regional carbon-free sources.
		Recycle heat from a data center to provide a low-carbon, low-cost source of heat for neighboring public sector, residential, and commercial customers.
	Improve Server Utilization	Increase the deployment of converged infrastructure to optimize the provisioning of compute and storage resources in multi-tenant environments.
		Adopt new server architectures that support new interfaces such as CXL to enable the sharing of memory resources over a distributed network of servers.
		Deploy accelerated computing platforms such as GPUs and DPUs (Smart NICs) for more efficient processing of select workloads.
	Optimize Life-Cycle Management	Practice circular economy to reuse components used for server upgrades and retrofits.
		Conduct internal reliability studies to extend useful life of data center equipment.
		Manage server and other equipment end-of-life by reselling equipment in secondary markets.
Improve Data Center PUE & WUE	Reduce Power Consumption	Utilize liquid and immersion cooling to reduce server power consumption through improved thermal management.
		Optimize ambient cooling conditions.
		Deploy servers with custom ARM processors to improve performance per watt.
		Employ AI simulation to optimize data center operational parameters.
	Reduce Water Usage	Harvest rainwater and reclaim condensate.
		Increase the use of direct evaporative cooling systems, as opposed to indirect evaporative systems.

Below are some actual sustainability use cases practiced by US-based and China-based cloud service providers. While the US-based cloud service providers generally have the largest data center installed base and footprint, the China cloud data center market is still in the early stages, with long-term growth opportunities ahead.

- **United States**

  - Google operates a vast network of data centers in nearly 40 global regions to support its core advertising and emerging public cloud businesses. As Google continues to scale its infrastructure, the firm needs to be mindful of driving greater efficiencies, while meeting strict environmental regulations and fulfilling its corporate social responsibility (CSR targets. Google has been utilizing circular economy and waste management best practices for many years to lower its environmental impact. For instance, up to a quarter of the components used for server upgrades were from refurbished inventory and millions of components were resold to secondary markets. As a result of these actions, up to 78% of the waste generated by this cloud service provider's data center operations were diverted from landfills. In addition, this cloud service provider has been conducting extensive reliability studies and optimizing server designs, extending server useful life by as much as five years. These actions have reduced server purchases, resulting in substantial CapEx savings and reduction of carbon emissions over time.

  - Meta (formerly Facebook), operating millions of servers in its global network of data centers, has been expanding its infrastructure at a heightened pace. The firm has also increased its investments of AI infrastructure in order to enhance engagement on its social media platforms, as well as to lay the foundation for the emerging metaverse opportunity. AI infrastructure is power hungry, with demanding sustainability requirements. Today, Meta's global footprint of data centers—running on 100% renewable energy— features six new data center buildings totaling 3 million square feet, which have been awarded the Leadership in Energy and Environmental Design (LEED) Gold certification. These new data centers have achieved a PUE of 1.09 and a WUE of 0.26. One method used to conserve water is through direct evaporative cooling, which utilizes ambient air to cool data center equipment. In environments that do not support the use of ambient air due to dust, humidity or salinity, indirect evaporative cooling is used. Meta uses a specific indirect evaporative cooling solution that uses less water than a typical indirect cooling system because it uses air to cool water instead of the reverse. Furthermore, in conjunction with direct evaporative cooling, optimizing the relative humidity resulted in substantial reduction in data center water usage.

- **China**
  - In 2020, Huawei and China Telecom collaborated on a project at a facility with 1,570 rack cabinets, covering 7,000 square meters intended for a carrier end user. Huawei developed an innovative method to reduce data center PUE through the use of AI-based optimization. AI optimization was able to optimize the data center operating parameters to fit a wide range of ambient conditions to reduce PUE from 1.57 to 1.42. This PUE improvement resulted in a savings of 2.89 million kWh of electricity, equating to 1.16 million RMB per year for the customer.
  - Huawei has promoted the value proposition of prefabricated modular construction and green, intelligent energy saving for the industry. A recent example was the Fujian data center for a telecom operator, a facility with approximately 110,000 square meters of floor space when complete, with first phase implementation of 2,910 IT cabinets. Through prefabricated data center construction and digital delivery in conjunction with utilizing an AI optimized design—adopting high-efficiency cooling, modular power distribution, and intelligent energy-saving optimization technologies—the data center has been able to reduce the annual PUE from 1.6 to less than 1.31. The prefabrication and digital delivery method also reduced the facility's TTM by six months.

These highly scalable use cases are complementary to various data center sustainability best practices for meeting stringent PUE, WUE, and carbon footprint requirements.

Conclusion

The data center sustainability journey has just begun. The scale of the sustainability challenge can be intimidating but it also creates countless opportunities to innovate in the process of achieving desired goals. The data center industry can be among the first, if not the very first, industry to not only set significant sustainability goals—ranging from carbon neutrality to consuming 100% renewable energy and eliminating waste—but to actually achieve them. The data center industry can set the sustainability blueprint for other industries to follow.

To achieve desired goals, it is essential that the power of collaboration be harnessed. We believe that one of the first and most important steps is for the data center industry to define, set, and standardize data center sustainability standards and metrics on which to report. Sustainability should not be a game of marketing, but rather one of accountability and action. Without that, the growth of the entire industry could be threatened by regulation. No solution, technology, or company can achieve sustainability without collaboration. We encourage you to have conversations with your partners and customers to develop ideas and create learnings in achieving sustainable data center growth.

About Authors

Baron Fung joined Dell'Oro Group in 2017. Since joining the firm, Mr. Fung has significantly expanded our coverage on data center infrastructure, with an emphasis in cloud service provider capex, and server technology and vendor trends. In addition, Mr. Fung built and launched coverage of Smart NICs and is responsible for the Data Center Capex, and Ethernet Adapter & Smart NIC market research programs. Mr. Fung's research and analysis have been widely cited in business publications. Mr. Fung has appeared as an invited speaker at investor conferences.

Lucas Beran joined Dell'Oro Group in 2021 and is responsible for coverage of Data Center Physical Infrastructure. Mr. Beran's research and analysis has been frequently quoted in leading trade and business publications including Bloomberg, CRN, Data Center Knowledge, Information Week, Mission Critical, SDxCentral, The NexPlatform and The Verge. Mr. Beran regularly speaks at industry conferences such as Data Center Dynamics and Data Center World, OCP (Open Compute Project) and also presents to senior executives at corporations. Mr. Beran graduated cum laude with a B.A. in Economics and Applied Mathematics minor from the Honors College of Boise State University.

About Dell'Oro Group

Founded in 1995 with headquarters in the heart of Silicon Valley, Dell'Oro Group is an independent market research firm that specializes in strategic competitive analysis in the telecommunications, security, enterprise networks, and data center infrastructure markets. Our firm provides world-class market information with in-depth quantitative data and qualitative analysis to facilitate critical, fact-based business decisions. Visit us at www.delloro.com.

About Dell'Oro Group Research

To effectively make strategic decisions about the future of your firm, you need more than a qualitative discussion – you also need data that accurately shows the direction of market movement. As such, Dell'Oro Group provides detailed quantitative information on revenues, port and/or unit shipments, and average selling prices – in-depth market information to enable you to keep abreast of current market conditions and take advantage of future market trends. Visit us at www.delloro.com/market-research.

Dell'Oro Group
230 Redwood Shores Parkway
Redwood City, CA 94605 USA
Tel: +1 650.622.9400
Email: dgsales@delloro.com
www.delloro.com