GENERATIVE AI AND ITS **IMPACT ON DATA CENTER DESIGN**

> "Recent developments in advanced computing systems configured for **generative AI and machine learning** have led to an explosion in demand for power and **enhanced design solutions** for powering and cooling modern IT compute deployments."

Data center developers and their design teams will face challenges in obtaining available power and will need to explore non-traditional ways to attain, produce, and deliver that power to the IT kit.

Read on to learn about the top emerging technologies driving the need for innovative solutions in data centers.

Artificial Intelligence (AI) and Machine Learning (ML) will occupy almost every data center to some degree moving forward. Facilities that were previously designed and constructed for traditional computing environments can accommodate some AI/ML builds, but their capabilities will be limited without renovation and upgrades to accommodate the unique power densities and load profiles of these systems.

Power Densities | The latest chipsets from NVIDIA are designed to draw as much as 1,200W or more per GPU. Certain configurations presently offered by NVIDIA consume as much as 120kW per cabinet, which is well beyond the limitations of air-cooled designs. These densities are only expected to increase in the coming years.

Cabinet/Row Configurations | AI applications often include unique network architecture to account for low latency connections between processing, network and storage equipment. These constraints will require consideration when determining row spacing, row length, any in-row cooling distribution equipment, and support systems.

Infrastructure Resiliency | AI applications are a subset of high-performance computing, which often incorporates varied levels of infrastructure resiliency due to the nature of the applications. The majority of the IT kit is composed of parallel processing across a high speed production environment which requires significant power but does not often demand redundant power sources, and often does not require UPS power with battery backup. This is in part due to the increased application resiliency, with automatic fail-over and load balancing occurring within the same environment or in a separate room or even building. More critical IT infrastructure such as network, storage, and master compute nodes require a higher level of MEP infrastructure resiliency to support the critical nature of these systems.

Cooling High-Density Compute Loads | The power densities being observed with AI applications loaded onto HPC clusters are driving the need for advanced cooling designs, focusing squarely on liquid cooling applications. The load densities are simply too high to rely solely on an air cooled solution, which has resulted in a variety of proposed solutions. Localized cooling bringing chilled water to the cabinet has been a solution deployed many years, utilizing either in-row fan coil units or rear door heat exchangers (RDHX). These solutions bring water to the row but not specifically to the chip. Water is over 23x as efficient as air in transferring heat and is therefore a more appropriate solution for the high density loads inherent in HPC and AI deployments. Recently, a variety of direct liquid cooling solutions have been introduced ranging from single-phase and two-phase liquid to the chip designs to complete immersion cooling solutions, both single and two-phase.

The direct liquid to the chip (DLC) solutions typically incorporate a cooling distribution unit (CDU) which is essentially a heat exchanger and pump package interfacing facility water with process (IT infrastructure) water, which is run directly to a plate mounted directly to the board housing the GPU chipset.

Several manufacturers have entered the CDU market offering packages that provide up to and even exceeding 1,000kW of cooling capacity in a single CDU. The liquid to the chip cooling design only provides about 80% of the heat rejection required at the server, because there are certain components within the server that require air as the cooling medium. Thus, it may not be uncommon to see both a direct liquid to the chip solution coupled with an in-row or RDHX application to address the air cooling component of the design.

There are some complexities associated with the direct liquid to the chip solution. The liquid used for the process water requires a minimum 25% glycol mixture (PG25) to prevent the formation of contaminants. It also requires a 50-micron filter to capture any particulate that would impede the fins within the cold plate. These items are to be inspected and maintained on a regular basis, though the frequency is not expected to be often, as the system is sealed and uses stainless steel piping for these reasons.

Another important consideration is resiliency and responsiveness of the process water loop should a power loss occur at the facility water loop or at the CDU itself. These can be addressed by incorporating UPS power into the mechanical infrastructure, as well as providing additional reserve cooling capacity via thermal storage. These upgrades have up-front and ongoing cost impacts and require additional real estate within the facility, and need to be considered and decided upon in the early planning process of facility design.

The immersion cooling solution includes the use of a tank filled with a dielectric liquid in which the servers will be submerged. In single-phase immersion cooling, a CDU similar to that used in DLC applications acts as a heat exchanger between a facility water system and the dielectric fluid used in the immersion tank where the servers are installed. Fluid is circulated from the CDU to the tank and flows across the server, removing heat directly from the components and to the facility water through the CDU. The immersion tank is typically filled with a fluorocarbon based fluid with a low boiling point, which results in a phase change when it is exposed to heat-generating components of the servers immersed in the fluid.

The vapor condenses on a coil at the top of the tank and returns to liquid form in the tank. The condenser coil at the top of the immersion tank is tied to a process liquid loop which in turn is pumped through the coil by the associated CDU. This is one of the most efficient means of removing heat from the servers, because the entire surface area of heat generating devices is immersed in the fluorocarbon based fluids. One of the challenges with immersion cooling is the continued focus on the environmental impact of PFAS (perfluoroalkyl and polyfluoroalkyl substance) liquids, which do not break down and can accumulate in the environment causing potential issues over extended periods of time. The potential for immersion cooling applications remains high because of its effectiveness in removing heat from processing equipment. However, it will lag the DLC cooling in global acceptance because it requires further analysis from multiple perspectives, including material compatibility (i.e., immersion fluid with server components), and facility adaptability to accommodate a very unique footprint, as well as familiarity with server maintenance using immersion tanks.

Powering High-Density Loads | The advent of AI in the data center is a very disruptive force, requiring another look at how to effectively distribute reliable power to the high performance IT systems being deployed to handle AI workloads. Over the past decade, the majority of designs for hyperscalers and cloud providers have migrated to higher voltage distribution, utilizing 415/240V to deliver power to the cabinets. Conventional IT equipment power supplies are manufactured with a wide voltage range, typically 100-240VAC, allowing them to be deployed globally with a simple cord change based on the local distribution standards.

Prior to the recent evolution of AI within the high-performance computing world, a single 415/240V 30A circuit (a second circuit if needed for redundancy) would meet most power requirements for a what would be described as a high-density (15-17kW) server cabinet. Equipment built to process AI workloads are being designed with GPUs that have a much higher TDP (thermal design power) than legacy designs. As an example, in order to meet the requirements for a fully populated NVIDIA Grace Blackwell DGX GB200 NVL72, a single 120kW cabinet would require ***Eight (8) 30A 415/240 circuits***, without accounting for any circuit redundancy.

This is impractical, and is forcing manufacturers to look at other options to reliably supply power to these large loads. One approach developed by the Open Compute Project (OCP) is the ORV3 NVIDIA MGX rack, specifically designed to support the 120kW NVIDIA GB200 NVL72, and incorporates built in busbar, network spine, and DLC cooling manifold within the cabinet (https://www.opencompute.org/products-chiplets/525/cheval-group-open-rack-v3-nvidia-mgx-rack-for-gb200-nvl72). The OPC is leading the industry in developing and deploying detailed solutions with collaboration from the IT equipment manufacturers, which results in custom solutions for specific computing applications.

Other concepts being considered to address the 100kW cabinets encountered today, as well as the 300-1000kW cabinets that will be available in the not too distant future, include raising the utilization voltage of the IT equipment. Distribution options ranging from 480V up to 600V and above are presently being considered. These options will require a change in IT equipment manufacturing to allow the server power supplies to receive higher voltages described previously.

This is not a simple undertaking; this would require product advancements and employing a higher equipment insulation class at the cabinet level to accommodate the higher voltage, and it may also result in higher incident energy levels that could impact the PPE requirements for operations staff maintaining the IT infrastructure.

A final item to consider when discussing the challenges that AI workloads will present to the infrastructure systems serving them is the load profile observed at the GPUs processing IT workloads, particularly when training an AI model. The unique load profile is transient in nature, resulting in erratic power consumption variations that change over the course of milliseconds. These load transients can jump from 10% rated power, up to 80%, then can jump from 80% up to 120% and even as high as 150% of rated power during the training exercise. Training of AI models can last for days or weeks and can span thousands of GPU clusters, creating large power variations on a building-wide level, and at scale, even at campus level. This is wreaking havoc with local utilities supplying power to these facilities, and causes significant concerns for a campus that may generate power locally during utility outages. There are solutions to these issues in the works, potentially being addressed at the rack level with capacitors or other stored energy devices, and also being considered at the centralized UPS level, looking at UPS performance under varying load conditions and determining responsiveness to these load conditions. Ultimately, some of these solutions will be developed as these deployments are installed, based on real-world experience with bespoke solutions tailored for the specific circumstances of the AI installation.

HED

Written by Jack McCarthy PE, DCEP
Principal | Market Sector Leadership
Explore more market and design insights at
www.hed.design/insights