Intersect360 Research White Paper: STAY COOL: DDC EASES THE RISK IN EMBRACING THE AI FUTURE

MARKET DYNAMICS: ENTERPRISE AI GETS HOT

It's no secret what big new technology is driving data center investments today. AI is the revolutionary anthem everyone is singing; in fact, there are multiple available chatbots that would quickly write you the lyrics. The cyber-generated verses might be good or might be bad (sometimes laughably so), but AI is getting better by leaps and bounds, and it holds extraordinary promise for almost any enterprise.

Business leaders see this promise and are eager to invest, but they quickly are faced with the challenges of any revolution. Once convinced of a destination, the next question is how to get there, and specifically, how to get there *from here*. Rarely will a company want to start totally new – new buildings, new staff, new real estate, new contracts. Those things were hard enough to amass the first time. Protecting existing capabilities and business operations is paramount.

Furthermore, AI is evolving so quickly, it represents a moving target. What will an enterprise AI data center need in 2030? By the time 2030 arrives, any plans made in 2025 may seem foolish. For any enterprise to go all-in on AI, it would be nice to have some kind of safety net, leveraging existing infrastructure and capabilities and moving flexibly into the future.

The AI Data Center

The current drive toward enterprise AI has been years in the making. Initial experiments with AI date back to Alan Turing, and his now-famous Turing test. In the 1990s, IBM's Deep Blue supercomputer used artificial intelligence to defeat chess grandmasters. The current wave of machine learning advancements included another IBM innovation, when its Watson AI competed on *Jeopardy!* in 2011. But what has truly enabled the modern AI revolution is hyperscale computing.

Hyperscale companies are those with massively scalable, internet-driven business models that scale arbitrarily to any size, bound only by how many users connect to them. Amazon, Google, Meta, and Microsoft are some of the largest and most noteworthy, but the category spans other cloud computing providers, search engines, content streaming services, gaming platforms, and enterprise service providers, including companies outside the U.S. such as Alibaba, ByteDance, Naver, and Tencent.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

AI was enabled by hyperscale, not so much by their scale-out computing infrastructures, but more so by their extraordinary access to data. The largest hyperscalers have billions of users across multiple platforms, constantly generating data that can be fed into AI training models.

As much as these organizations were already accustomed to provisioning and managing large-scale computing, AI brought about a new challenge. While applications like content streaming, search, and social media are all manageable on loosely coupled, scale-out clusters, AI required much more computational power, tightly coupled in dense configurations.

The most critical enabling technology has been GPUs. Originally designed for graphics – what the G in GPU stands for – GPUs have been deployed as computational accelerators for generations of high-performance computing (HPC) applications. An Intersect360 Research survey found that 89% of users deployed accelerators – usually GPUs – as part of their HPC environments.[1]

Having already been established in HPC, GPUs proved to be an ideal architectural fit for large-scale AI training and inference, and hyperscale organizations have been consuming them as fast as providers can churn them out. (Nvidia is the dominant supplier, but AMD also has a competing GPU for HPC and AI, while other companies, including Intel, have specialized processors that stand in as alternatives.) Today, the top-spending hyperscale organizations are spending tens of billions of dollars annually on AI infrastructure.

These high-density GPU clusters came with a new pair of related challenges: power and cooling. GPUs tend to be power-hungry components, and putting as many as eight of them into a single server node can lead to individual server racks that draw 100 kilowatts, or more. And of course, the more power the chips consume, the hotter they tend to run, creating tremendous challenges for cooling at each level: component, node, rack, and data center.

On-Premises Enterprise AI Heats Up

AI is not exclusive to hyperscale. Today, most large organizations have designs to benefit from "enterprise AI," a general term that encompasses a grab bag of possible benefits, such as streamlined operations, lower costs, higher product quality, or enhanced customer service. A 2024 survey of U.S.-based large corporations found that 95% had budgets for HPC, AI, or (usually) both. (See Figure 1.)

Early rounds of enterprise AI deployment concentrated mainly into companies with HPC experience. Beyond the simple fact that majority of large companies have reason to do modeling and simulation – major industries such as manufacturing, finance, energy, and pharmaceuticals are significant HPC consumers – these companies also had the experience and expertise to manage large scale systems, and most of them had already deployed GPUs

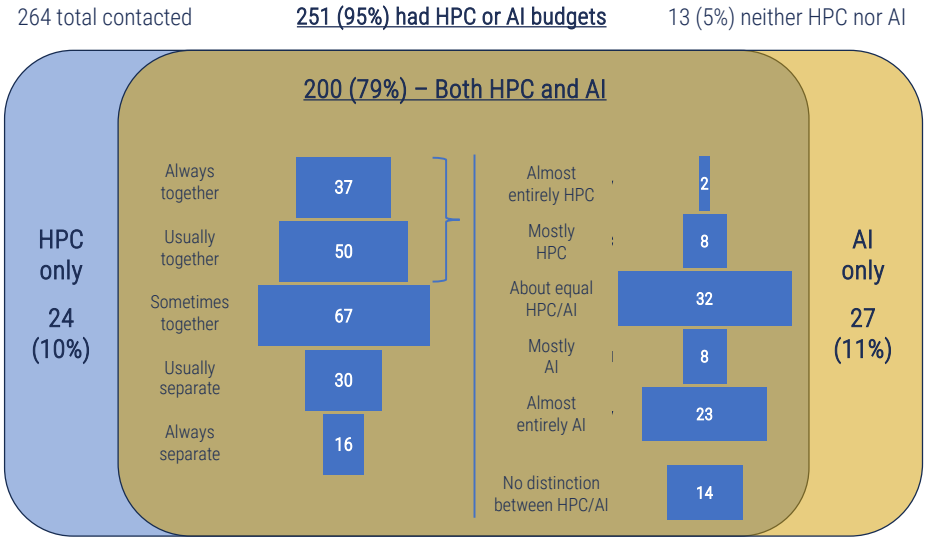
A 2024 survey of U.S.-based large corporations found that 95% had budgets for HPC, AI, or (usually) both.

[1] Intersect360 Research HPC-AI Technology Survey, 2024.

for at least a portion of their HPC environments. When other organizations outside of HPC invested in AI, they tended toward cloud computing for the resources. This also made logical sense; the hyperscale companies that offered cloud services had the resources for AI, and preexisting cloud computing contracts could be extended to cover them.

Figure 1: Budget Overlap of HPC and AI

Intersect360 Research Budget Map Survey of Large U.S.-based Companies, 2024



In 2024, as generative AI took hold and AI deployments accelerated, we began to see a new trend toward more on-premises AI deployments, often independent from HPC. For example, more organizations became interested in designing their own custom large language models (LLMs), often for internal usage. (See Figures 2 and 3.) There are multiple arguments in favor of moving these deployments on-premises. Beyond the straightforward cost of cloud computing, the notion of data locality – often influenced by regulatory factors – leads many companies to possess and control their own data, without the cost and latency of moving it in and out of the cloud.

Figure 2: Generative AI Adoption Considerations

Intersect360 Research HPC-AI Software Survey, 2024

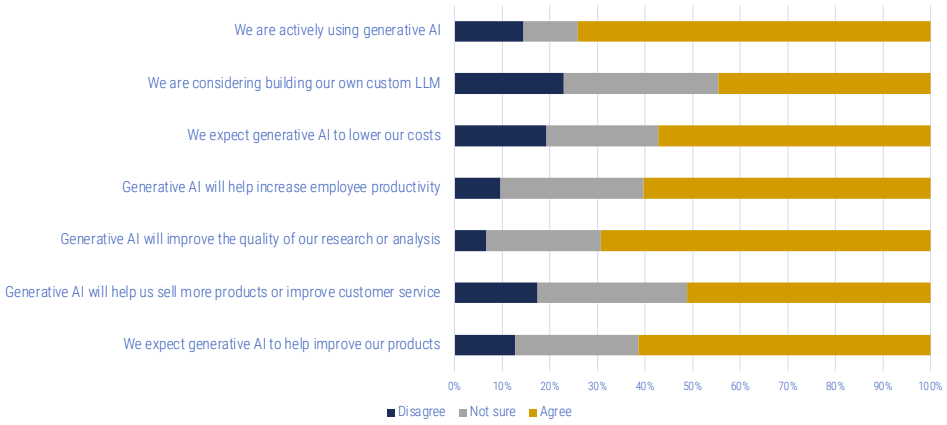
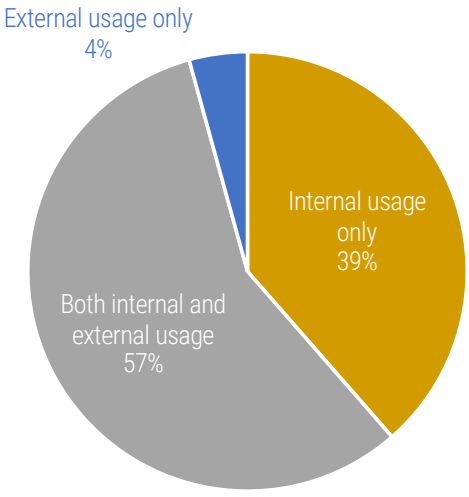


Figure 3: Planned Deployments of Large Language Models (LLMs)

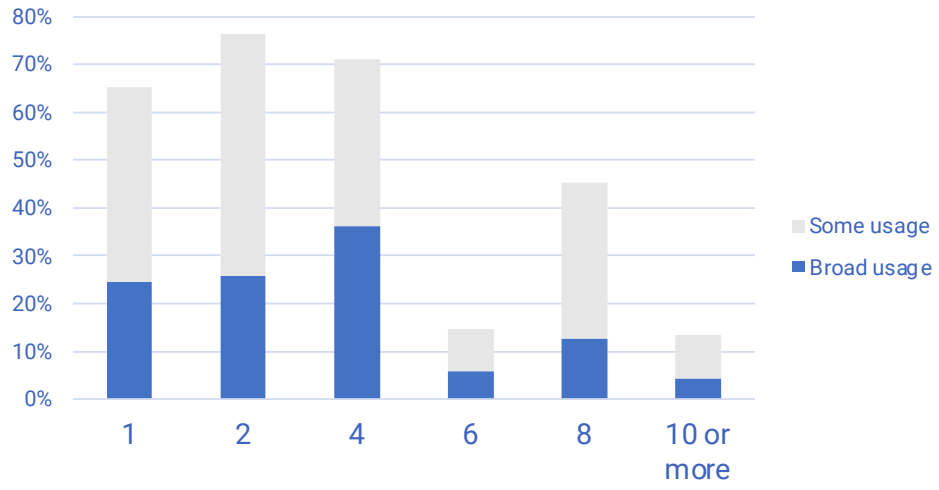
Intersect360 Research HPC-AI Software Survey, 2024



This combination of trends brings the facilities challenges first experienced by hyperscale companies into the everyday enterprise. Those with HPC experience are finding their GPU densities rising from the needs of incorporating AI workloads. Four GPUs per node is now the most common HPC-AI configuration in "broad usage," and configurations with six, eight, or more GPUs are deployed with increasing frequency. (See Figure 4.) Those without HPC experience are likelier to dive directly into 8 GPU configurations designed exclusively for AI.

Figure 4: Accelerators per Node for HPC-AI

Intersect360 Research HPC-AI Technology Survey, 2023



Unlike hyperscale companies, most other organizations have neither the money nor the inclination to build new data centers from the ground up, hundreds of megawatts at a time. With AI moving into the on-premises enterprise data center, the biggest question is how the power and cooling challenge can be managed.

Even among those with HPC experience, most racks and data centers are air-cooled. Liquid cooling, at the rack or room level, is less common (see Figure 5), but many have been convinced (or at least told) that liquid cooling will be the only viable option for the GPU clusters required for AI. Among organizations that do not currently deploy rack-scale liquid cooling, 68% say they "would need to significantly upgrade facilities" to incorporate it. 36% say they are looking at liquid cooling as a "pilot project."

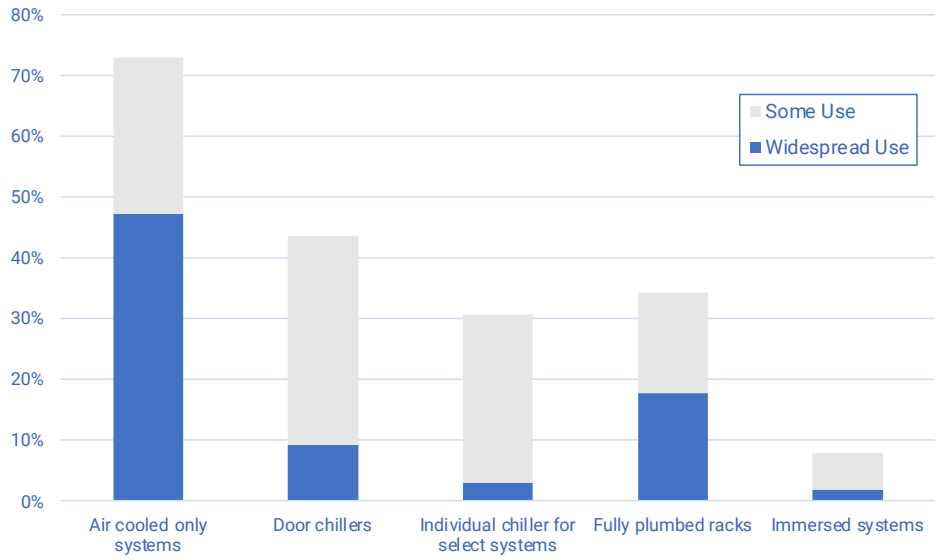
Most companies do not already use liquid cooling in the data center.

Among those that do not, 68% say they "would need to significantly upgrade facilities" to incorporate it.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Figure 5: Air Cooling Only vs. Various Liquid Cooling Options, Among HPC Users

Intersect360 Research HPC-AI Technology Survey, 2023



These would be sufficient challenges without any external pressures, but for many companies, all this additional planned power consumption comes against a backdrop of environmental governance strategies to be greener. (Consuming electricity in the cloud instead of on-premises does not "save" electricity any more than it would use less water to shower at the gym instead of at home.) Datacenter solutions for AI not only have to power and cool systems at scale, but also handle electricity, water, and waste heat as efficiently as possible.

Being convinced to invest in AI isn't the hard part. Envisioning exactly what an enterprise can achieve with AI is harder, and it is a major challenge to specify what infrastructure will be needed to support AI over even a moderate, ten-year horizon. The greatest challenge of all, therefore, is addressing how to get there, starting from here. Managing risk is the mantra of most CIOs or CTOs – even more than enabling new capabilities. At this dawn of the AI era, enterprises don't need to be sold on AI as much as they need lower-risk ways to get there.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

INTERSECT360 RESEARCH ANALYSIS

DDC Cabinet Solutions: Modular Answers for the Modern Enterprise

As AI has taken off, it has created a need for facilities experts who specialize in high-performance, high-density computing. For ten years, DDC Solutions has done exactly that, with over 1,100 global deployments of its cabinet technology for enterprise environments. Beginning in 2024, the company began focusing exclusively on the high-density solutions, driven by the move toward enterprise AI.

DDC's founders were previous enterprise data center operators themselves, and if moving toward AI solutions seems obvious, DDC has distinguished itself by challenging assumptions that introduced risk. For example, it is common knowledge that AI will require liquid cooling, and organizations will simply have to overhaul their facilities to incorporate it. Maybe so. Maybe not. This is where DDC's expertise comes in.

True to its focus, DDC recently introduced the S-Series, the newest purpose-built solution in the company's lineup. The DDC S-Series can handle up to 100 kilowatts per rack, *with air cooling*. The ultra-high density air cooling solution opens the possibility for organizations to move to many AI-optimized configurations, without the need to convert to liquid cooling. Where AI is in its early stages of deployment, this is a tremendous factor in reducing cost and risk.

Of course, nothing is guaranteed. Perhaps an organization's next generation of AI clusters will demand even more power. What if that liquid cooling is needed after all? The same S-Series cabinets have liquid-to-chip cooling capability built in, supporting up to 400 kilowatts per racks. The liquid elements are completely self-contained, with no water waste.

It's common knowledge that organizations will simply have to overhaul their facilities to incorporate liquid cooling. Maybe so. Maybe not.

This is where DDC's expertise comes in.

Figure 6: DDC Solutions S Series Cabinet Features

Source: DDC Solutions

These S-Series racks come in a variety of form factors, either 30-inch or 36-inch width, with various heights supporting a range of rackmount units (RMUs). They optimize PUE (Power

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Usage Effectiveness) – exact figures will depend on the specific deployment – a key figure in measuring and optimizing the power efficiency of any data center deployment. And, true to a data center operator's heart, they check the necessary boxes for safety and regulations, with NEMA 3R certification, UL listing, fire suppression, sound attenuation, and biometric access control. If the sprinklers in the building go off, no worries, your AI infrastructure is safe; the cabinets are completely sealed. (See Figure 6.) Perhaps most importantly, DDC's S-Series solutions are completely modular, scaling from one to 1,000+ cabinets, depending on organizational needs.

As these high-density systems are introduced into the enterprise, monitoring and management are essential. As a key part of its S-Series solutions, DDC provides DCIM (Data Center Information Manager) software for dynamic management and real-time monitoring. DCIM consolidates tools for managing and monitoring power, cooling, and environmentals, and it also dynamically adjusts airflow and water temperature as workloads increase and decrease to keep each cabinet with optimal conditions. (See Figure 7.) The dashboard and visualization tools in DCIM give data center operators a command center for everything from carbon footprint optimization to fault detection, all part of the peace of mind in managing computing and data resources at scale.

Figure 6: Features of DDC DCIM: Data Center Information Manager

Source: DDC Solutions

For organizations looking to deploy enterprise AI but seeking to mitigate cost and risk along the way, DDC Solutions provides a potential lifeline. Convert to liquid cooling later, and only if you need to. Add racks modularly, not all at once. Maintain the regulations and environmental goals of existing data center facilities. And do it all with improved monitoring and control. If AI is the revolution that's coming, this is the way many enterprises will want to get there.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com