
Aligned in Northern Virginia

THE FUTURE IS FLEXIBLE

Get (and stay) ahead with agile designs

By Michael Welch, Vice President, Procurement and Design, Aligned Data Centers

The growing significance of next-generation computing and applications in our personal and professional lives going forward is immense.

Let's consider AI. The generative AI segment in particular is poised to explode, thanks to an increase in the use of deep learning (DL) training and machine learning (ML) inference models across an ever-expanding range of business and consumer applications. According to Bloomberg Intelligence, generative AI will grow to 1.3 trillion USD over the next 10 years from a market size of just 40 billion USD in 2022. One of the largest drivers of incremental revenue will be generative AI infrastructure-as-a-service (247 billion USD by 2032) used for a variety of business applications. Many other projections from credible sources are similarly bullish.

The growth of AI and other next generation, high-performance applications will fundamentally alter the design and operation of data centers. Faster, hotter chipsets, larger servers, cluster networks, and rising densities will push the limits of legacy data center infrastructure and designs.

There's a common belief that next generation infrastructure requires immense capital investment and complete data center overhauls. At Aligned, we're taking a slightly difference approach: Stay ahead of product innovation and make agility innate in your design. The future is flexible.

IT'S GETTING HOT (AND DENSE) IN HERE

Let's go back to AI to provide some background. At data centers all over the world, ML training workloads are being used to build various AI models. These models can perform many types of tasks, including translating languages and chatbot conversations. Not only can these models understand complex textual data, but they can also generate and respond with new text that is coherent and grammatically correct.

Training workloads require massive amounts of data fed to specialized servers with processors known as accelerators. A graphics processing unit (GPU) is an example of an accelerator. Along with servers, training also requires data storage and a network to connect it all together. These elements are assembled into an array of racks known as an AI cluster.

The larger the AI training model, the more accelerators are required, and rack densities in large AI clusters can range from 30 kW to 100+ kW , depending on the GPU model and quantity. Clusters are generally limited to a handful of racks by connectivity limitations but can be expanded based on the GPU type and application.

Once AI models are trained, they are ready to be utilized for various inference applications. Inference signifies that the previously trained model is engaged in production to predict the output of new queries. Training is the necessary building block for inference, whether the application is ChatGPT, a recommendation engine suggesting your next Netflix binge, or an autonomous vehicle.

Consider that a driverless vehicle's system would first have to be trained on the rules of the road. This is a crosswalk; that is a stop sign. But when driving in traffic, the car itself would need to make split-second inferences based on the dynamics of its surroundings. Inputs and outputs. Due to the nature of these inference models being business critical they require mission-critical type resiliency.


Racks in an Aligned data center

There's a common belief that next generation infrastructure requires immense capital investment and complete data center overhauls. At Aligned, we're taking a slightly difference approach: Stay ahead of product innovation and make agility innate in your design. The future is flexible.

The GPU and tensor processing unit (TPU) clusters essential to running AI—as well as high-performance computing and ML applications—requires massive processing power and extremely fast cluster communication, which pushes densities far beyond the average kW per rack. Historically, lower densities could be cooled with air. However, as densities continue to rise, the industry will see a significant shift to liquid cooling, starting with air-assisted liquid cooling (AALC)—a hybrid of air and liquid—and eventually moving to entirely liquid-cooled for ultra-high-density deployments. Liquid cooling can

A close-up of Aligned's trademarked Delta³ cooling technology

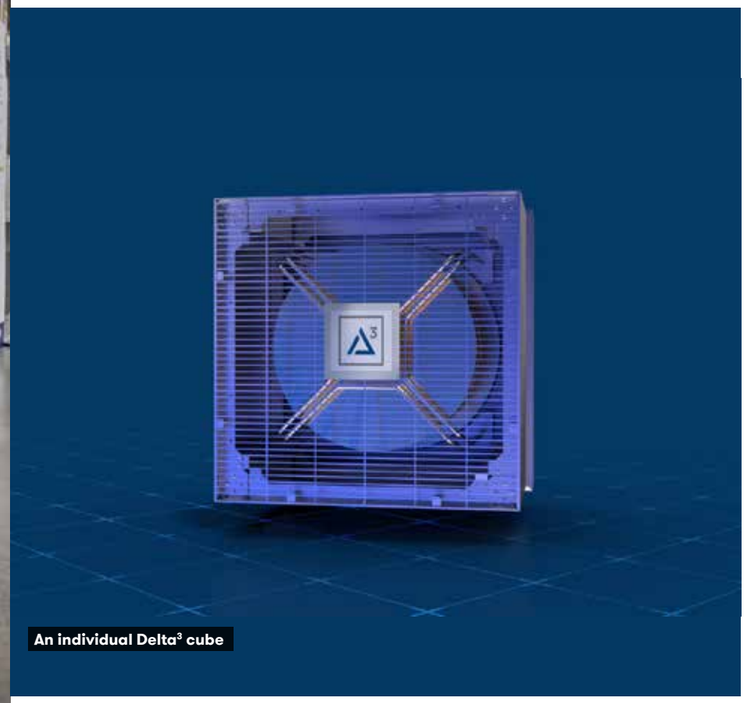
dissipate heat up to 1,000 times more efficiently than air. Data center designs should be flexible enough to allow customers to make the transition from air to liquid, or deploy a hybrid of both, with minimal risk and lift—especially in a live environment.

LIQUID, WITHOUT THE LIFT

These significant changes to power and cooling will pose significant challenges for more rigid designs and legacy infrastructures, which will likely require massive rebuilds to existing data centers and completely new designs for facilities purpose-built to house higher-density applications.

However, data center rebuilds in legacy facilities face several hurdles, including physical site constraints, a lack of standardized designs and components / equipment, staff inexperienced with operating the newly installed infrastructure, and changing customer requirements for specific liquid technologies (direct-to-chip, rear-door heat exchangers, immersion, etc.). Plus, with new products launching continuously, the risk of cooling and design obsolescence runs high as thermal profiles continue to rise. The same can be said for building new dedicated facilities with a rigid design—just tack on added time to market, CapEx, and being locked-in to supporting that specific application until the next industry shift.

The key to supporting the workloads of the future is designing for change and scale. For existing deployments, that means the ability to grow cooling capacity incrementally to support rising densities and temperatures while maintaining existing water temperatures and unchanged power requirements—all in in your existing environment,


An individual Delta³ cube

GROW WITH THE FLOW

The key to supporting the workloads of the future is designing for change and scale. For existing deployments, that means the ability to grow cooling capacity incrementally to support rising densities and temperatures while maintaining existing water temperatures and unchanged power requirements—all in in your existing environment, and all done without any downtime for your deployment. For new data center projects, that means the ability to support a myriad of high-density applications and expand on demand when your business requires it.

Customers come to Aligned for scalable infrastructure and flexible cooling technologies that are designed to meet the increased density requirements and rising temperatures of next-generation technologies. That means data center designs that can seamlessly pivot and scale in support of shifting compute environments, no matter the application, density requirement, or cooling solution.

A DECADE OF COOLING INNOVATION

Aligned has been innovating cooling technologies for more than a decade, from the days when rack densities topped out at 30kW per rack to now, when we provide customers with a pathway to transition from air to liquid or hybrid cooling with proven solutions to cool more than 100kW per rack.

To design effectively for change, our teams work closely with leading server and chip manufacturers as well as our customers to ensure we're solving both for what's now and what's next. With the right balance of turnkey and customization, we can provide customers with flexible data center designs within a standardized delivery process to streamline deployment while reducing cost impacts, time delays, and risk.

With our newly launched DeltaFlow™ liquid cooled and Delta³™ air-cooled technologies, our customers can make the move from air to liquid, or deploy hybrid cooling systems, in the same data hall. There is no need for massive rebuilds of existing infrastructure.

Designed with efficiency in mind, our cooling technologies leverage Aligned's standard closed loop system, using no outside air or water for heat rejection. This universal architecture can support all current liquid cooling solutions—liquid to the chip, liquid to the rack, or liquid to the tank. DeltaFlow™ easily integrates with Delta³™, which means unchanged power requirements and compatibility at existing facility water temperatures.

Aligned is redefining the future of flexible data center design to support the requirements of next-generation workloads.

ABOUT THE AUTHOR

Michael Welch is a forward-thinking executive with over 13 years' experience in data center engineering and design. As VP of Procurement and Design for Aligned Data Centers, he is responsible for leading all design, pre-construction, strategic sourcing, and supply chain management for the company's data center projects across the Americas.

Prior to his current role, Welch held numerous leadership roles at Aligned, including Senior Director of Pre-construction & Delivery, Director of Platform Procurement, and Product Development Engineer. He was also instrumental in the launch of the company's proprietary and award-winning Delta³™ air-cooled and DeltaFlow™ liquid-cooled technologies.

Welch holds an M.S. in Engineering Management from Ohio University as well as a B.S. in Environmental Engineering from the University of Connecticut.