iceotope

WHITE PAPER

From Cloud Cost Crisis to Deterministic Performance:

Iceotope Precision Liquid Cooling for Edge HPC Workloads

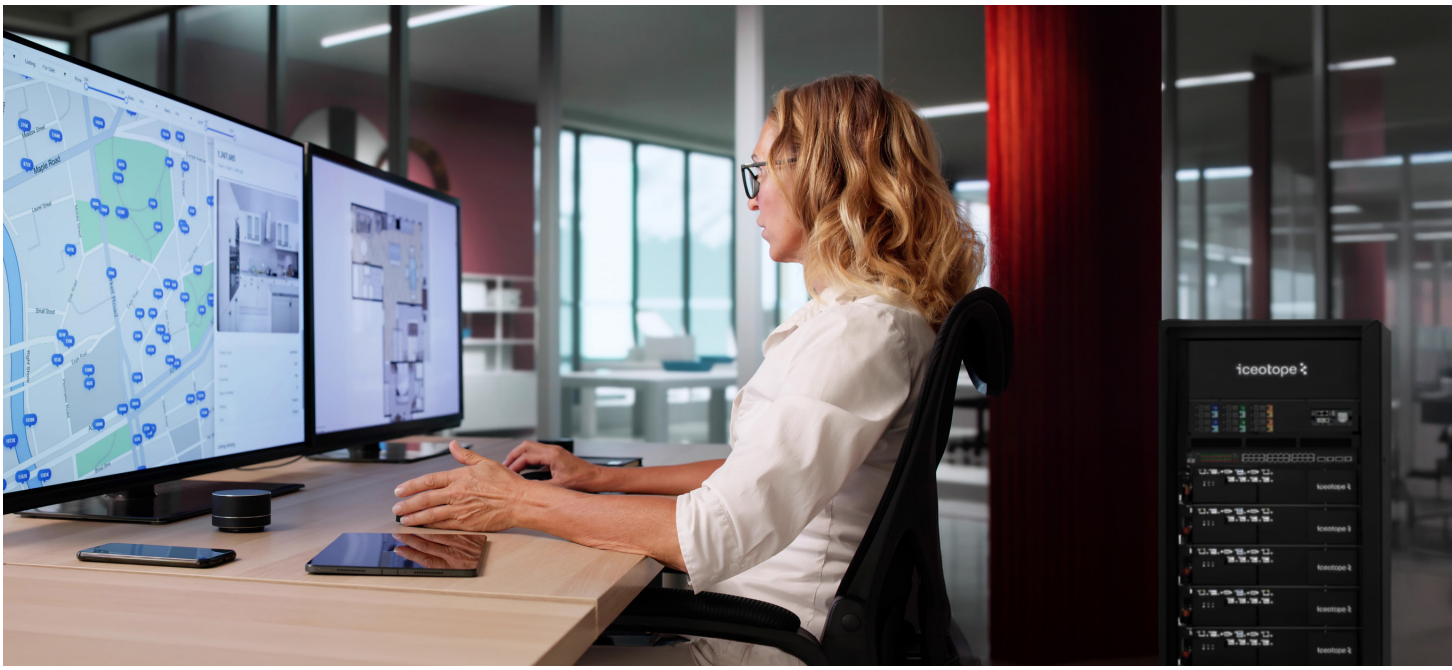
By Ian Ferguson

Executive summary

Enterprise HPC and AI workloads are hitting a cloud cost and data gravity wall. As simulation, rendering, and inference pipelines scale, organizations face runaway GPU IaaS bills, egress charges on multi-terabyte datasets, and tightening compliance regimes that make the public cloud increasingly unattractive for core workloads.

Iceotope precision liquid cooling technology supports liquid-cooled infrastructure platforms for PCIe GPU clusters in customer-owned datacenters and colocation facilities, delivering 1.7× higher performance per watt than air-cooled alternatives, up to 2× GPU density per rack, and predictable 3–5-year TCO. For sustained HPC and edge HPC workloads—CFD, FEA, rendering, and real-time analytics—Iceotope technology eliminates thermal throttling, reduces cooling overhead by 40–50%, and unlocks 10–60% savings versus cloud GPU IaaS, while keeping sensitive data on-premise.

This paper sets out the economic and technical case for liquid-cooled compute solutions across three dimensions: the shifting economics of HPC and cloud compute, the physical limits of air cooling at modern GPU power densities, and the practical TCO and performance advantages available to organisations that deploy on-premise infrastructure today.



1. The new economics of HPC and edge HPC

1.1 INFERENCE ERA, HPC REALITY

By Q1 2026, inference spending in the cloud surpassed training for the first time, reaching 20.6 B USD (55% of AI cloud infrastructure spend) and is projected to reach 70–80% of total AI compute costs by the end of 2026. While this inflection point is discussed mainly in the context of LLMs, the underlying economic shift directly affects HPC users who increasingly integrate AI into simulation and engineering workflows.

For most production AI systems, inference accounts for 75–90% of lifetime cost; training is an episodic capital event, but inference—alongside traditional HPC jobs—drives a recurrent cost with every query, transaction, or frame rendered. Enterprises now report monthly cloud inference bills growing 100–200% year-on-year, with multi-million-dollar exposure locked into long-term cloud agreements. In parallel, classic HPC workloads (CFD, FEA, rendering) are generating ever more local data, making egress-based cloud models increasingly unsustainable from a financial standpoint.

The implication for HPC infrastructure buyers is clear: the unit economics of cloud GPU compute, already becoming hard to justify for sustained high-utilisation workloads, are deteriorating further as demand outpaces supply as inference surcharges become embedded in platform pricing. On-premise liquid-cooled infrastructure is no longer a niche choice for the cost-conscious; it is increasingly the rational default for any organisation running HPC at meaningful scale.

1.2 DATA GRAVITY AND EGRESS: WHY HPC IS COMING HOME

For many engineering and media environments, the dominant constraint is no longer raw compute, but data gravity and IP protection. Workloads such as high-resolution CFD, structural FEA, video rendering, and design-space exploration routinely generate or consume hundreds of gigabytes per day at a single site. At these scales, cloud egress costs alone can justify on-premise deployments.

In regulated sectors, data cannot easily be exported to shared cloud infrastructures: ITAR/EAR constraints, HIPAA, GDPR and sector-specific sovereign-AI mandates often require that critical datasets remain in customer-owned facilities or trusted colos. HPC centers must therefore deliver high sustained performance while maintaining strict data residency, which in turn pushes them toward dense, efficient on-premise GPU clusters rather than burst-only cloud models.

Iceotope precision liquid cooling platforms are designed for precisely this environment: sovereign, high-density, thermally efficient compute that sits inside the customer's own four walls.

2. Why traditional air-cooled HPC struggles at scale

2.1 THERMAL THROTTLING UNDER SUSTAINED LOAD

Conventional air-cooled HPC nodes were not designed for sustained 100% GPU utilization at modern power envelopes. As GPU TDP rises and racks approach 20–30 kW, many air-cooled deployments exhibit 15–25% performance loss on long-running jobs due to thermal throttling and conservative power limits set to protect hardware.

For HPC users, this means that the theoretical peak performance of a cluster is rarely delivered in practice on multi-hour CFD or rendering runs. Not only does this elongate job turnaround times but it also inflates effective cost per simulation, since more energy and calendar time are consumed per completed workload.

In competitive engineering environments where simulation cycle time is directly linked to product development velocity, this hidden performance tax has material business consequences that go well beyond the energy bill.

2.2 POWER, DENSITY, AND FACILITY LIMITS

Air cooling also constrains density. Typical air-cooled GPU racks are limited in GPU count and rack-level power densities to maintain acceptable inlet temperatures and airflow. As a result, organizations quickly run into space and power constraints: more racks, more floor space, higher PUE (1.4–1.6 is common), and higher operational costs.

At the same time, many enterprise datacenters and colo suites cannot be rebuilt from scratch to accommodate hyperscale-style aisles or exotic mechanical plant. HPC teams need a way to double GPU density and improve efficiency within existing facility envelopes, without multi-year construction projects or wholesale architectural changes.

2.3 THE UPGRADE TRAP: NEXT-GENERATION GPU TDP

The thermal challenge is not static. Next-generation GPU platforms from Nvidia and AMD are tracking toward 1-1.5 kW TDP per card. Air-cooled facilities designed around today's 300-700 W GPU envelopes will face a binary choice when they upgrade hardware: accept severe throttling or undertake costly mechanical plant upgrades. Iceotope precision liquid cooling eliminates this upgrade trap: the cooling architecture scales with GPU TDP without requiring facility reconstruction.

3. Iceotope: engineered for enterprise PCIe HPC

3.1 PURPOSE-BUILT FOR PCIE GPU CLUSTERS

Iceotope liquid cooling technology is ideally suited for on-premise PCIe GPU clusters in the 8–64 GPU range per deployment, with typical enterprise clusters scaling to dozens of GPUs per rack and multiple racks per site. It targets the 60–70% of deployments that are PCIe-based rather than SXM, matching the dominant form factor in enterprise and edge environments through at least 2027–2028.

Instead of chasing hyperscale “AI factory” designs that ship as complete SXM racks with integrated liquid cooling and no aftermarket opportunity, Iceotope focuses on the retrofit-eligible, PCIe-centric enterprise and colo segment—where customers own the racks, choose the GPUs, and control the facility.

This positions Iceotope liquid cooling solutions as a vendor-agnostic infrastructure layer: as GPU generations evolve, the liquid cooling architecture remains constant, protecting the customer’s facility investment across hardware refresh cycles.

3.2 PRECISION LIQUID COOLING FOR SUSTAINED PERFORMANCE

By bringing precision liquid cooling directly to PCIe GPU servers, Iceotope enables sustained 100% utilization without thermal throttling—the Achilles’ heel of air-cooled HPC nodes. Liquid cooling delivers approximately 1.7× better performance per watt compared to air-cooled alternatives, by keeping GPUs at optimal operating temperatures even under continuous load. An additional advantage is that by eliminating external air, external contaminants that degrade systems are also kept out of the system.

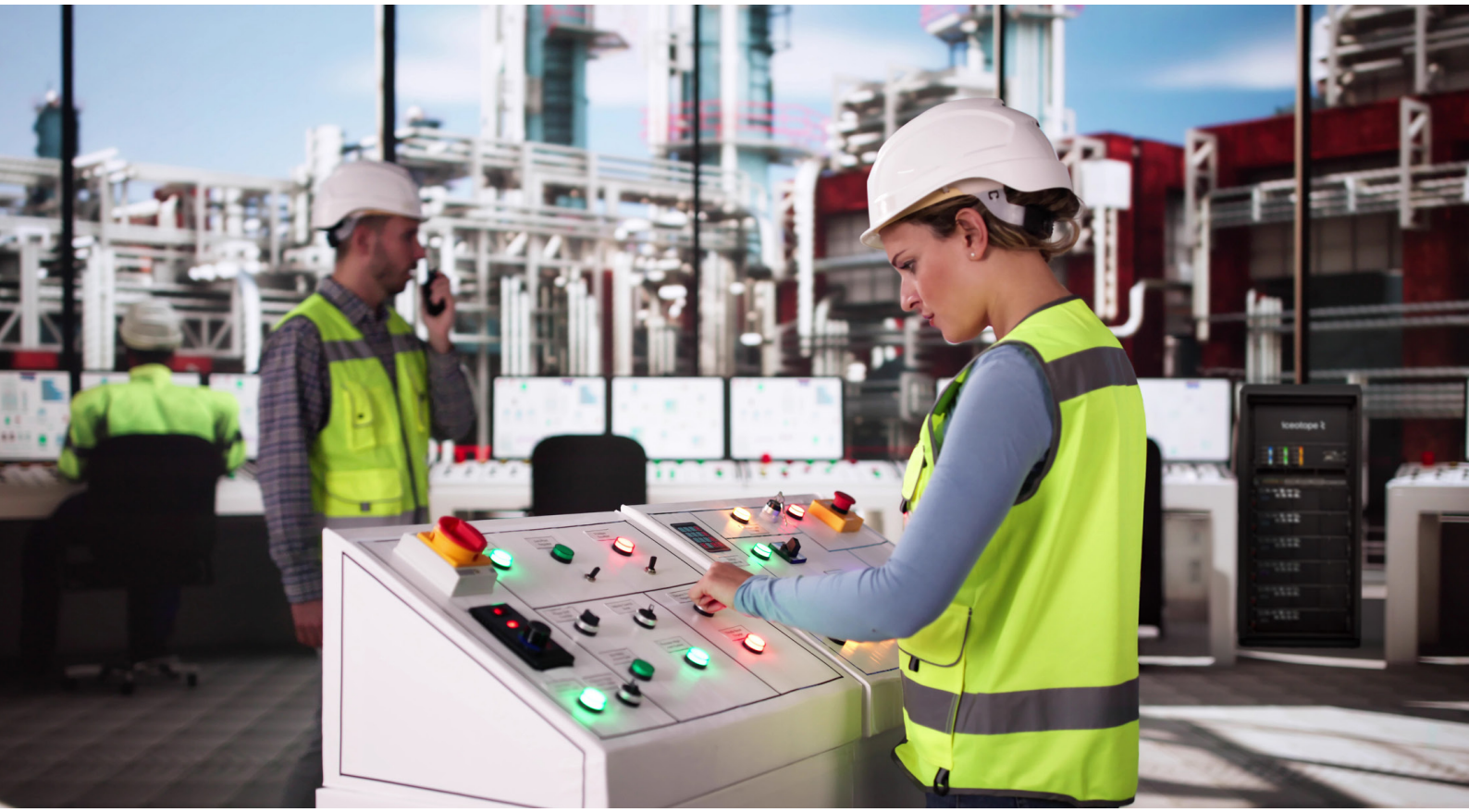
This improvement compounds across large HPC workloads: a cluster that avoids 15–25% performance loss can complete more simulations per day, improve engineering iteration speed, and increase the effective ROI of each GPU purchased. For edge HPC scenarios, such as manufacturing quality control or real-time engineering analysis, this means consistent latency and throughput even in thermally constrained sites.

3.3 RETROFIT-COMPATIBLE, FULL-STACK SOLUTION

Iceotope liquid cooling platforms are designed to integrate into existing datacenter and colo environments without requiring a facility rebuild. Iceotope enables a full-stack, turnkey solution spanning liquid-cooled servers, manifolds, coolant distribution units, power and networking integration, rather than isolated cooling components that customers must assemble themselves.

This approach reduces integration risk, shortens deployment timelines, and ensures that HPC teams can focus on workloads rather than fluid dynamics. With up to 2× GPU density per rack compared to air-cooled equivalents, and PUE improvements from typical 1.4–1.6 down to 1.2–1.3, our liquid cooling technology allows organizations to expand HPC capacity within existing power and space constraints.

Deployment timelines are measured in weeks, not quarters, enabling teams to respond to business demand without the capital planning cycles typically associated with facility upgrades.



4. HPC and edge HPC use cases

4.1 EDGE HPC COLOCATION

Edge HPC colocation is a distinct market niche beyond pure AI inference, where data gravity, sustained compute performance, and IP protection drive the economics. Typical workloads include video rendering and VFX pipelines in media and entertainment, CFD and FEA simulations in engineering and automotive, and high-resolution imaging and analytics in healthcare.

These workloads generate or consume large volumes of local data at a single site, making cloud egress costs alone a compelling argument for on-premise deployments. For these customers, Iceotope enables sustained 100% GPU utilization with no thermal throttling, dense, efficient PCIe GPU clusters located near data sources, and a secure, compliant environment for sensitive design data and intellectual property.

4.2 ENGINEERING SIMULATION AND DESIGN

Engineering teams running CFD, FEA, and multiphysics simulations are highly sensitive to job completion time and queue length. Liquid-cooled HPC clusters allow these teams to sustain high utilization and throughput across long-running jobs, which directly translates into faster design iteration and reduced time-to-market.

Iceotope liquid cooling technology supports a range of GPUs suitable for both traditional HPC and AI-enhanced simulations. Combined with data-gravity benefits—keeping CAD and simulation results on-premise—this yields a holistic performance and cost advantage over cloud or air-cooled clusters.

For automotive OEMs and aerospace primes operating under ITAR constraints, this is not merely a cost optimisation—it is a compliance requirement that Iceotope satisfies out of the box.

4.3 HYBRID AI + HPC ENVIRONMENTS

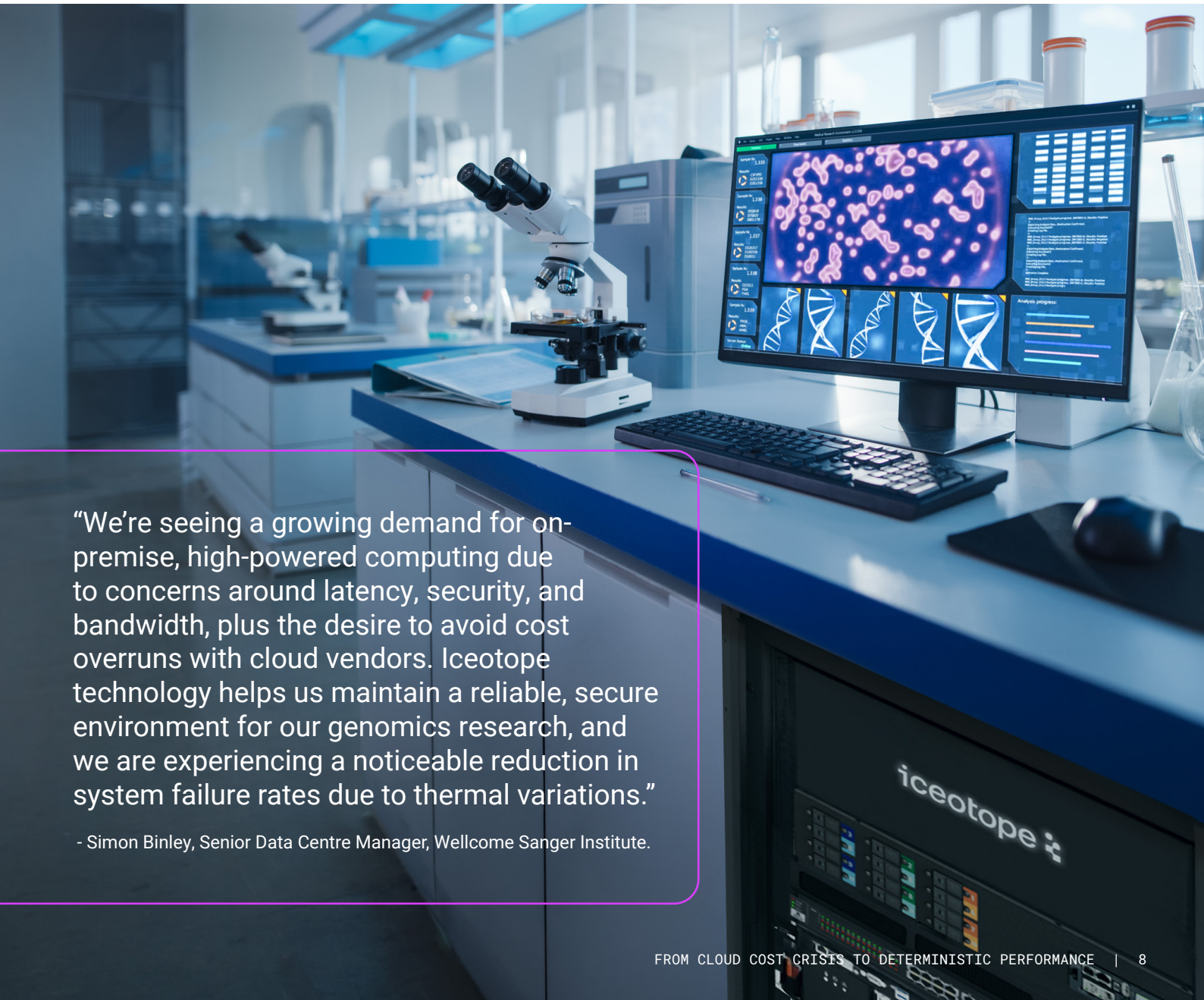
Many HPC centers are evolving toward hybrid environments where traditional simulation runs alongside AI workloads such as surrogate modeling, anomaly detection, or LLM-based post-processing. Iceotope's focus on GPU-based architectures makes its liquid cooling technology suitable for both these classes of workloads.

In these hybrid environments, enterprises can use the same on-premise cluster to run core HPC workloads at high utilization, serve AI models that accelerate or augment those workloads without incurring external token or egress fees, and maintain compliance and sovereignty by keeping both simulation data and AI inference on the same controlled infrastructure.

This convergence of HPC and AI on a single liquid-cooled platform also simplifies the infrastructure stack: rather than managing separate air-cooled HPC nodes and GPU inference servers, organisations run a unified, high-density cluster with a single operational model.

4.4 LIFE SCIENCES AND GENOMICS

Life sciences is an emerging high-value vertical for on-premise liquid-cooled clusters. Genomic sequencing pipelines, protein structure prediction (AlphaFold-class workloads), and clinical imaging AI all combine large-dataset locality requirements with sustained GPU utilisation profiles. HIPAA and GDPR constraints make cloud egress of patient-adjacent data legally fraught, while the compute intensity of these workloads makes cloud GPU pricing prohibitive at production scale. Iceotope's support for both sovereign on-premise deployment and sustained thermal performance directly addresses both constraints.



"We're seeing a growing demand for on-premise, high-powered computing due to concerns around latency, security, and bandwidth, plus the desire to avoid cost overruns with cloud vendors. Iceotope technology helps us maintain a reliable, secure environment for our genomics research, and we are experiencing a noticeable reduction in system failure rates due to thermal variations."

- Simon Binley, Senior Data Centre Manager, Wellcome Sanger Institute.

5. TCO and performance benefits versus cloud and air-cooled alternatives

5.1 VERSUS TRADITIONAL AIR-COOLED ON-PREMISE

Against traditional air-cooled on-premise deployments, precision liquid cooling offers lower cooling overhead, higher throughput per watt, higher GPU density per rack, and lower PUE. For HPC users, this translates into more simulations per rack, lower energy bills, and the ability to extend the usable life of existing datacenter space without structural changes.

Quantified: moving from a PUE of 1.5 to 1.2 on a 500 kW HPC cluster reduces annualised power overhead by 100 kW—at European commercial energy rates, this represents a six-figure annual saving that compounds across the asset lifetime. Combined with the elimination of thermal throttling, the effective cost per simulation hour on an Iceotope-cooled system is materially lower than an air-cooled equivalent even before factoring in density gains.

5.2 VERSUS CLOUD GPU IAAS

Compared with renting GPUs as IaaS, liquid-cooled on-prem HPC deployments can deliver significantly lower cost over a 3-year horizon for sustained high-utilization workloads. A comparable deployment converts variable OPEX into predictable CAPEX and controlled OPEX, eliminating token pricing volatility and egress charges while providing better performance per watt.

A useful framing for procurement teams: a single Nvidia H100 PCIe GPU rents on major cloud platforms at $2–4 per GPU-hour. At sustained 70% utilisation over three years, that equates to $12,000–$24,000 per GPU in cloud spend—typically two to four times the on-premise all-in cost including hardware, power, and cooling. For clusters of 32 GPUs or more, the cumulative delta is measured in millions of dollars.

5.3 SAVINGS PROFILE FOR HPC VERSUS AI INFERENCE

While AI inference workloads can see order-of-magnitude savings versus cloud APIs once volumes exceed certain thresholds, HPC and edge HPC workloads typically realize more moderate, but still material, savings versus cloud GPU IaaS. The value driver is different: HPC buyers primarily care about sustained throughput and performance consistency, with cost savings and egress avoidance as important but secondary benefits.

The most compelling TCO case for HPC buyers is therefore not a single savings headline, but a composite: lower cost per simulation hour (from elimination of throttling and lower PUE), lower egress cost (from on-premise data residency), and lower risk (from predictable CAPEX versus volatile cloud pricing). On-prem, Iceotope-cooled systems deliver all three simultaneously.

6. Qualification and positioning guidance for HPC buyers

6.1 WHEN ON-PREM IS A STRONG FIT

Iceotope liquid cooling technology delivers strongest value when three conditions are present. First, sustained GPU utilisation: workloads running at 60%+ utilisation over multi-hour or multi-day jobs, where throttling and cloud pricing variability have measurable cost impact. Second, high data gravity: datasets of hundreds of gigabytes or more per day generated or consumed on-premise, making cloud egress economically or operationally impractical. Third, IP or regulatory sensitivity: ITAR, HIPAA, GDPR, or sector-specific sovereign-AI mandates that constrain data export to public cloud environments.

6.2 WHEN CLOUD BURST STILL MAKES SENSE

On-prem is less suitable when workloads are bursty and unpredictable, dataset sizes are small with minimal egress sensitivity, or the customer has no existing IT footprint—no datacenter, colo, or operations team.

In practice, many HPC organisations operate a hybrid model: an on-prem cluster handles the sustained baseline workload on-premise, while cloud GPU capacity absorbs peak demand spikes that exceed local capacity. This architecture captures the cost and performance benefits of on-premise liquid cooling for the predictable workload majority, while retaining cloud flexibility for episodic peaks—without the risk of over-provisioning on-premise hardware.

6.3 DEPLOYMENT AND OPERATIONAL CONSIDERATIONS

Prospective buyers should assess operational factors alongside the financial case. Iceotope has developed an edge cabinet configuration with support for up to 32 GPUs that includes an external chiller with a sealed coolant connection. There is no open water loop to manage or integrate with existing facility infrastructure. The configuration can deployed in any location with adequate power and space.

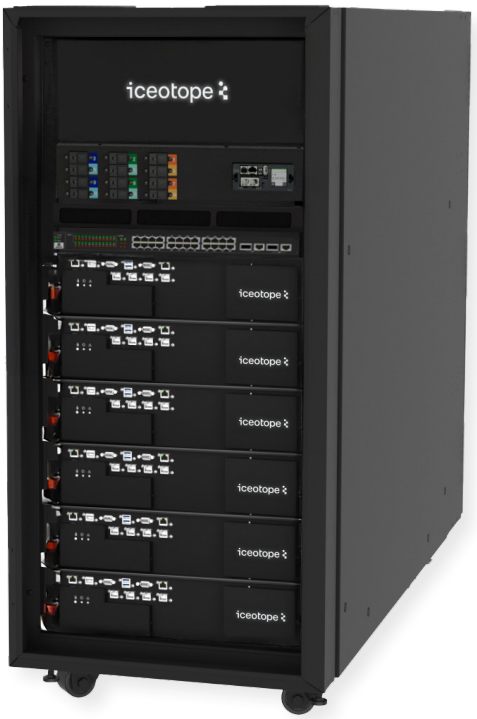
Conclusion

HPC and edge HPC users are under pressure from three converging forces: the shift of AI economics toward inference, explosive growth in data gravity and egress costs, and rising regulatory demands for data sovereignty and IP protection. Traditional air-cooled clusters and cloud-only strategies cannot simultaneously satisfy these constraints at sustainable cost.

Iceotope offers a pragmatic path forward: a liquid-cooled, PCIe-centric infrastructure platform that delivers sustained performance, improved density and efficiency, and predictable TCO for on-premise and colocation environments. For organizations running serious HPC workloads—and increasingly, hybrid AI + HPC pipelines—Iceotope turns the economics and operational profile of high-end compute from a cloud cost crisis into a controllable, high-performance asset.

The window to act is narrowing: as next-generation GPU TDPs exceed 1 kW and cloud GPU pricing continues to reflect constrained supply, the cost of deferring a liquid-cooling decision rises with every hardware refresh cycle. Organisations that deploy Iceotope liquid cooling technology today lock in the infrastructure foundation for the next five years of HPC and AI compute growth.

To discuss your HPC workload profile and model a precision liquid cooling deployment for your environment, contact the Iceotope team.



iceotope

AI is heating up.
We keep it cool.

SALES@ICEOTOPE.COM
WWW.ICEOTOPE.COM