FROM HOTSPOTS TO HEADROOM:

Phonronic Thermal Fabric™ Unlocks Performance, Efficiency, and Asset Life in AI Data Centers

Distributed Thermal Control Systems™ enable precision hotspot cooling across the data center, unlocking design degrees of freedom and delivering quick ROI

Ryan Brown, Product Director, Phononic

August 2026 V_1.0

Executive Summary

Thermal Control Systems™ for GPU HBMs:

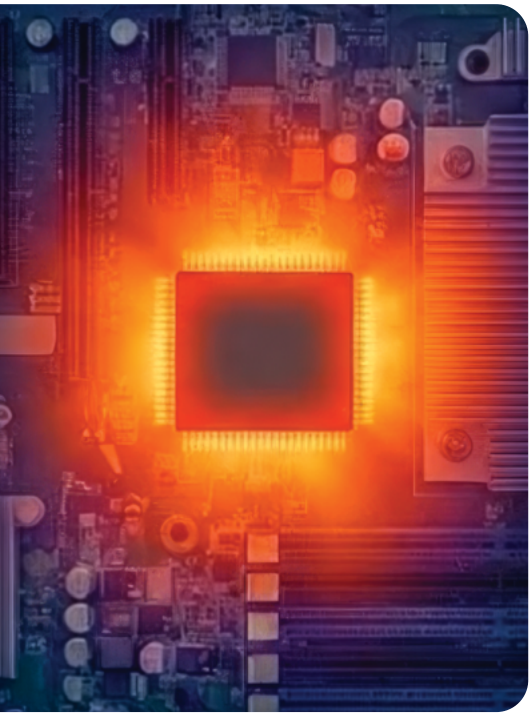
- Deliver precise, TEC-enabled hotspot cooling that avoids thermal throttling, extends useful device life and improves PUE up to 0.15
- Software-enabled thermal control layer delivers predictive cooling only when and where it is needed—boosting COP of 9+
- Facility-wide energy savings of up to 14%, allowing facilities to run water temperatures up to 50°C with no throttling or performance losses
- Deliver 3x ROI across the entire data center

AI has fundamentally transformed the data center into a high-density, high-variability compute platform, where thermal behavior is dynamic, workload-driven, and no longer safely abstracted through static design margins.

While effective at bulk heat removal, traditional air and liquid cooling lacks the responsiveness and spatial granularity required to manage localized hotspots and rapid thermal fluctuations inherent to AI workloads. As a result, operators are forced to overprovision cooling capacity, throttle performance, or accept accelerated asset wear—trade-offs that constrain both economic and computational efficiency.

Phononic is breaking this paradigm. When Thermal Control Systems™ are deployed at critical hotspots across the entire data center, this thermoelectric (TEC)- enabled, software-defined control layer is capable of millisecond-scale cooling adjustments, a 600x improvement over legacy times. The seamless integration of Phononic's Thermal Control Systems™ with system software and infrastructure management frameworks transforms cooling into a closed-loop optimization system—continuously balancing performance, power efficiency, and hardware longevity.

When deployed at the heart of AI compute, Phononic's Thermal Control Systems™ for GPU HBMs reduce thermal throttling, extend device useful life by up to 5x, improve PUE by up to 0.15, provide facility-wide energy savings of up to 14%, and deliver up to 3x return on investment. Data center operators can now mitigate thermal transients at the source, sustain peak GPU performance under burst workloads, reduce facility-level power overhead, and extend asset life.



Why Thermal is Now a First-Order Design Variable

AI workloads fundamentally change the thermal profile of the data center. Large-scale training and high-throughput inference are characterized by:

- Extreme power density at the silicon and package level (GPUs, HBM stacks, accelerators)
- Highly transient, burst-driven utilization patterns driven by dynamic scheduling and model execution phases
- Tightly coupled compute, memory, networking, and power delivery subsystems, where localized thermal conditions propagate system-wide performance impacts

High-bandwidth memory (HBM) exemplifies this shift. As bandwidth and stacking increase, HBM operates with minimal thermal margin. Temperature directly affects:

- Memory access latency and refresh overhead
- Throttling thresholds across the GPU-memory complex
- Signal integrity and error rates
- Long-term degradation and reliability

PHONONIC'S APPROACH COMBINES SOLID STATE THERMOELECTRIC COOLING AND CONTROL SOFTWARE THAT EMBEDS DYNAMIC, WORKLOAD-AWARE THERMAL CONTROL DIRECTLY INTO THE AI STACK

The Limits of Current (and Emerging) Cooling Approaches

The industry's transition from air to liquid cooling has delivered meaningful improvements in heat removal capacity and facility-level efficiency. Even liquid cold plate systems, while a clear advancement over air cooling, are still inherently reactive and coarse-grained. **They are designed to manage average thermal loads across devices, not the localized, millisecond-scale hotspots that drive performance behavior in modern AI systems.**

To be clear, what is at stake is mission critical. When High Bandwidth Memory (HBM) devices overheat, they enter thermally driven functional and reliability protection modes because elevated junction temperatures accelerate leakage rates, reduce retention time in DRAM cells, increase the probability of data corruption and significantly impact the silicon's life expectancy. As temperatures rises beyond specified thresholds (typically ~85–95 °C depending on the HBM generation), the memory controller and on-die thermal sensors cooperate to mitigate errors by increasing refresh activity—most notably through high temperature refresh modes such as "double refresh". While this preserves data integrity, it consumes additional bandwidth and power, reducing effective throughput and increasing latency due to refresh interference with normal accesses. Concurrently, thermal throttling mechanisms may be triggered at the stack or controller level, reducing command issuance rates or clock frequency cuts to limit further heating. If all of these standard protection strategies still fail to bring junction temperatures under control, even more aggressive protections will be initiated to protect the hardware, up to and including a full system shut-down. The challenge is significant and the cost unbearable.

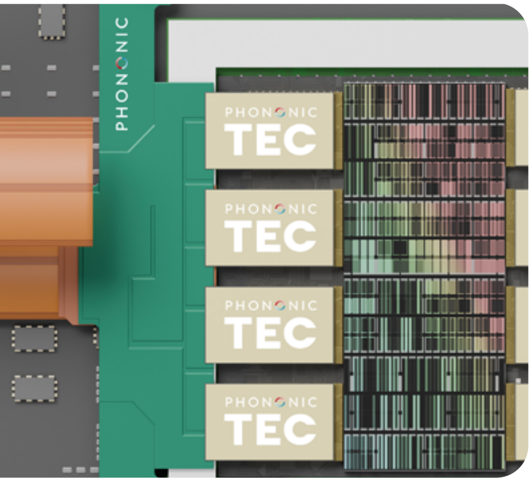
Introducing the Thermal Control Layer – Precision Solid State Cooling Where and When it is Needed

Phononic's approach combines solid state thermoelectric (TEC) cooling, Redfish-compatible control software, and modular, rapidly deployable hardware to deliver fine-grained thermal management at the point of need. Rather than treating cooling as a static facility function, Thermal Control Systems™ enable dynamic, workload-aware thermal control directly into the AI stack. When deployed throughout the data center, the deployed Thermal Control Systems™ form the underpinning of the Thermal Fabric™, data center-wide predictive thermal optimization and refined workload placement.

Unlike conventional approaches that maintain a fixed thermal envelope, Thermal Fabric™:

- **Delivers millisecond control of cooling, responding in real time to workload behavior:** Actively tracks and responds to workload-driven power fluctuations, mitigating transient hotspots before they trigger throttling or performance degradation
- **Applies precision cooling exactly where it is needed, when it is needed:** Targets high-intensity regions—GPUs, HBM, optics, and power delivery—eliminating inefficiencies associated with bulk or package-level cooling
- **Allows for real-time thermal telemetry and analytics, unlocking truly workload-aware orchestration:** Provides continuous thermal visibility and integrates with orchestration frameworks to enable workload-aware cooling optimization aligned with scheduling, utilization, and performance objectives

This architecture transforms thermal management from a passive constraint into an active, software-defined control plane—capable of dynamically balancing performance, efficiency, and hardware longevity in response to real-time AI workload demands.



Solving the Hardest Problem First: GPU HBM

High-bandwidth memory (HBM) has emerged as the primary thermal bottleneck in modern AI systems. As models scale and memory bandwidth becomes the limiting factor for training and inference throughput, HBM is increasingly the component that dictates achievable system performance. And the industry is taking notice, and striving for power efficient solutions that scale, and fast. JEDEC's recent unveil of the foundational LPDDR6 standard in July 2025 is the next in their ongoing roadmap of enabling higher densities while striving to reduce power per bit.[1]

This challenge is only intensifying with the transition to HBM4 and HBM5, where higher stack densities, wider interfaces, and increased power per bit drive disproportionate thermal sensitivity—all within already constrained package footprints. Unlike GPUs, which retain some dynamic headroom, HBM operates with minimal thermal margin, making it acutely sensitive to transient temperature excursions.

[1] JEDEC Solid State Technology Association. (2026, April 22). *JEDEC® previews LPDDR6 roadmap: Expanding LPDDR into data centers and processing-in-memory.*

Why HBM Becomes the Limiting Factor

HBM exhibits several characteristics that make it vulnerable under AI workloads:

- **Temperature-dependent refresh overhead**
  Above ~85 °C, DRAM refresh rates increase due to leakage, directly reducing effective memory bandwidth and increasing latency
- **Hard thermal throttling thresholds**
  Near ~100–105 °C, HBM triggers protective throttling, constraining GPU performance regardless of available compute capacity
- **Accelerated reliability degradation**
  Sustained high temperatures and rapid thermal cycling drive electromigration, TSV stress, and interconnect fatigue, reducing module lifespan
- **Tight coupling to GPU performance states**
  Thermal limits within HBM propagate across the GPU–memory complex, constraining boost clocks, interconnect efficiency, and overall system throughput

Memory-intensive AI workloads—large model training, attention-heavy architectures, and high-throughput inference—systematically drive HBM into worst-case thermal regimes. By some estimates, in fact, the reduction of throttling could prevent bandwidth degradation of up to 30-40%.[2]

These are exactly the conditions where conventional, bulk cooling approaches lack the precision and responsiveness to intervene effectively.

[2] Waseem, U. (2025, June 29). *What is High Bandwidth Memory 3 (HBM3): Complete engineering guide 2025*. Wevolver.

PHONONIC'S THERMAL CONTROL SYSTEMS™ CAN PROVIDE HIGHLY STABLE TEMPERATURE CONTROL IN STATIC OPERATING MODES, SIGNIFICANTLY IMPROVING THERMAL STABILITY BY AS MUCH AS 70%.

Phononic's Thermal Control Systems™ for GPU HBMs: Unlocking Performance

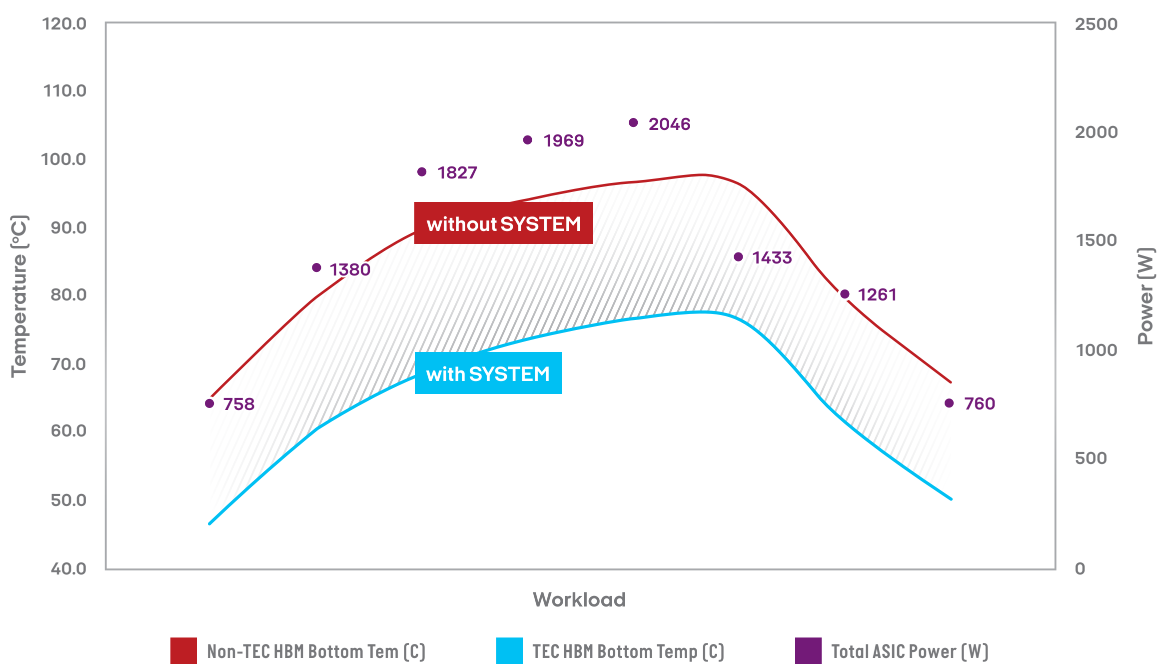
Solutions equipped with Phononic Thermal Control Systems™ demonstrate significantly improved thermal stability compared to non-TEC configurations, by as much as 70%. By actively managing transient temperature excursions at the source, Thermal Control Systems™ maintain HBM operation within optimal thermal bounds—preventing entry into temperature regimes that trigger refresh overhead increases or hard throttling.

Phononic's Thermal Control Systems™ can provide highly stable temperature control in static operating modes. As shown below, the most basic implementation is to establish a fixed temperature offset (ΔT), with the control package maintaining that differential across the workload to keep the device below the 85°C thermal ceiling.

For this evaluation, the test profile was defined as High GPU + Maximum HBM utilization with an inlet water temperature of 35°C, executed over a 20-minute test interval to represent a sustained, thermally demanding workload.

PHONONIC THERMAL CONTROL SYSTEMS™

Significantly improve thermal stability vs. non-TEC configurations



The outcome is clear:

- **Elimination or substantial reduction of thermal throttling events** under memory-intensive workloads
- **Sustained effective memory bandwidth** and lower latency
- **Improved GPU utilization** and overall system throughput consistency
- **Improved cooling efficiency (COP 1.4-1)**

Phononic's Thermal Control Systems™ for GPU HBMs: Unlocking Performance (continued)

While the static implementation provides a meaningful improvement in thermal performance, activating the embedded two-way telemetry, software, and firmware control layers within the Phononic Thermal Control Package delivers substantially greater gains.

Using the same test profile—High GPU + Maximum HBM utilization, a 35°C inlet water temperature, and a 20-minute sustained workload interval—the dynamic control mode continuously monitors and adjusts cooling performance in real time. Rather than maintaining a fixed thermal offset (ΔT), the system adapts to changing workload conditions, applying cooling capacity only when and where it is needed.

By leveraging this intelligent, closed-loop control architecture, the Thermal Control System™ moves beyond precision cooling to provide predictive, workload-aware optimization. This enables a more sophisticated approach to thermal management—one that not only suppresses temperature excursions as they emerge, but anticipates and mitigates them before they impact system performance.

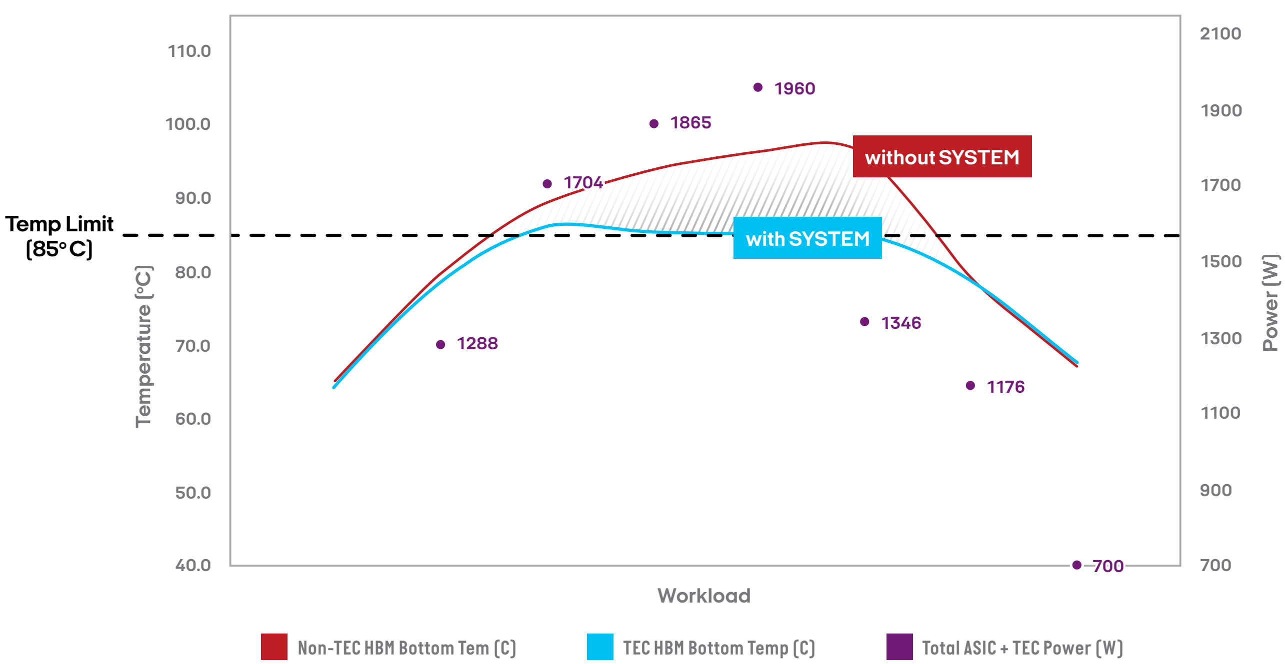
This intelligent control layer unlocks an additional dimension of value:

- **Proactive throttling avoidance:** Cooling is dynamically modulated ahead of thermal thresholds, maintaining both HBM and GPU operation within their optimal performance envelopes.
- **Energy-proportional cooling:** TEC power is applied precisely when and where required, eliminating unnecessary energy consumption during periods of lower thermal demand and improving overall system efficiency, with a measured **COP ranging from 1.9:1 to 9.6:1.**
- **Continuous optimization across performance and efficiency:** Real-time telemetry and analytics enable fine-grained balancing of compute throughput, thermal headroom, and power consumption.

The result is a fully adaptive thermal system that aligns cooling behavior directly with AI workload dynamics, maximizing performance while minimizing energy consumption.

PHONONIC THERMAL CONTROL SYSTEMS™

Predictive, workload-aware optimization, COP of 9+

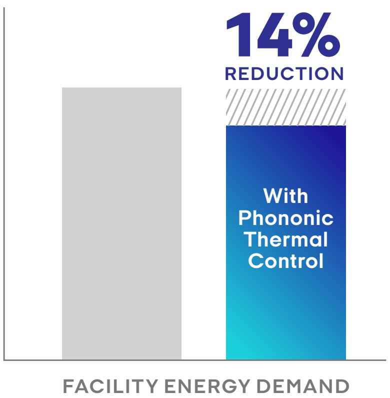


Thermal Control Systems™, Delivering Net Data Center Savings up to 14% in Energy Savings

These results validate that precision, real-time thermal control at the HBM level directly translates into recoverable performance and efficiency gains.

By reducing peak cooling demand through precision, real-time thermal management, rather than traditional cooling overprovisioning, the **Phononic Thermal Control System™ enables more efficient operation of the facility cooling infrastructure, resulting in a modeled net data center energy savings of up to 14%, allowing facilities to run water temperatures up to 50°C with no throttling or performance losses.** This analysis includes the power consumed by the Thermal Control Systems™ themselves, ensuring that the reported savings reflect total system-level performance.

This evaluation is intentionally conservative, focusing primarily on the direct impact of HBM thermal management. Additional system-level benefits are expected as precision thermal control is extended beyond HBM to other thermally constrained components and integrated into broader platform and facility cooling strategies. In particular, the ability to safely operate at higher facility water temperatures, while dynamically optimizing thermal resources across the platform, rack, and row, has the potential to deliver substantially greater improvements in overall data center energy efficiency.

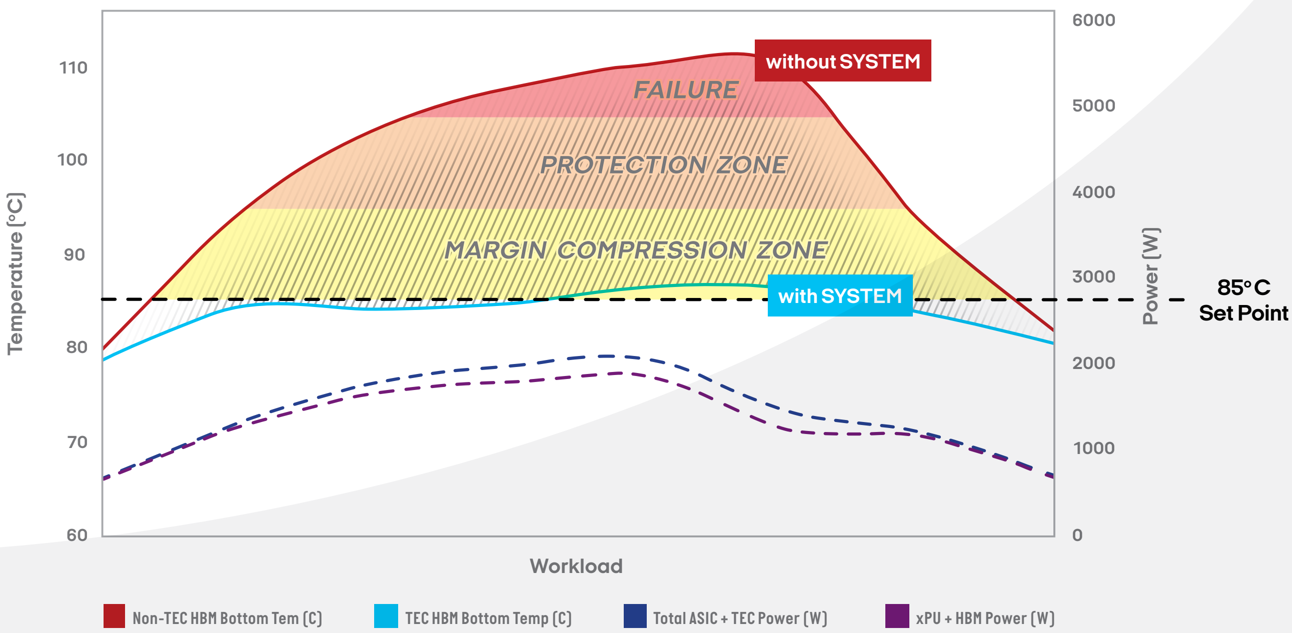


With Thermal Control Systems™, facility cooling power is reduced from 40% to 22% of total data center power, representing an approximately 18% reduction. Realizing this benefit requires a slight increase in server power consumption, from 40% to 44% (+4%). After accounting for this increase, the net impact is an overall data center power reduction of approximately 14%.

DATA CENTER IMPACT

Up to 14% reduction in Facility Energy Demand with HBM Thermal Control

Simultaneous improvement in OPEX, performance, reliability



Optimize Performance, Power, and Lifetime without System-Level Tradeoffs

PERFORMANCE WITHOUT THROTTLING:
Unlocking up to 40% incremental performance under throttling

By maintaining thermal stability through transient events, systems sustain higher clocks for longer durations. Performance gains come not from overclocking, but from eliminating forced slowdowns.

Facility Power Efficiency – Up to 0.15 PUE Improvement

Traditional metrics like PUE fail to capture the real impact of active thermal control. By reducing unnecessary cooling and enabling higher coolant temperatures, Phononic's Thermal Control Systems™ deployed throughout the data center deliver meaningful reductions in absolute facility power. For context, this is the same improvement the industry realized with the transition from air to liquid cooling solutions.

Extending Asset Longevity – Extending Useful Life by 5x

Thermal stress is a dominant factor in silicon aging. Reducing peak temperatures and temperature cycling materially extends component life, enabling:

- Full asset amortization
- Increased resale or repurposing value
- Lower replacement CAPEX over time

Predictive Maintenance as a Native Capability

Changes in cooling power required to maintain temperature provide early indicators of component degradation. This enables:

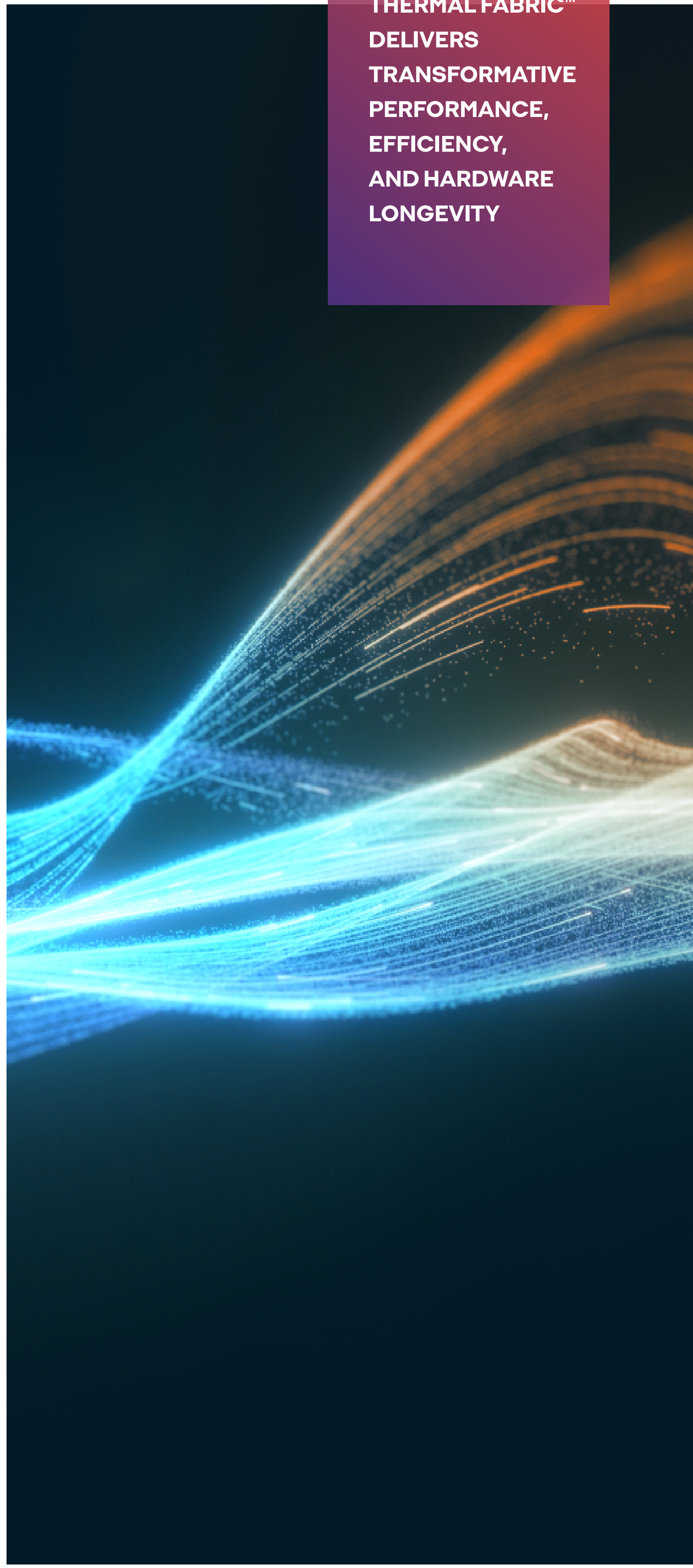
- Predictive failure detection
- Proactive maintenance
- Reduced downtime and operational risk

Why This Changes the Industry

Thermal Fabric™, enabled by Thermal Control Systems™ deployed strategically at pivotal hotspots throughout the data center, represents a shift from reactive thermal management to proactive system control. It transforms cooling into a strategic lever that:

- Accelerates adoption of next generation silicon
- Enables precise, predictive cooling
- Improves data center economics
- Improves data center orchestration and optimization opportunities
- Establishes durable differentiation through controls and software
- Allows greater flexibility in design and standard run operations, even enabling an increase of facility water from 35°C to as much as 50°C

PHONONIC'S THERMAL FABRIC™ DELIVERS TRANSFORMATIVE PERFORMANCE, EFFICIENCY, AND HARDWARE LONGEVITY



Conclusion

AI has forced a reckoning in data center design. Cooling can no longer be treated as a background utility—it must become an intelligent, responsive control layer.

Thermal Fabric™ delivers that transformation, unlocking performance, efficiency, and longevity simultaneously. Just as liquid cooling reshaped the modern data center, software defined thermoelectric cooling enables the next step forward.

To learn more about Phononic's Thermal Control Systems™ for CPO, GPU HBM and more, please visit us at: phononic.com/co-packaged-optics and phononic.com/cooling-for-gpu-hbms.

Or, request a live demo and discussion on the opportunity for active thermal optimization for your specific activation at: phononic-thermal-kit-solid-state-cooling-for-gpu-hbms/#speak-to-an-expert.

About Phononic: Phononic, Inc. is the preeminent leader in solid state cooling solutions for data centers, deployed across every major hyperscaler today with a suite of thermoelectric devices and designs, for networking, GPUs, and AI data centers. Phononic's portfolio of cooling solutions for AI data centers deliver cooling when and where it is needed so as to optimize AI infrastructure, minimizing overprovisioning, delivering cooling in millisecond timing, unlocking performance, and dramatically improving ROI, useful lifetime, and energy efficiency.

Learn more at: **www.phononic.com**