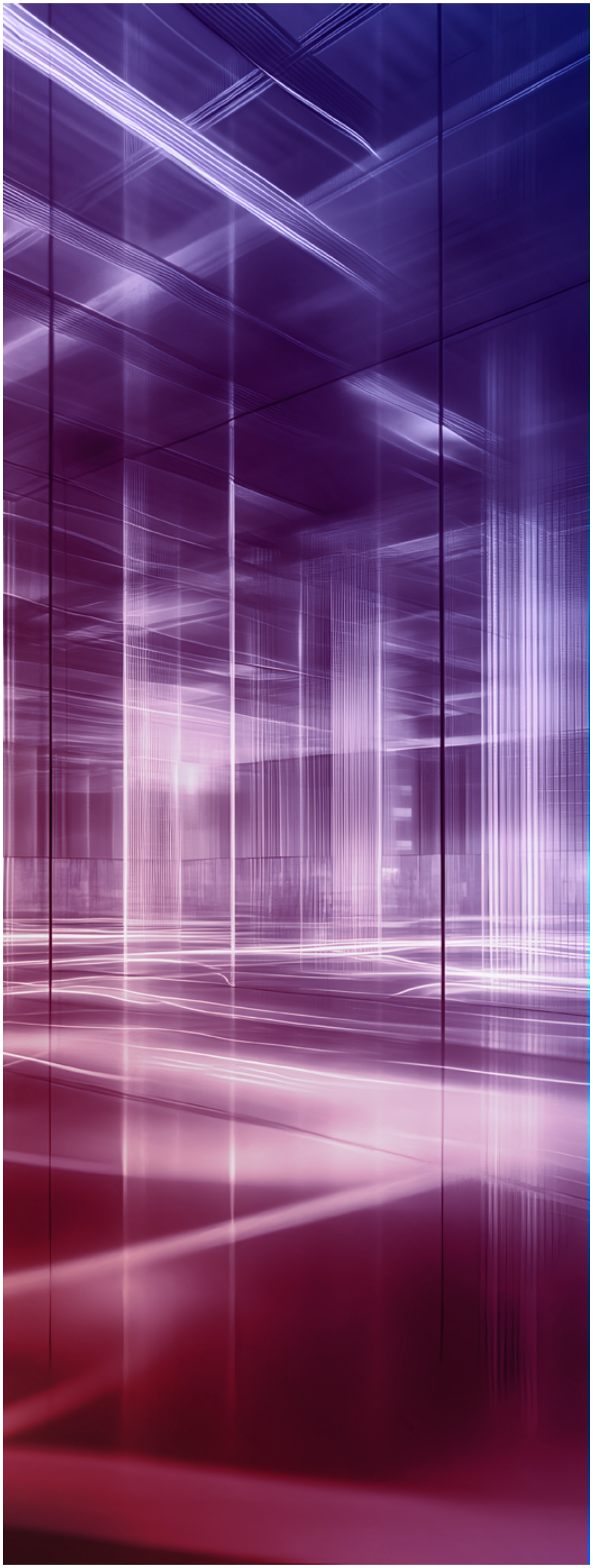
Accelerating the Future of AI –

PHONONIC'S RESPONSIVE THERMAL MANAGEMENT IMPROVES PERFORMANCE ACROSS AI DATA CENTERS



INTRODUCTION

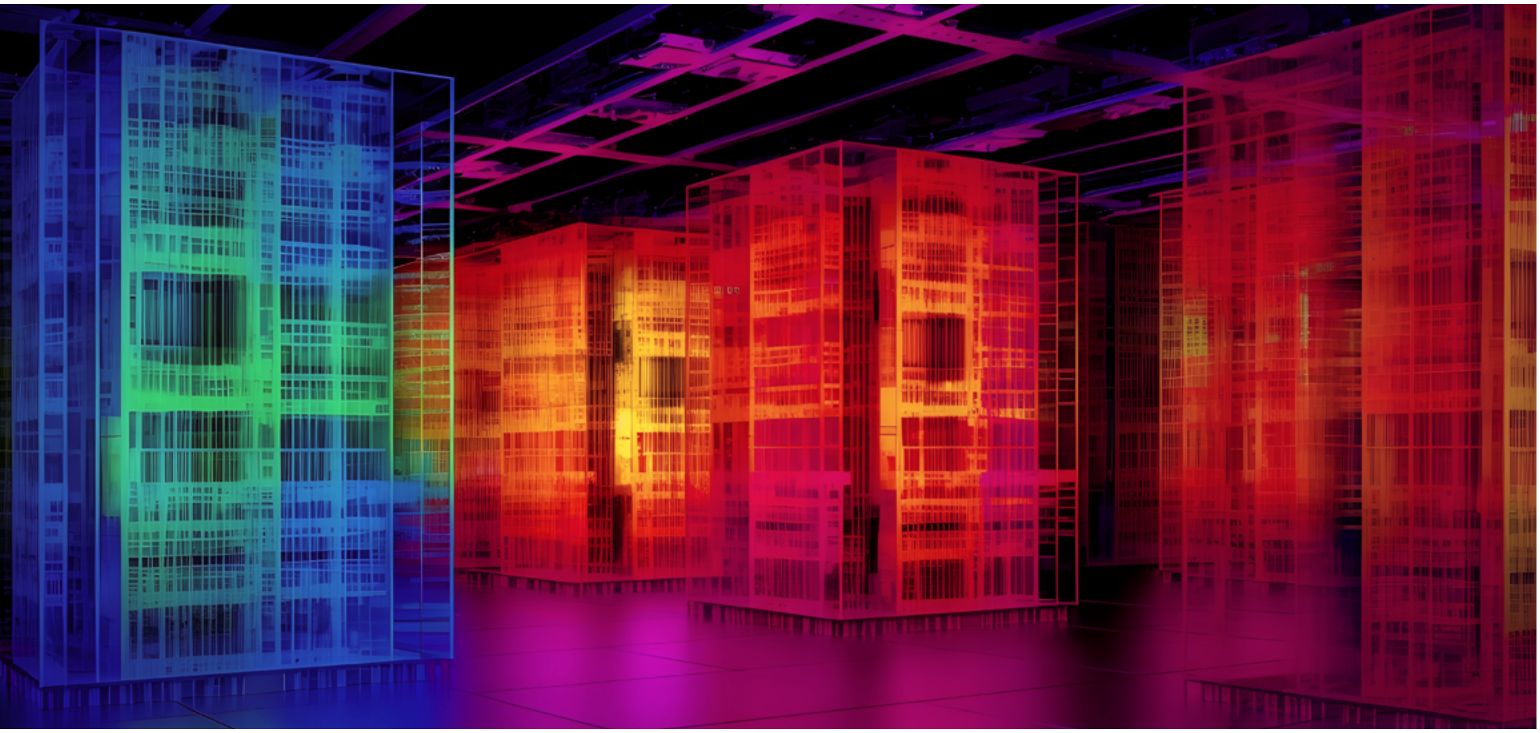
Data center investment has surged globally, driven by relentless demand for AI and a broader global reliance on digital infrastructure. Capital investment in data center construction reached $31.5 Billion USD in 2024 and is projected to reach $1 Trillion USD by 2028.

With AI models growing exponentially in complexity and scale, the race is on to build data centers that are immensely powerful while simultaneously raising the bar for energy efficiency. AI infrastructure is particularly challenging to manage, featuring higher-power density and non-uniform workloads, often in space constrained environments. To put this in context, AI racks are already pushing 120 kW, and heading toward 1 megawatt.

As data centers continue to scale, they will necessarily face revenue-limiting power constraints. Computational requirements for modern AI will continue to grow exponentially. At a very basic level, if data center operators and architects can't cool these extraordinarily dense racks, they certainly can't scale it.

TO LEARN MORE, CONTACT:

Larry Yang, CPO, Phononic
Jesse Edwards, Distinguished Engineer, Phononic



It's Not Just the Amount of Heat, It's the Density of Heat

We are at an inflection point—thermal constraints are fundamentally limiting the scaling and realization of the full potential of AI. With every new GPU generation, power density doubles. The associated cooling has not similarly scaled and is becoming the bottleneck to performance realization. The current approach to cooling the whole data hall to fix these targeted, performance constraining hotspots is like blasting AC through a house just to keep the refrigerator cold. That doesn't scale.

This translates into a serious business challenge for data centers—including hyperscalers, co-location facilities, and beyond. What AI needs is precision cooling, right where the heat lives.

At Phononic, we embrace this challenge. We are 'digitizing' cooling in a similar fashion to what has already occurred in compute. We've embedded TECs, electronics, sensors, firmware, and orchestration into the cooling layer itself—transforming cooling from a passive, sluggish mechanism to a real-time, proactive, predictive, data-driven platform. It's cooling that effectively thinks, adapts, and acts, real-time.

Just like modern compute stacks wouldn't function without dynamic scheduling or real-time telemetry, next-gen AI infrastructure will require a thermal fabric that is just as responsive.

Responsive Cooling Approaches: Take What You Know, and Make it Better

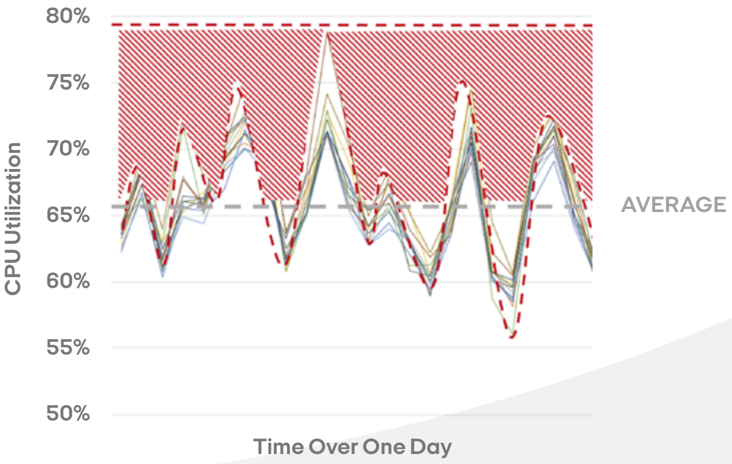
On a day-to-day or even minute-to-minute basis, processor utilization varies widely as applications start, stop, and ramp up and down. Current cooling systems have very slow response times to frequent changes in loads, because they rely on pumped liquid or forced air. Today's industry standard is to cool the infrastructure for maximum demand and minimize risk of downtime or loss. This approach results in over-provisioning cooling to meet the needs of the 'worst case' load. This approach is only logical if there isn't a more precise and responsive way to manage transient workloads.

However, this need has now been met. Phononic's Responsive Cooling™ can be deployed throughout data centers to precisely monitor workloads and provide instantaneous cooling responses, ensuring maximum performance during periods of peak demand and energy-savings during periods of lower demand.

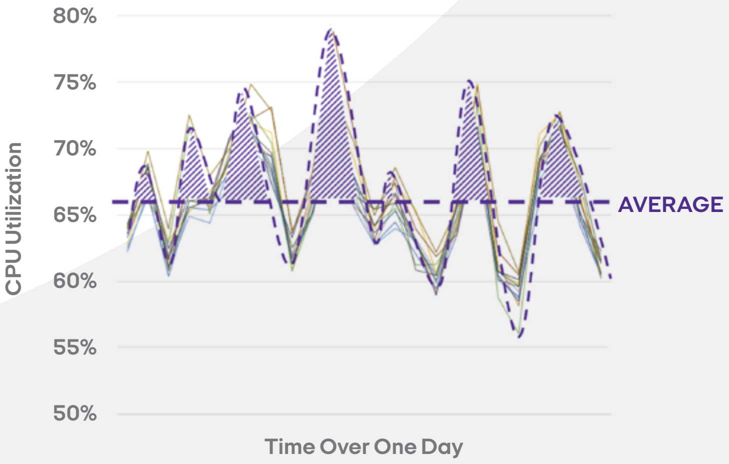
TRADITIONAL COOLING IS BULKY AND MECHANICAL, WITH MOTORS, PUMPS, AND COMPRESSORS. OUR SOLID STATE TECH IS SILENT, SCALABLE, AND RUNS ON ELECTRONS, NOT MOVING PARTS, FOR REAL-TIME RESPONSIVENESS.

PROVISION FOR AVERAGE, NOT FOR PEAK

TRADITIONAL COOLING: wasted cooling energy due to overprovisioning infrastructure for worst case



RESPONSIVE COOLING: reduce cooling energy with dynamic cooling during peak utilization



RESPONSIVE COOLING™ ENABLES A HIGHER THERMAL DESIGN POWER (TDP) FOR XPUS AND SWITCH ASICS, MAINTAINING THE SAME MAX OPERATING TEMPERATURE WHILE USING THE SAME HEAT EXCHANGER AND AIRFLOW.

READ PRODUCT BRIEF HERE:
Increase TDP while Maintaining Todays Thermal Management Architecture

Don’t Rip and Replace, Extend Air-Cooled Solutions, Increase TDP Up To 20% or more

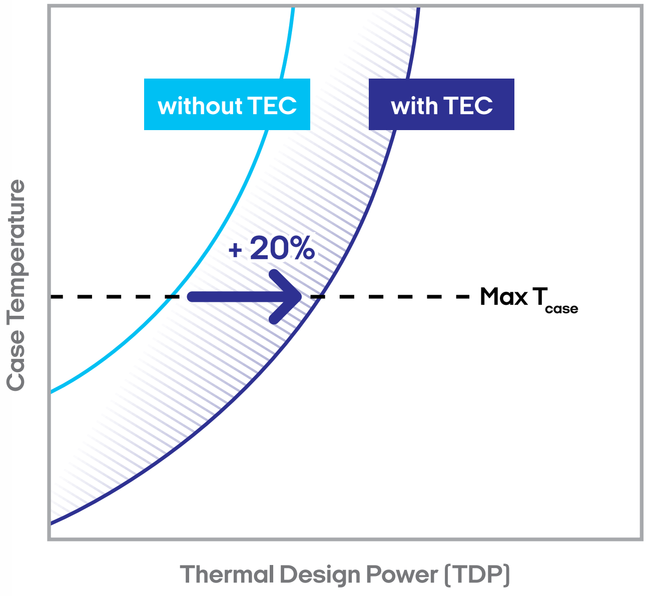
Nearly 80% of all data centers today rely on traditional hot aisle/cold aisle air cooling approaches with high-speed fans drawing air through racks. This represents a classic opportunity for Responsive Cooling™ to capitalize on existing infrastructure and unlock incremental improvements in performance. The demands of CPUs and switch ASICs are continuing to rise dramatically on the back of AI growth.

As associated TDP requirements also rise, current air-cooled approaches can’t keep pace with the thermal management requirements of next-gen CPUs and switch ASICS. Enter Phononic’s Responsive Cooling™ approach, which embeds thermoelectric chips (TECs) into the CPU or Switch ASIC heat spreader to increase TDP of chips by as much as 20% while maintaining identical air flow and cooling architecture.

These chips can be deployed to ensure precision temperature control, delivering high performance while simultaneously extending the useful life of existing infrastructure.

RESULTS:

Responsive Cooling™: Unlock ~20% increase in TDP while holding T_{case} constant.



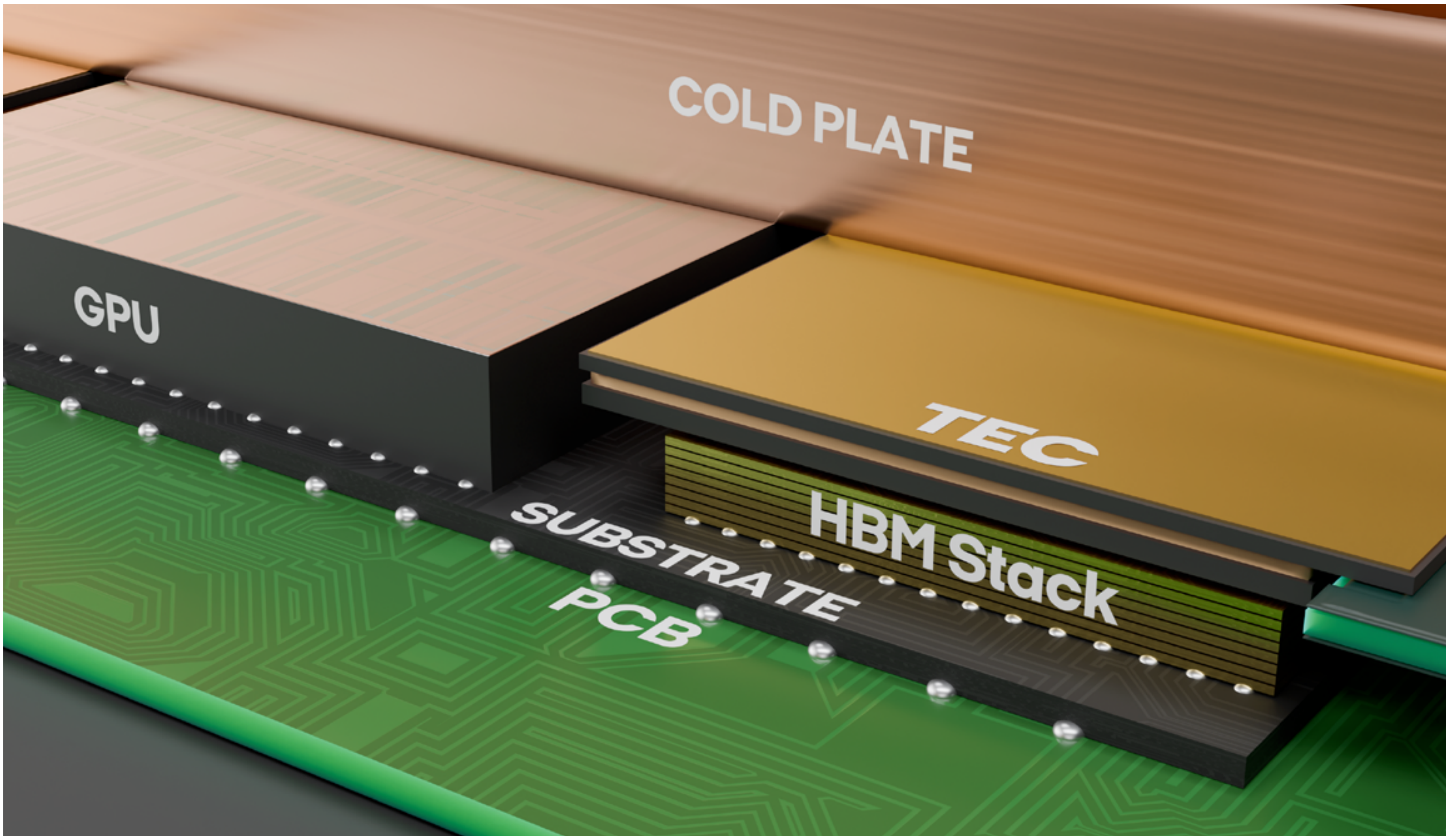
Unleashing Compute Performance, at the Source

GPU and High Bandwidth Memory (HBM) have become the heart of AI compute performance. GPUs have several compute die surrounded by eight or more stacks of HBM, each composed of a dozen or more memory die connected through silicon vias (TSVs). AI performance is constrained by both the memory bandwidth – how fast data can be moved in and out of the computing engines – and by the memory capacity – how much data can remain local to the computing engines.

The current approach of cooling GPUs and HBMs aims to reduce the liquid temperature running through the cold plates that contact both these units, both increasing energy consumption as well as forcing a sub-optimal temperature maintenance of at least one of the GPU or HBM units.

This approach is sub-optimal, because in order to keep that bottom-most HBM die cool, the cold plate must be overcooled. This gap is a direct result of our current inability to responsively decouple these two temperatures. But what if we could, and thereby unlock incremental TDP?

You could then raise the cold plate temperature and only cool the HBM stack on demand. A TEC could then lock the HBM at its ideal temperature while allowing its power dissipation to increase—unlocking increased performance!



RESPONSIVE COOLING™ ALLOWS SELECTIVE COOLING OF EACH DIE, TO THEIR PREFERRED TEMPERATURE, INCREASING BOTH PERFORMANCE AND ENERGY EFFICIENCY.

READ PRODUCT BRIEF HERE:
Responsive Direct Liquid Cooling for xPUs

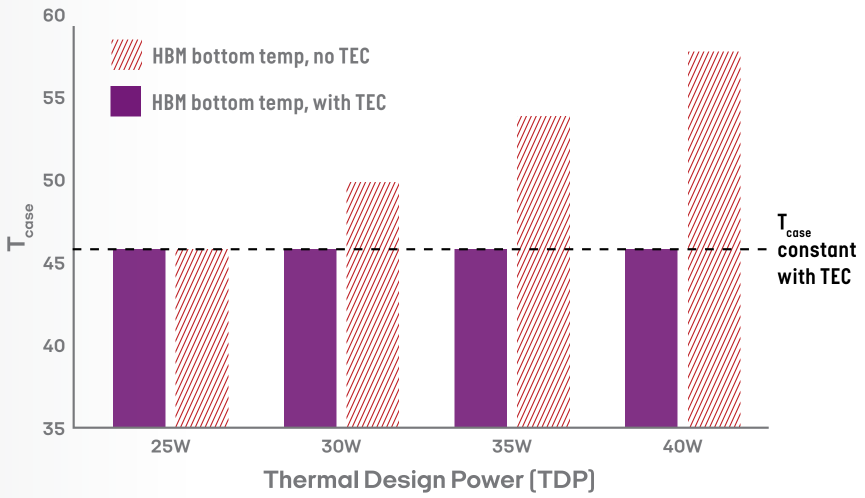
Save or Redeploy Energy for Incremental Gains

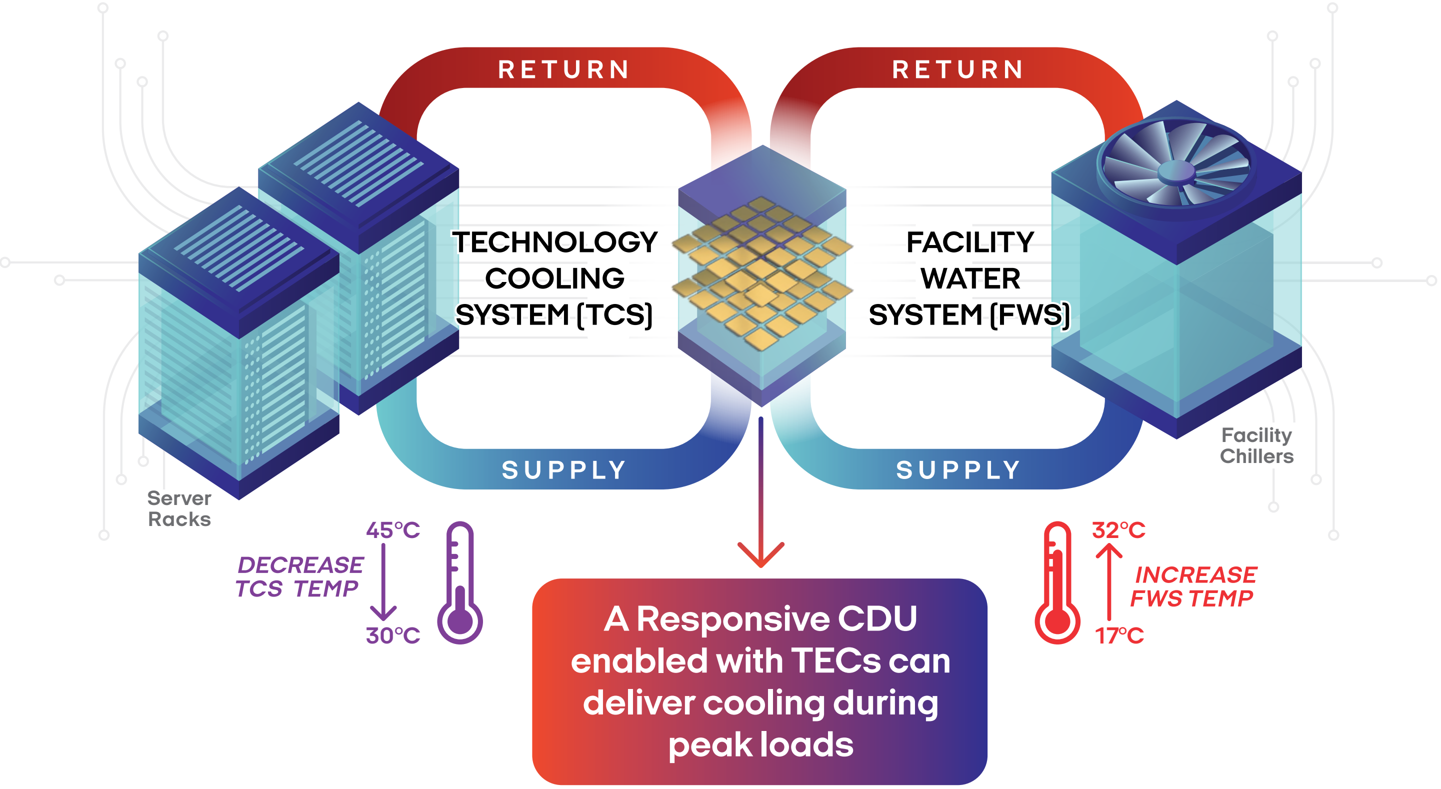
Specifically, by applying a modest ~5% increase in overall module power, only when needed, HBMs can operate at higher bandwidth and dissipate 33% more TDP vs. no TEC, without exceeding their maximum operational temperature. With a higher liquid temperature in the cold plate, the system can utilize the responsiveness of thermoelectric cooling to use cooling energy only when an HBM nears its TDP limit, reducing overall energy consumption while improving performance. Our chips are small, so they go exactly where the heat is. As racks densify, we deliver cooling that fits the footprint and the power draw.

Phononic is not just keeping up with the AI revolution, we're enabling it. We are not fundamentally retrofitting old tech, rather, we are building the thermal layer that AI infrastructure has been waiting for. Modular, solid state, and ready to scale from core to edge.

RESULTS:

60% gain in TDP with constant T_{case} versus industry standard





A Responsive CDU enabled with TECs can deliver cooling during peak loads

Solving the Technical vs. Facility Water Temperature Conflict

Phononic's Responsive Cooling™ architecture is at the intersection of precision-temperature, responsive action, and smart scalability. The Technology Cooling System (TCS) loop that cools the server racks and the Facility Water System (FWS) that evacuates heat from the building have conflicting demands as higher power servers drive a decrease in TCS temperatures to keep them cool and data center facility operators want to increase FWS temperatures to save energy and avoid the costs of operating mechanical chillers like vapor compressors. Eventually these temperatures will cross over, resulting in a "negative" approach temperature between these two loops, where traditional passive heat exchange will no longer work.

Every 1°C increase in FWS temp results in a 2-3% decrease in energy consumption. Phononic's Responsive Cooling™, built into Coolant Distribution Units (CDUs) can help solve this problem by utilizing passive cooling during periods of lower demand and activating thermoelectric cooling assistance during period of peak demand. By leveraging the responsive control of thermoelectric cooling, the TCS temperature can be held steady while the FWS temperature rises, allowing facilities to redeploy the energy that was being spent on cooling towards increasing compute performance.

RESULTS:

A Responsive CDU enabled with TEMs can deliver ~25% decrease in cooling energy at the facility level.

Solving the Cooling Problems of Tomorrow through Partnership

Phononic partners with companies across the data center infrastructure ecosystem to design and deploy Responsive Cooling™ solutions for a variety of applications—from chip packaging, heat spreaders, air-to-air heat exchangers, liquid-to-air heat exchangers, direct liquid cooling cold plates and CDUs to facility chillers. We provide design talent, IP, devices, control electronics, firmware, and software solutions that take difficult cooling challenges and provide intelligent solutions that improve performance and save cost.

It's time to take your compute performance to the next level with Phononic.

The time to respond to market demands is now. If you're only solving for power constraints, you're only solving for less than half the true constraint to high performance.

Phononic's Responsive Cooling™ approach to cooling throughout the data center integrates thermoelectric chips (TECs), existing cooling platforms, and proprietary control electronics to deliver distributed cooling that optimizes to temperature management needs. The approach decouples temperature where needed to eliminate hot spots, lowers total power requirements and unleashes meaningful TDP improvement quickly and reliably.

AT PHONONIC, WE BELIEVE COOLING SHOULDN'T BE A CONSTRAINT, IT SHOULD BE A CATALYST. WE'RE BUILDING THE THERMAL FOUNDATION FOR THE NEXT ERA OF COMPUTE: SMARTER, FASTER AND READY TO SCALE WITH THE FUTURE.

PHONONIC

About: Phononic's Responsive Cooling™ transforms cooling into a performance-critical, intelligent platform. The company is the global leader in solid state cooling, and its thermoelectric designs, IP portfolio, devices and reference design kits are the underpinning of the company's thermal management fabric of AI infrastructure. Phononic's platform technology comprised of Phononic's Intelligent software, TECs, and reference design kits can be deployed throughout all areas where thermal management is mission critical for compute performance within the data center, and serve to multiply AI performance, prevent throttling and fundamentally optimize data center performance, energy use and deployment.

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

Copyright © 2025 Phononic, Inc. All rights reserved. V01_0825 The information in this document is subject to change without notice.

Address: 800 Capitola Dr #7, Durham, NC 27713 **Website:** phononic.com **Twitter:** @phononic_inc