

Responsive Direct Liquid Cooling for xPUs

Improve performance.
Save or redeploy energy for incremental gains.

In the pursuit of performance gains, semiconductor designers are using advanced packaging techniques to create multi-chip modules of compute and memory die, each of which operate at distinct ideal temperatures. Responsive Cooling™ allows selective cooling of each die, to their preferred temperature, increasing both performance and energy efficiency.

Maintaining the distinct temperatures of compute and memory die simultaneously creates clear challenges for thermal engineers, and to date many have chosen to over-cool, while others have accepted performance compromises as a natural consequence of diverse temperature requirements. One such example is Direct Liquid Cooling [DLC] with cold plates, which are often used in these applications to maintain operating temperatures.

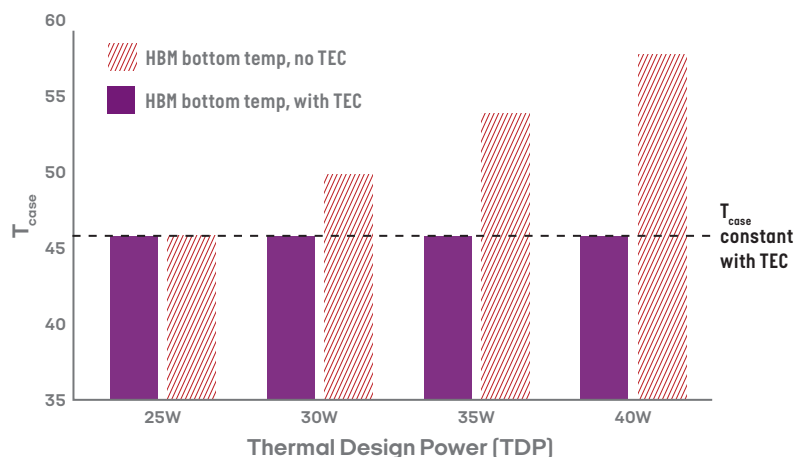
A prime example is graphics processing units [GPUs] in AI data center infrastructure. GPUs have multiple computing die surrounded by eight stacks of High Bandwidth Memory [HBM] that are each composed of a dozen or more memory die connected with through-silicon vias [TSVs]. AI performance is constrained

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. There are challenges in scaling to higher TDPs – commonly exceeding 1000W per module – and managing the hot spots across different chips.

To maintain acceptable operating conditions, the trend is to reduce the liquid temperature running through the cold plates – increasing the energy consumed at the data center level to provide colder water. An alternative approach is to utilize a Responsive Cooling™ system with warmer water that integrates thermoelectric chips [TECs], cold plates, and control electronics. This system thermally decouples the computing die and HBM stacks, using TECs to manage the temperature of the HBMs. By applying a modest ~5% increase in overall module power, 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.

RESULTS:

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



PHONONIC RESPONSIVE COOLING SYSTEM™

Take Your Compute Performance to the Next Level with Phononic.

Thermal management has always been a constraint to achieving higher memory and compute performance, but it's increasingly becoming a bottleneck to the step-change progress required to support the future of AI. AI performance is delivered by faster GPUs and GPU performance is constrained by the power and cooling capabilities of the system. At the heart of the issue? The ability to respond immediately and precisely to changes in workload demand, at a system-level, not merely component to component. This stands in stark contrast to current approaches that over-provision cooling, resulting in energy 'wasted' on cooling that could have been deployed to compute.

Phononic's Responsive Cooling System™ 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. This approach decouples temperature where needed to eliminate hot spots, lowers total power requirements and unleashes meaningful TDP improvement quickly and reliably.

Scale your compute performance, with Phononic.

Multiply Your AI Performance with Phononic. Increasing future AI performance requires a fundamental change in cooling approach. Phononic's Responsive Cooling™ transforms cooling into a performance enhancing, intelligent platform.

Design Capabilities that Push the Cooling Boundaries to Enable AI

Phononic has a deep understanding of data centers and the components that are powering the future. With more than 30M+ devices in field today, deployed across all major US hyperscalers, Phononic's engineering team has been setting the standard for performance, efficiency and cost-effectiveness in our TECs.

Our IP library is robust and growing, with hundreds of patents covering materials, software integrations, thermal management approaches and more.

A wealth of reference design kits, along with the backing of ISO Quality Management Systems, IATF and Telcordia certifications ensures that our designs are consistently real-world predictive, reliable and deployable.

Learn more:
<https://phononic.com/datacenter-cooling>

Industry-Leading Design, Delivered Consistently and at Scale

At Phononic, as part of our design approach, we work closely with our customers to make system level tradeoffs that optimize not just the custom TEC we deliver, but the entire product for the end customer.

By leveraging our proprietary TEC technology and scalable device architecture, we are uniquely positioned to reliably deliver high performance cooling that meets the most rigorous demands, regardless the sector.

Together with our partners in Thailand, Phononic has the ability to scale availability of devices, and fully integrated solutions to our partners and licensees in a manner that both leverages the deep R&D and engineering expertise in HQ in RTP, NC while enabling full global scale and supply chain flexibility.