REPORT

In-Depth Report of Thermal Management Solutions for I/O Modules

Emerging Advancements in the Data Center Ecosystem

MAY 2024

TABLE OF CONTENTS

Introduction 01

The State of Cooling: Legacy Thermal Solutions 03

Thermal Challenges for Optical I/O Modules 06

Innovative Thermal Management Solutions for Data Center Architecture 09

Standardization and Testing for Next-Generation Cooling Strategies 12

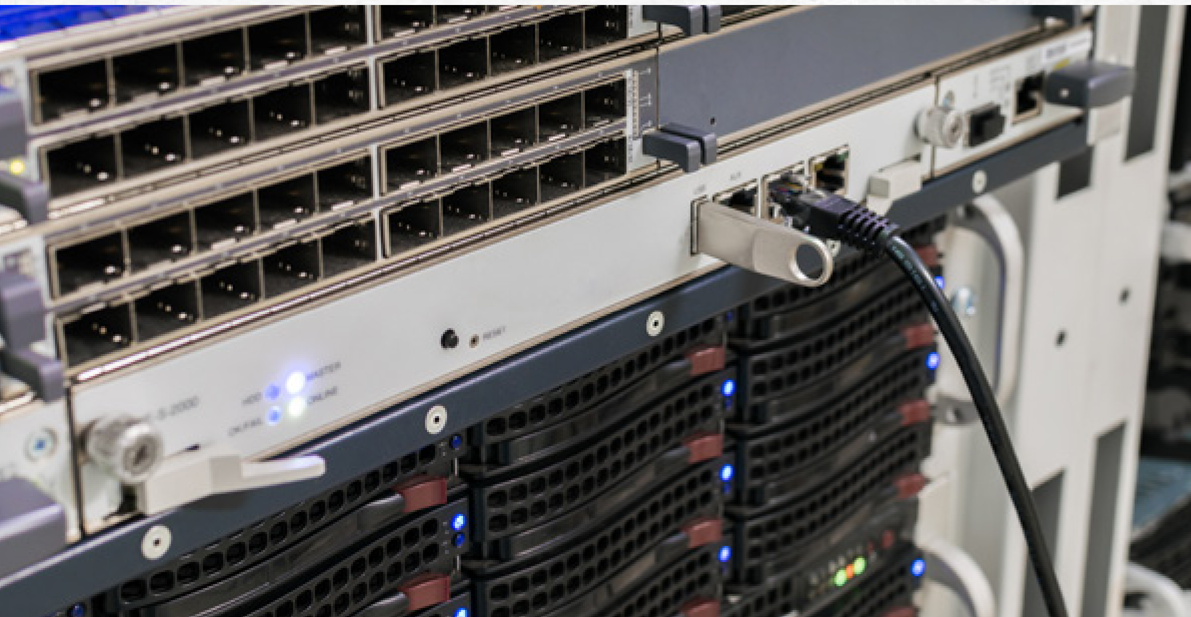
Enabling Innovation for the Future of Data Center Cooling 14

INTRODUCTION

Cloud computing in data centers has become the dominant enabler of digital products and services, ranging from basic email to sophisticated generative artificial intelligence (AI). This computing power isn't free, with each server in a data center requiring electricity to operate. Power consumption can reach high values, especially in data centers supporting high data processing demands for advanced areas such as AI, machine learning and more. The major consumers of this power are the GPUs and accelerator cards that power these advanced services. Data centers continue to rapidly increase compute density as much as possible, which inherently creates more thermal challenges. Focusing on effective thermal management strategies has never been more important.

With more companies embracing digital transformation, data centers are under additional pressure to provide high-efficiency compute power while minimizing maintenance and operating costs. Thermal management represents one of the major costs of operating data centers, and effective thermal management reduces long-term maintenance costs by extending the lifetime of components. According to IT solutions provider Enconnex, **operating expenses** for modern liquid cooling systems can reach $2,000 per kW of power being cooled, and investment for enterprise data center cooling systems can easily exceed $100,000. Clearly, these expenses can challenge today's broader corporate initiatives focused on cost efficiency, and thermal management may be identified as a natural place to start tackling capital expenditures (CAPEX) and operating expenses (OPEX).

The untold story of thermal management in data centers lies in the optical modules used for communication between rack-mount servers, networking switches and between data centers. Servers don't operate in isolation — they need to communicate with each other in clusters via fiber optic links to enable next-generation services such as generative AI. Scaling these services requires scaling these server clusters — and the data rates at which they communicate. As new technologies become available and enable higher data rates, the power demands in optical I/O modules and Active Electrical Cable (AEC) transceivers are also increasing. For example, current power levels at 112 Gbps-PAM4 data rates are approximately 15 to 25W, and just the optical I/O modules in a large enterprise switch with 32 ports would consume up to 0.8kW. If coherent (800G) optics are used for 112G communication over long distances, the power levels can reach as high as 30W per module. At these power levels, the I/O modules are pushing traditional forced-air cooling systems to their operational limits.



The shift to **224 Gbps-PAM4** interconnects represents a doubling of the per-lane data rate. Power consumption also increases, with optical modules alone reaching as high as 40W over long-range coherent links. This is challenging because optical I/O module power requirements have increased from 12W to 40W over just a few years, yet module form factor has not changed. This essentially represents nearly a 4X increase in power density, demanding new approaches to cooling. Implementing liquid cooling solutions carries additional investment and maintenance costs, but an integration of creative liquid cooling solutions within existing form factors can address these greater power and thermal demands in I/O modules.

Due to the increasing power demands in optical I/O modules, systems designers and data center architects are now considering the use of liquid cooling for optical I/O modules to support upcoming 224G implementations. Beyond liquid cooling lies more advanced approaches to module design and characterization, which can enable the next generation of high-speed network interconnects.

This report will examine the limitations of legacy approaches for thermal characterization and management, and explore new innovations in server cooling and optical module cooling being implemented in systems requiring 112G and 224G links.

THE STATE OF COOLING: LEGACY THERMAL SOLUTIONS

High-power systems generally use active cooling, or cooling that requires an active power system to remove heat from network infrastructure. The use of active cooling brings the investment and maintenance costs discussed earlier, as well as the need for experienced technicians who can install and maintain active cooling systems. The common set of active cooling measures in data center architecture includes:

Forced airflow (or directed airflow): These systems pump air directly from a plenum into server racks, including in elevated floor configurations. Servers and switches can feature their own dedicated fans, which also assist in pulling air through the enclosure. The ability of these systems to fully cool specific components in a server — including processors and optical modules — is limited.

Liquid cooling: In this method, a liquid with high thermal mass is circulated onto a cold plate, which then interfaces with the heat-generating components in a rack-mounted system. Water is one option for these systems but other dielectric fluids, such as oils or propylene glycol (PG-25) mixtures, are also commonly used.

CURRENT ACTIVE COOLING APPROACHES

All modern data center deployments rely on active cooling due to the high thermal demands in processors and ASICs. It is the most effective method in terms of heat dissipation capacity, which is enhanced when fluid flow is directed to target components and assisted with additional passive components. Both forced airflow and liquid cooling are found in modern data center deployments.

These systems stand out for their exceptional heat dissipation capabilities, particularly when they incorporate mechanisms that allow fluid to be directed onto heated components, thus facilitating the exchange of heat with a cooler medium. Direct-to-chip liquid cooling takes active cooling a step further, especially in data centers where high-performance compute processors are generating most of the heat in a server.

Forced air: Air cooling is a low-risk active cooling approach and includes methods of directing airflow to heat sinks that are in direct contact with hot components as needed. When power demands are on the order of 10kW per rack, forced air systems can often handle the thermal load. Although liquid cooling systems may be present on the high-power components, forced air will remain part of a cooling strategy even when power demands in chips and I/O modules scale to high levels.

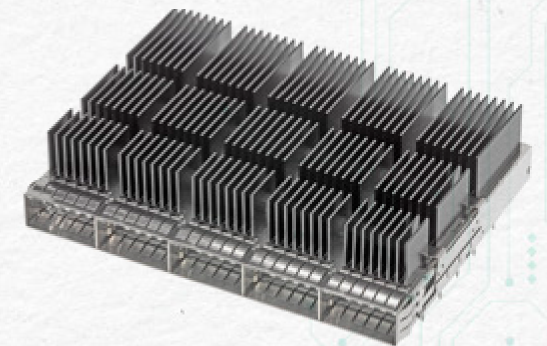
Direct-to-chip liquid cooling: One liquid cooling option for use in data centers is direct-to-chip liquid cooling, which is often used in high-compute processors required for today's cloud environments. In direct-to-chip cooling, fluid flows through a cold plate which interfaces with the chip's exposed rear surface and thus pulls heat from the hot component. According to Jeff Schuster from Enabled Energy, Inc., **direct-to-chip cooling** is needed to provide heat dissipation once power demands reach 25kW to 50kW per rack.

"Direct-to-chip liquid cooling takes active cooling a step further, especially in data centers where high-performance compute processors are generating most of the heat in a server."

While these active cooling options are the most effective, they also present complexity and require the most maintenance. Many liquid cooling solutions exist for processors, accelerators and power systems in data centers, but the need for solutions to cool optical I/O modules is now also emerging. For these portions of a server or switch, most operators currently rely on forced air or passive cooling for I/O modules.

PASSIVE COMPONENTS ENHANCING ACTIVE COOLING

Some passive components assist in active cooling strategies, aiding heat transport and providing some additional thermal mass. Principally, the common passive components used with active liquid cooling or forced airflow are heat sinks and heat pipes. Chips and GPUs are often deployed with a heat sink and an active cooling option, such as a fan or liquid cooling. Riding heat sinks on optical I/O modules can also assist forced airflow in transporting heat away from hot modules.



QSFP-DD cage with heat sink

To assist with heat transfer to integrated heat sinks and riding heat sinks on optical transceiver modules, one solution is the integration of thermoelectric cooling that targets internal components with the highest temperature sensitivity (e.g., the laser). The Peltier effect draws heat into the heat sink where it can be dissipated via airflow. While useful as an internal heat transport mechanism, it does not alleviate the higher heat dissipation demands of 224G optical I/O modules.

IMMERSION COOLING

Arguably the most effective liquid cooling option in data centers is immersion cooling, where an entire server is cooled by being submerged in a non-conductive liquid. The liquid provides significant thermal mass and can be circulated to a heat exchanger. Immersion cooling provides very effective thermal cooling that exceeds roughly 50kW per rack.

While highly effective, immersion cooling carries significant risk and cost, as outlined below.

- **Investment:** Equipment and installation costs for immersion cooling systems can be more expensive than forced air cooling or liquid cooling. This is largely because it requires a complete overhaul of data center architecture, whereas air and liquid cooling may be deployed with a retrofit approach.
- **Space requirement:** Racks that are compatible with immersion cooling tanks are typically wider and deeper than standard rack units.
- **Compatible I/O modules and connectors:** The dielectric constant of the fluid impacts the electrical impedance of the connectors. Since connectors are typically designed with the assumption that air is going to be the fluid during operation, special connectors and transceiver modules are required.
- **Compatible servers:** Servers that will work with immersion cooling are purpose-built and are not available from all server vendors.
- **Fluids:** While effective in terms of their thermal mass, immersion cooling fluids require special circulation systems to cool the fluid.
- **Maintenance:** Due to specialized equipment, these immersion cooling systems tend to incur high maintenance costs.
- **Risk of leaks:** If there is a catastrophic leak in an immersion cooling system, flooding could damage other areas of a facility.
- **Component failure:** Insufficient flow near some components results in high temperatures, which can accelerate aging and lead to early failure.
- **Environmental impact:** The fluids used in immersion cooling need to be replaced periodically and require correct disposal procedures.

Immersion cooling often requires hardware to be designed or adapted for submersion. Components need to be assessed for their ability to withstand being in a fluid environment over long periods. When evaluating thermal demands in optical I/O modules in 112G and 224G systems, extending liquid cooling directly to the modules can address the thermal demands without the expense of specialized immersion cooling systems.

THERMAL CHALLENGES FOR OPTICAL I/O MODULES

Optical I/O modules inside of servers and rack-mount network infrastructure systems are always receiving direct cooling from an active cooling system, specifically from forced airflow coming from the front panel of rack-mount equipment. Thermal design in rack-mount equipment requires balancing thermal management of I/O modules with heat dissipation in processors or ASICs to avoid excess margin for either the I/O or ASIC operating temperatures. Optimizing the cooling strategy to account for processor cooling demands and overall optical I/O module power can help strike the right balance, maximizing the power efficiency of the system.

Link length vs. data rate: Optical I/O modules used for 56G and 112G can currently get by with air cooling. When implementing coherent optics at 112G data rates or beyond, pluggable optical I/O module power levels (33W+) may require extending liquid cooling measures to the modules.

The 112G and 224G generations of transceivers are still targeting standard link lengths defined in **IEEE 802.3 standards**, so systems designers and data

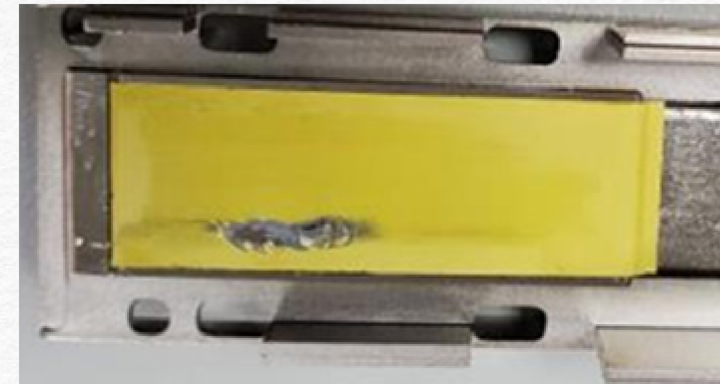
> The thermal demands already present in prior generations of optical modules are expected to increase, and the old approaches to thermal management may underperform.

center operators should have little expectation that standards will change simply to accommodate higher power requirements in optical modules. This means the thermal demands already present in prior generations of optical modules are expected to increase, and the old approaches to thermal management may underperform.

Form factor: Pluggable optical modules bring challenges in that the form factor has not changed since the implementation of fiber optic transceiver modules nearly 20 years ago. Now that the industry is moving to 224G, the new generation of optical I/O modules requires backward compatibility with existing rack-mount equipment to enable upgrades. This means heat density will continue to increase, and this may lead to the exhaustion of forced air as the only method for cooling optical I/O modules.

Heat sinking: Heat sinks attached to optical I/O modules bolster cooling abilities of forced airflow systems, but they are constrained by metal-to-metal contact — due to durability requirements — to maximize heat transfer. Bare metal contact is undesirable for any heat sink contact, but this is particularly true on I/O modules given the significant increase in optical module power levels over the past several years. The projected increase in power demands reaching as high as 40W per module further aggravates this bottleneck. To improve thermal contact resistance at bare metal contact surfaces, a thermal interface material (TIM) can be mounted to a riding heat sink that makes intimate contact with the pluggable module and helps to increase heat transfer efficiency.

The problem with attaching TIMs to a riding heat sink is the reliability of the TIM. When being plugged or removed from a cage, the sharp edges of the module will scrape away the TIM and cause the thermal efficiency to decrease with each mating cycle, making it ineffective after the first couple of insertions — if not the first insertion. This durability challenge is further intensified when these modules are exposed to aggressive field conditions such as angled insertion due to cable loading, as this makes the fragile TIM surface even more exposed to sharp edges on the module. To ensure high reliability with repeated mating, the heat sink contact method needs to be re-engineered so that TIMs can withstand up to 100 mating cycles.



Damaged thermal pad on cage/heat sink

Monitoring module temperature: Increasing power density requires a re-evaluation of the traditional thermal characterization approach for the optical modules. Traditionally, a blanket 70°C case temperature requirement was used as the thermal specification (i.e., as a proxy of the digital optical monitoring (DOM) temperature). However, recent studies are showing that even at case temperatures of 70°C, more than a couple degrees of margin are left with the temperature-sensitive components inside the module. This leads to inaccurate conclusions about thermal feasibility of systems and overdriving the cooling system. For example, in a system where I/O thermal performance is the limiting factor, the fans would run at a higher-than-needed speed simply to meet case temperature requirements — even when excess margin is available based on internal component temperature of the modules. A new thermal characterization (discussed later in this report) will help resolve this limitation of today's approach.

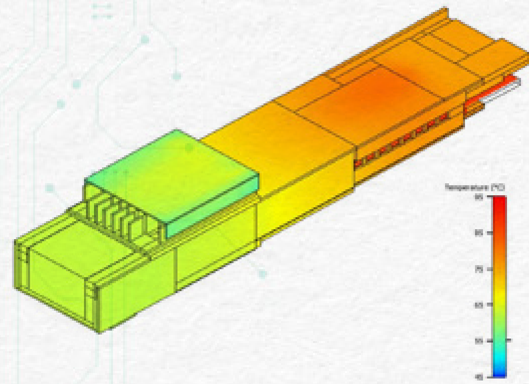
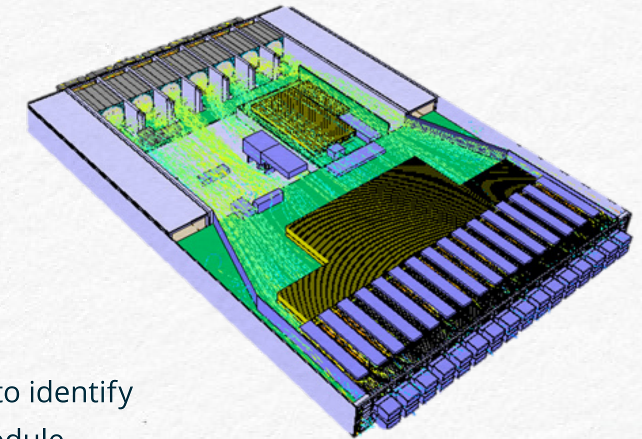


Illustration of module temperature

Simulation and testing: Simulation/predictive engineering is used to optimize a system design, component placement and cooling strategy before build and deployment. Optimizing a heat sink plus a forced air approach on an optical module often requires simulating airflow in the entire chassis before a mechanical design is finalized. Rack-mount servers are standardized in terms of their height and width, with most deployments using the 1RU form factor.

Placement of other components (e.g., chips, add-in cards, SSDs, etc.) can affect the airflow path through the enclosure and along a bank of I/O modules, thus affecting cooling effectiveness.



Thermal simulation system

Component-level simulation is also important for optical I/O modules to identify hot spots along the body of the module. Simulations need to consider the internal structure of the module itself, followed by correlation to measurements of isolated modules. When running in isolation, temperature tests range from contact measurements to infrared camera measurements. Once the thermal profile in a transceiver is understood, it can be used as an input for system-level simulation, followed by system-level testing and correlation.

Immersion cooling: High-power 112G and 224G optical modules can be effectively cooled in an immersion cooling system. While this is the most effective method for cooling from a thermal load perspective, the dielectric fluid creates challenges with module connectors, primarily in terms of signal integrity. Optical modules and I/O connectors are most commonly designed assuming the surrounding dielectric is air, so replacing it with an alternative dielectric creates coupling inefficiency. The result is that 112G and 224G channels in immersion-cooled rack-mount equipment will require specialized modules that are compatible with the dielectric fluid. Lower supply and more specialized construction lead to greater cost per rack unit when immersion cooling is preferred.

INNOVATIVE THERMAL MANAGEMENT SOLUTIONS FOR DATA CENTER ARCHITECTURE

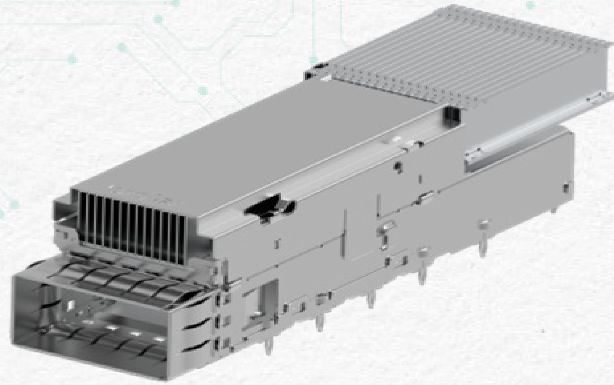
Given the increasing thermal loads, as well as the form factor limitations due to backward compatibility in servers and optical I/O modules, liquid cooling solutions that are already present in servers and switches may need to be extended to the modules to support higher data rates and greater compute requirements in data centers. For I/O specifically, new solutions can be integrated into servers and switches that provide greater heat sinking without compromising reliability. This is made possible through mechanical changes and innovative liquid cooling directly on modules, which maintains the standard form factors used in rack-mount networking systems and pluggable modules.

DROP DOWN HEAT SINKS

To maximize the heat transfer capability of a riding heat sink, the dry metal-to-metal contact between the heat sink pedestal and the pluggable module must be improved through implementation of a TIM. As highlighted earlier, when plugging in an optical I/O module, its sharp corners can damage the TIM and reduce the number of allowed mating cycles. This requires an alternative contacting mechanism for the heat sink to preserve the mechanical and thermal integrity of the TIM over numerous insertion cycles.

Molex has developed an innovative solution using drop down heat sinks (DDHS) on optical I/O modules to enhance thermal management. The breakthrough design of the DDHS ensures that there is no direct contact between the module and the TIM, effectively creating a levitating heat sink that only makes contact when the module is nearly 90% inserted into the receptacle. During the final 10% of insertion, the heat sink “drops down” onto the TIM and makes complete contact with the surface of the module that is free from any sharp edges. This allows successful implementation of the TIM for more than 100 insertion cycles. Drop down heat sinks can be implemented in different single row and stacked cage configurations.

The Molex DDHS is a drop-in replacement for today’s traditional riding heat sink solution. When compared to an already optimized zipper fin heat sink solution, DDHS can offer up to a 9ºC improvement at 35W.



Drop down heat sink system from Molex

This drop down heat sink solution provides a reliable heat management option that fits within standard module and rack-mount form factors. System designers can choose to take advantage of this 9ºC improvement in one of two ways:

- Use modules with the same power — 30W, for example — and simply lower the system fan speed to use up the thermal margin from the DDHS, allowing for greater power efficiency.
- Cool 5 to 7W higher power modules (35-37W instead of 30W) while running the fans at the same speed.

The DDHS solution enables systems to cool higher power modules with a simple drop-in replacement.

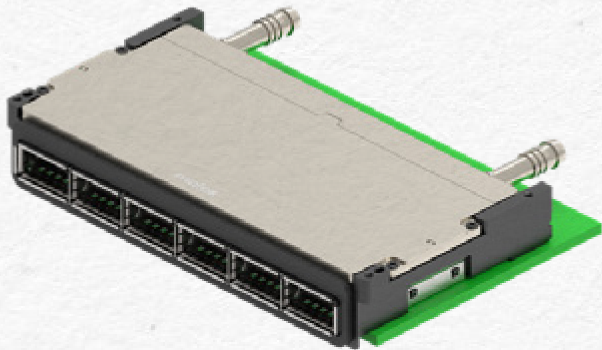
ADVANCED LIQUID COOLING SOLUTIONS

At 112G data rates, optical I/O modules are operating at power levels that push the ability of forced air cooling nearly to its limit. In 224G implementations, liquid cooling may be required to manage heat generated in optical I/O modules. Because high-compute processors are already using liquid cooling, it makes sense to integrate a solution for high-power optical I/O modules into the existing cooling system. This then enables retrofitting of existing equipment as new technologies are implemented at higher data rates.

While liquid cooling is not new to the data center industry, it does present some inherent challenges when it comes to its implementation for pluggable I/O. The natural path to implementing liquid cooling is to substitute the individual riding heat sinks with individual cold plates. However, that results in as many as 32 inlets and outlets. This level of plumbing is unmanageable in the constrained 1RU/2RU system space. The next step is to implement a single cold plate that can cool multiple I/O ports. The challenge with this approach is that each of the I/O ports has a different tolerance stack up depending on module height, module positioning inside the cage, pedestal height, etc. While it may be possible to ensure good thermal contact with one port, the differing stack up for each port makes it impossible to guarantee adequate thermal contact for each one of the ports. For example, in a 1x6 cage configuration, this would essentially require perfect coplanarity for all the cold plate pedestals as well as all the module surfaces that contact the cold plate. This indicates the need for a compliant pedestal that can reliably address each port's tolerance while providing sufficient force to make adequate thermal contact.

To solve these challenges, Molex developed a liquid cooling solution called the integrated floating pedestal. For this solution, each pedestal that contacts the module is spring-taught and moves independently, allowing implementation of a single cold plate to different 1xN and 2xN single row and stacked cage configurations. The independently moving pedestals can compensate for different tolerance stack ups for each port while still providing the desired downforce for good thermal contact.

An example of this is the 1x6 **QSFP-DD** liquid cooling solution shown below. This solution features six independently moving pedestals which can compensate for the varying stack up for each port — while ensuring good thermal contact (with desired downforce).



Example integrated floating pedestal from Molex

With this integrated floating pedestal, I/O liquid cooling can be achieved without thermal or mechanical gap fillers. Gap fillers add thermal resistance to the conduction path. In this solution, heat directly flows from the module generating heat to the pedestal, and the pedestal directly interfaces the liquid flowing through the cold plate. This is theoretically the shortest possible conduction path a liquid cooling solution can achieve, helping to minimize thermal resistance and maximize heat transfer efficiency.

While strongly dependent on the boundary conditions, Molex has demonstrated that with this liquid cooling solution, modules as high as 40W can be cooled to within specifications.

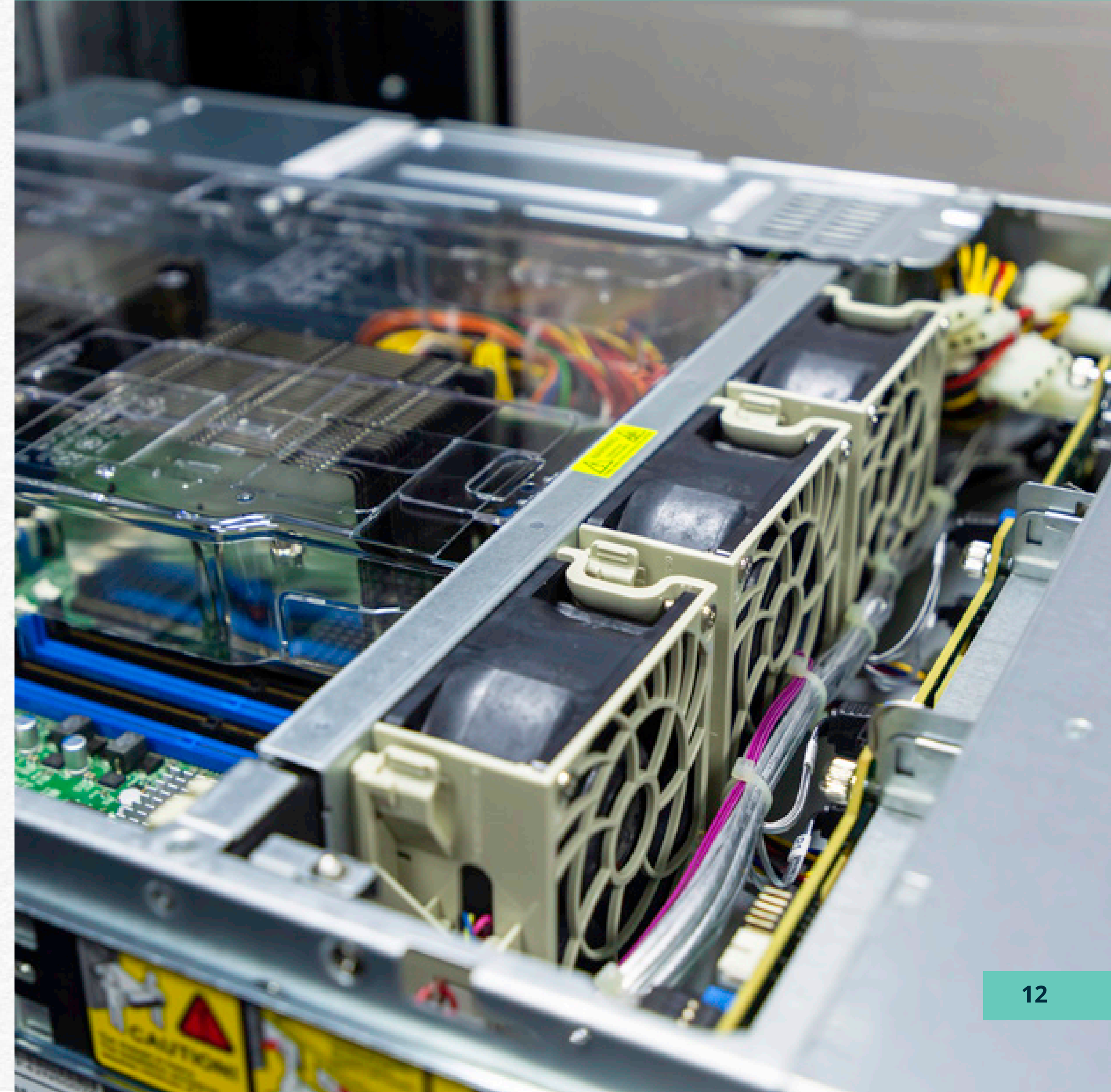
Molex liquid cooling solution demonstration

STANDARDIZATION AND TESTING FOR NEXT-GENERATION COOLING STRATEGIES

One important factor affecting the design of cooling strategies for optical modules is the use of case temperature as a specification or limit on the module temperature. These modules are complex designs, and a case temperature specification alone will not provide an accurate reflection of the internal temperature of the critical components in the module. It is the temperature limits of internal components which will define whether a module will operate to specification. Using case temperature as the specification potentially leaves significant operating margin on the table.

The traditional approach to monitoring module temperature is to select a monitoring point on the module case, which is likely beneath the heat sink. The system cooling strategy is designed such that the maximum case temperature specification (Tcase, typically 75°C) is not exceeded during operation. This monitor point is typically inaccessible during operation without disturbing the heat sink and is not a direct reflection of the actual temperature of internal components. However, internal sensors report the Tcase value using the Digital Optical Monitoring (DOM) value, which can be read by the software management interface (i.e., CMIS).

"Using case temperature as the specification potentially leaves significant operating margin on the table."



An example of the margin lost using the case temperature approach is shown in the table below. The table shows the readings from a typical module located in the lower port of a stacked cage. The temperature limit for the module case is compared with the actual temperatures of critical internal components required to ensure module operation and performance. When the internal component temperatures are examined, we can see that there is still excess design margin available.

Module	Limits	Actual	Margin (ΔT)
Tcase (above DSP)	75°C	72.6°C	2.4°C
Laser	85°C	76.4°C	8.6°C
TIA/driver	105°C	81.4°C	23.6°C
DSP	105°C	93.5°C	11.5°C

In this case, the cooling strategy could be redesigned such that the load on fans is reduced and thus the case temperature could run hotter; this would allow the system to exploit some additional margin. Using the module case temperature as the module's temperature limit yields only a 2.4°C margin. If, instead, the laser is used as the critical component (with least thermal margin) defining temperature limit, one finds that there is actually 8.6°C of available margin before any performance impact on the laser would be noticed.

Therefore, it is proposed that module DOM reading for optical modules be redefined based on the lowest temperature margins of the internal components, as illustrated in the formula below. As mentioned earlier, additional margin can be exploited in the cooling system design while maintaining backward compatibility with existing CMIS and system software. The value reported in the DOM register becomes:

$$DOM = 75°C - Min(\Delta T(laser), \Delta T(DSP), \Delta T(TIA), etc.)$$

This proposed definition for DOM has a simple interpretation: the DOM value, and thus the actual temperature margin, should be based on the internal component (e.g., laser, optics, TIA, DSP chip, etc.) that has the smallest margin in the module's operating environment. This simple change in reported DOM values helps system designers eliminate excess margin in the cooling system architecture and provides much better module control for system management.

ENABLING INNOVATION FOR THE FUTURE OF DATA CENTER COOLING

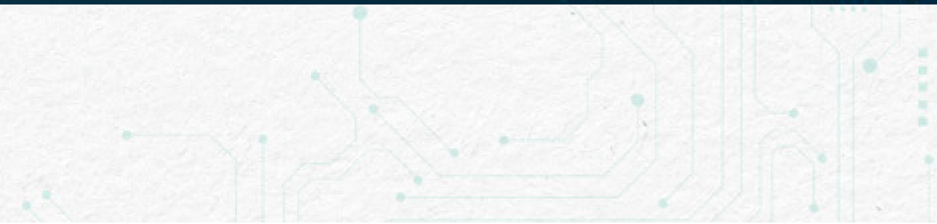
Leveraging decades of experience and extensive expertise in thermal management for complex data center environments, Molex is bringing innovative approaches to the increased heat challenges that come with greater data rates. While the pace of change is rapid, the constraints on system design and implementation remain. Standardization of equipment form factors demands inventive solutions that meet space constraints while preserving a high I/O count.

Molex has developed industry-leading cooling solutions for optical I/O modules, providing greater reliability for systems running at high speeds with demanding power requirements. Uniquely designed heat sinking and contact methods for pluggable optical I/O modules provide much more reliable performance with lower complexity than legacy thermal management solutions. This paves the way for scaling up the next generation of data center interconnect architecture without resorting to cumbersome immersion cooling methods.

Choosing the right provider to navigate the complexities of data center thermal management is pivotal in moving forward confidently. Molex brings advanced capabilities for next-generation data center architecture and a collaborative customer-first approach for optimizing both performance and efficiency.

Share

creating connections for life



molex

987652-7012 ©2024 Molex