molex

报告

深度分析报告：I/O光模块的热管理解决方案

数据中心生态系统的新发展

5月24日

目录

介绍

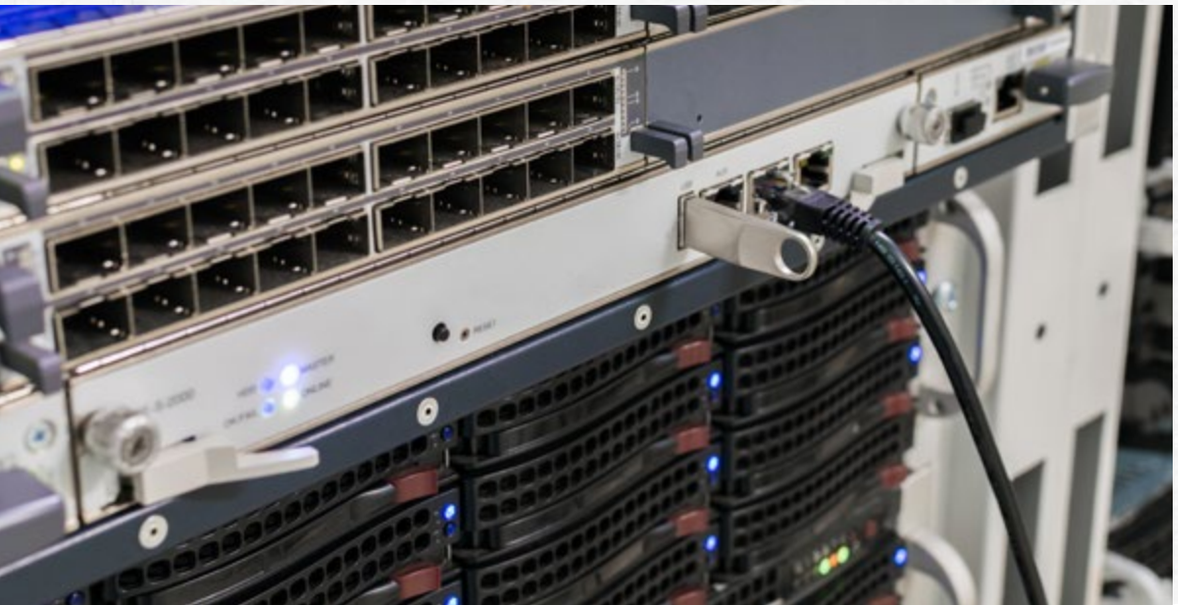
数据中心的云计算已成为推动数字产品和服务发展的主要动力，涉及从基本的电子邮件服务到复杂的生成式人工智能（AI）在内的各种数字产品和服务。这种计算能力不是免费的午餐，数据中心中的每台服务器都需要耗费大量电力。有些数据中心支持人工智能、机器学习等先进领域的高数据处理需求，其功耗会特别高。这些数据中心中的耗电大户主要是为上述高级服务提供动力的GPU和加速卡。随着数据中心不断提升计算密度，散热挑战也随之增大。采取有效的热管理策略变得空前重要。

随着越来越多的公司进行数字化转型，数据中心面临着更大的压力。在提供高效计算能力的同时，它们需要尽量降低维护和运营成本。热管理成本是数据中心运营中的主要支出之一。有效的热管理可以通过延长设备使用寿命来降低长期维护费用。根据IT解决方案提供商Enconnex的数据，现代液冷系统的**运营费用**高达每千瓦冷却功率2000美元，而企业级数据中心冷却系统的投资很容易超过10万美元。显然，这些费用对当前专注于提升成本效益的企业来说是一个挑战。为了节省资本支出（CAPEX）和运营费用（OPEX），自然可以从热管理方面着手。

数据中心热管理的不为人知的故事在于光模块，这些模块用于实现机架式服务器、网络交换机以及数据中心之间的通信。服务器并非孤立运行，它们需要借助光纤链路相互通信，构成服务器集群，以提供下一代服务，如生成式人工智能。为了扩大这些服务，我们需要扩展服务器集群并提高它们之间通信的数据速率。随着新技术的出现和更高数据速率的采用，我们需要更大功率的光通道I/O光模块和有源电缆（AEC）收发器。例如，在112 Gbps-PAM4数据速率下，功率范围约为15至25瓦，而具有32个端口的大型企业交换机中的I/O光模块将消耗高达0.8千瓦的功率。如果使用相干（800G）

光学器件进行112G长距离通信，则每个模块的功率可以高达30瓦的水平。在这样大的功率下，I/O模块正在将传统的加压风冷系统推向其运行极限。

向**224 Gbps-PAM4**互连方式的转变将使每个通道数据速率翻倍。然而，这也会导致功耗的增加，特别是光模块在远距离相干链路上的功耗会高达40瓦。这是一个具有挑战性的问题，因为I/O光模块的功率在短短几年内从12瓦增加到40瓦，但模块的外形尺寸并没有改变。这实际上意味着功率密度增加了近4倍，因此需要采用新的冷却方法。实施液冷解决方案需要额外的投资和维护成本，但将创新型液冷解决方案整合到现有外形尺寸的设备中可以满足I/O模块中更大功率带来的散热需求。



随着I/O光模块的功率需求不断增长，系统设计人员和数据中心架构师正在考虑采用液冷技术来支持即将到来的224G速率设备。除了液冷之外，还有其它更先进的模块设计和特性化方法可以实现下一代高速网络互连。

本报告将探讨传统热特性管理方法的局限性，并研究在需要112G和224G链路的系统中对服务器和光模块采取的创新型冷却方式。

冷却技术现状：
传统散热解决方案

大功率系统通常采用主动冷却技术，即通过电动制冷系统从网络基础设施中移除热量。然而，主动冷却的使用带来了之前讨论的投资和维护成本。此外，为了实施主动冷却，需要有经验丰富的技术人员来安装并维护这些系统。在数据中心架构中，常见的主动冷却措施包括：

加压气流（或定向气流）是一种将空气直接从增压室泵入服务器机架的系统，包括在架空地板机房中。为了进一步促进空气流动，服务器和交换机可以配备专用风扇。然而，这些系统的能力有限，无法完全冷却服务器中的特定组件，如处理器和光模块。

液冷是一种常见的散热方法，它通过将具有高热容的液体循环到冷板上，并把冷板与机架安装系统中的发热组件连接来实现。这些系统可以使用水作为冷却液体，但通常更倾向于使用其它介电流体，例如油或丙二醇（PG-25）混合物。

当今采用的主动冷却方法

由于处理器和专用集成电路（ASIC）的高散热需求，现代数据中心的部署都依赖于主动冷却。就散热能力而言，主动冷却是最有效的冷却方法。当流体流向目标组件并有额外的无源组件作为散热辅助手段时，散热能力会得到增强。加压气流冷却和液体冷却都是现代数据中心设施中常见的散热方式。

这些系统以其卓越的散热能力而闻名，特别是当它们结合了将流体引导到发热部件上的引导装置时，散热效果更好。引导装置能够促进部件与冷却介质之间的热交换。更进一步的主动冷却方式是使冷却液直接作用于芯片上，特别是在数据中心中，高性能计算处理器会产生服务器中的大部分热量。因此，使冷却液直接作用于芯片上的冷却方法将会被广泛采用。

加压风冷是一种低风险的主动冷却方法，其方式包括根据需要将气流引导至与热部件直接接触的散热器上。当每个机架的功耗约为10千瓦时，通常可以采用加压风冷系统来处理热负荷。尽管大功率部件上可能采用液冷，但功耗很高的芯片和I/O模块仍然可以采用加压风冷作为一种冷却策略。

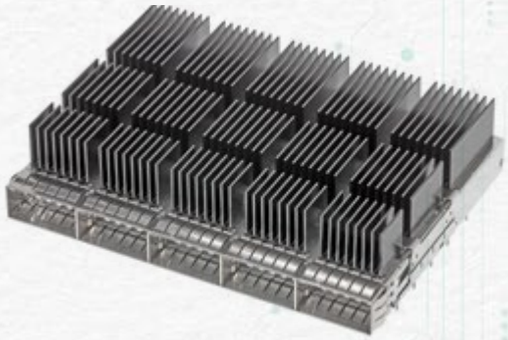
直接作用于芯片的液冷：该方案是一种用于数据中心的高效散热方案，特别适用于满足当今云环境中高性能处理器的散热需求。在这种冷却方式中，流体流经冷板，该冷板与芯片裸露的后表面连接，从而有效地从发热芯片中吸收热量。Enabled Energy, Inc.的杰夫·舒斯特尔（Jeff Schuster）认为，当每个机架的功耗达到25至50千瓦时，采用**直接作用于芯片的冷却方式**是一种理想的散热方案。

> 直接作用于芯片的液冷方式进一步提升了主动冷却的效果，尤其是在数据中心中，高性能计算处理器产生会大部分热量，这种方式就更加重要。

尽管这些主动冷却方式是最有效的，但它们也相当复杂，需要最大的维护量。在数据中心中，处理器、加速器和电源系统可以采用多种液冷方案进行冷却。然而，如今I/O光模块也需要进行冷却。对于服务器或交换机的这些装置，大多数运营商目前都依赖于加压风冷或被动冷却来对其I/O模块进行冷却。

无源部件可增强主动冷却的效果

一些无源器件在实现热传递和提供额外热容方面发挥着重要作用，从而增强了主动冷却的效果。常见的散热器和热管等无源器件主要用于主动液冷或加压风冷方案。芯片和GPU通常配备有散热器和主动冷却装置，如风扇或液冷系统。此外，在I/O光模块上安装散热器还有助于加压气流的进入，从而将热量从发热模块中带走。



带散热器的QSFP-DD屏蔽罩

为了有助于热量传递到散热器和光收发模块上集成的骑乘式散热器，业界针对温度最为敏感的内部组件（如激光器）采用了集成式热电冷却技术。通过帕尔帖效应将热量吸入散热器，然后热量通过气流消散。虽然这种方法在内部热传导机制中很有用，但并不能缓解224GI/O光模块更高的散热需求问题。

浸没式冷却

可以说，数据中心中最有效的液冷方式就是浸没式冷却，即，将整个服务器浸没在非导电液体中进行冷却。液体有足够的热容，并可循环到热交换器中。浸没式冷却提供非常有效的热冷却，粗略来看，每个机架的热冷却功率超过50千瓦。

浸没式冷却虽然效果非常好，但会带来如下所述的高风险和成本。

- **投资：** 浸没式冷却系统的设备和安装成本可能比加压风冷或液冷更高昂。这主要是因为它需要对数据中心架构进行全面革新，而空气和液冷系统可以通过改造方法进行部署。
- **空间要求：** 与浸没式冷却槽相匹配的机架通常比标准机架设备更宽更深。
- **相兼容的I/O模块和连接器：** 一定介电常数的流体会影响连接器的电阻抗。由于连接器的设计通常假设空气是冷却过程中的流体，因此浸没式冷却需要采用特殊的连接器和收发器模块。
- **相兼容的服务器：** 使用浸没式冷却的服务器是专门设计的，并非所有服务器供应商都提供该种服务器。
- **流体：** 虽然浸没式冷却流体在热容方面是有效的，但却需要特殊的循环系统来冷却流体。
- **维护：** 由于设备是专用的，这些浸没式冷却系统的维护成本往往很高。
- **泄漏风险：** 如果浸没式冷却系统发生灾难性泄漏，泛滥的液体可能会损坏其它区位的设施。
- **部件故障：** 某些部件附近的液体流量不足会导致部件高温，这会加速部件老化并造成其在早期出现故障。
- **环境影响：** 浸没式冷却中使用的流体需要定期更换，并需要通过正确的处置程序进行处置。

为了适应浸没式冷却，通常需要对硬件进行设计或调整。评估部件能否长时间浸泡于流体环境中运行是必要的。在评估112G和224G系统中I/O光模块的热需求时，可以直接将液冷技术应用于模块上以满足其散热需求，从而避免使用专门的浸没式冷却系统，并节省相关投入。

I/O光模块的散热挑战

服务器和机架式网络基础设施系统中的I/O光模块始终被主动冷却系统直接冷却，特别是受到来自机架式设备前面板的加压气流的冷却。机架安装设备中的散热设计需要平衡I/O模块的热管理与处理器或专用集成电路（ASIC）的散热，以避免I/O或ASIC工作温度的余量过大。应优化冷却策略以考虑处理器冷却需求和整体I/O光模块功率，这样做有助于实现适当的平衡，从而最大限度地提高系统的能耗效率。

链路长度与数据速率的关系：用于56G和112G的I/O光模块目前可以使用风冷。当以112G或更高的数据速率部署相干光学设备时，考虑到可插拔I/O光模块的功耗程度（大于33瓦）我们可能需要将液冷措施延伸到模块层级。

112G和224G这一代收发器仍以**IEEE 802.3标准**，中定义的标准链路长度为目标，因此系统设计人员和数据中心运营商不应期望标准化组织会仅仅为了适应光模块的更高功耗情况而改变链路长度标准。这意味着前几代光模块中已经存在的冷却需求预计将增加，而旧有的热管理方法可能会表现不佳。

> 预计新一代光模块的散热需求将超过前几代，而传统的热管理方法可能无法满足这种散热需求。

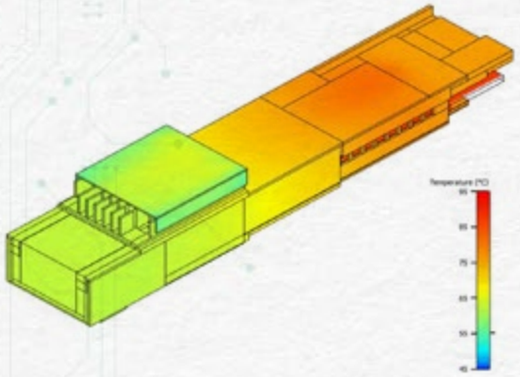
外形尺寸: 自光模块问世近20年来，可插拔光模块的外形尺寸一直没有改变，这给光模块带来了挑战。现在业界正在向224G发展，新一代的I/O光模块需要与现有的机架安装设备相兼容，即向后兼容，以实现升级。这意味着热密度将继续增加，这可能会导致加压空气作为冷却I/O光模块的唯一方法将不再可行。

散热: 将散热器贴附到I/O光模块上可以增强加压气流系统的冷却能力。然而，由于金属与金属之间的直接接触无法满足导热性能持久不变的要求，因此无法最大限度地传递热量。散热器与被散热器件之间的任何金属裸接触都是不可取的。过去几年光模块的功耗显著增加，尤其是在I/O模块上，预计每个模块的功耗将高达40瓦，这种情况进一步加剧了上述瓶颈。为了减小裸金属接触表面的热阻，可以将热界面材料（TIM）安装到骑乘式散热器上，以实现散热器与可插拔模块之间的紧密接触，从而提高传热效率。

将热界面材料连接到骑乘式散热器时，存在一个关键问题，即热界面材料的可靠性问题。当把模块插入其保持架中或从保持架中拔出时，模块的锋利边缘可能会刮掉热界面材料，导致每次插拔都会降低热效率。即使一次插拔操作不会使热界面材料失效，但是多次插拔后，热界面材料就会失效。此外，当这些模块暴露在严苛的现场环境中（例如由于电缆应力导致的倾斜插入），这种在导热耐久性方面的挑战会进一步加剧，因为这会使脆弱的热界面材料表面更加暴露于模块的锋利边缘。为了确保在反复配接情况下具有很可靠的导热性能，需要重新设计散热器接触方法，以便热界面材料能够耐受多达100次插拔操作。

保持架/散热器上的导热垫损坏

监测模块温度: 随着功率密度的增加，我们需要重新评估光模块的传统热鉴定方法。传统上，我们使用70摄氏度的外壳温度作为散热规范温度，即作为数字光学监测（DOM）温度的替代。然而，最近的研究表明，即使在70摄氏度的外壳温度下，模块内部的温度敏感组件仍然会留下几度以上的余量。这导致了关于系统热可行性的不准确结论，并导致冷却系统过度冷却。例如，在以I/O发热表现为限制因素的系统中，风扇将以高于所需速度运行，以满足外壳温度的限制要求，即使根据模块的内部组件温度有多余的余量（即被过度冷却）。新的散热鉴定方法(**本报告稍后将讨论**) 将有助于解决当前方法中存在的这一局限性。

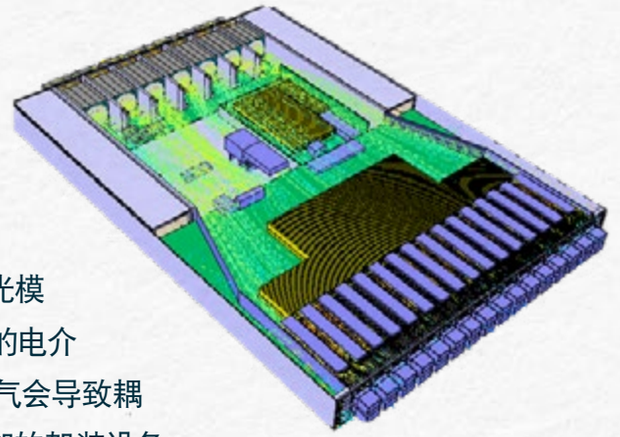


模块温度图示

仿真和测试: 在构建和部署散热系统之前，仿真/预测工程可用于优化系统设计、部件布局和冷却策略。通常，在最终确定机械设计之前，需要对光模块上的散热器和气流强制引入方式进行模拟，以优化整个机箱中的气流方式。架装服务器在高度和宽度方面是标准化的，大多数装置的外形尺寸是1RU标准尺寸。然而，其它部件（如芯片、外接卡、SSD固态硬盘等）的放置会影响流经机箱和一组I/O模块的气流路径，从而影响冷却效果。

对于I/O光模块而言，部件级仿真具有至关重要的作用，因为它能够发现模块主体上的热点。在进行模拟时，需要综合考虑模块本身的内部结构及其与各独立模块尺寸之间的关联性。当各个模块单独运行时，温度测试包括从接触式测量到红外热像仪测量等一系列测试方法。一旦了解了收发器中的热分布情况，就可以将其作为系统级仿真的输入，用于进行系统级测试和关联分析。

浸没式冷却: 浸没式冷却是一种有效冷却大功率112G和224G光模块的冷却方式。从热负荷的角度来看，这种方法是最有效的。但介电液体给模块连接器的连接带来了挑战，这主要体现在信号完整性方面。光模块和I/O连接器通常设计为假设周围的电介质是空气，因此用其它电介质替代空气会导致耦合效率低下。因此，在采用浸没式冷却的架装设备中，112G和224G通道需要采用与介电液体兼容的专用模块。当首选浸没式冷却时，设备供应量的减少和更专业的结构会造成每个机架单元的成本升高。



发热情况模拟系统

面向数据中心架构的创新型热管理解决方案

考虑到热负载的增加以及服务器和I/O光模块的向后兼容性对外形尺寸的限制，现有的液冷解决方案可能需要引用到模块上，以支持数据中心更高的数据速率和计算需求。在I/O方面，新型解决方案可以集成到服务器和交换机中，进而在不影响设备可靠性的前提下提供更好的散热效果。这是通过直接在模块上进行机械革新和液冷方式创新实现的，借助这种方式，我们可避免改变架装网络系统和可插拔模块中使用的标准外形尺寸。

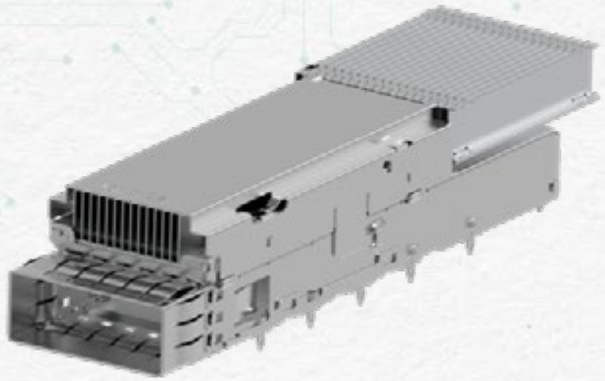
下落式散热器

为了最大限度地提高骑乘式散热器的传热效率，我们必须采用热界面材料来优化散热器基座与可插拔模块之间的金属对金属干式接触。如前所述，当将I/O光模块插入保持架时，其尖锐边缘可能会损坏热界面材料，从而降低模块的额定可插拔次数。因此，我们需要为散热器提供一种替代的散热接触机制，以确保在多次插拔模块的操作中保持热界面材料的机械稳定性和热稳定性。

Molex莫仕推出了一种创新的解决方案，通过在I/O光模块上使用下落式散热器（DDHS）来提升热管理效果。DDHS的突破性设计确保模块和热界面材料之间没有直接接触，有效地形成了一个悬浮散热器。只有在模块插入插座接近90%时，散热器才会与模块接触。在最后10%的插入过程中，散热器会“下落”到热界面材料上，并与没有任何锋利边缘的模块表面完全接触。这种设计使得TIM能够耐受100

多次的插拔操作。此外，下落式散热器还可以部署在不同的单排和堆叠屏蔽罩式配置中。

Molex莫仕出品的DDHS是当今传统骑乘式散热器解决方案的直接替代品。与已经优化的拉链翅片散热器解决方案相比，DDHS可以把35瓦设备的温度降低高达9摄氏度。



Molex的下落式散热器系统

这种下落式散热器解决方案提供了一种可靠的热管理方式，适用于标准外形尺寸的模块和架装设备。系统设计人员可以选择以下两种方式之一来利用这种可降温9摄氏度的改进型冷却措施：

- 使用具有相同功率（例如30瓦）的模块，只需降低系统风扇速度以利用DDHS的热余量，从而实现更高的功率效率。
- 在风扇以相同速度运行的情况下，冷却功率比30瓦高出5至7瓦的更大功率的模块（35-37瓦模块）。

DDHS解决方案使系统能够通过简单的直接更换操作来冷却更大功率的模块。

先进的液冷解决方案

即使在112G数据速率下，I/O光模块的功耗水平也使得加压风冷系统几乎达到了其能力极限。因此，在实施224G速率时，可能需要借助液冷来管理I/O光模块产生的热量。由于具有强计算能力的处理器已经在使用液冷，因此将大功率I/O光模块解决方案集成到现有冷却系统中是有意义的。这样，我们就能够对现有设备进行改造，以采用更高数据速率的新技术。

尽管液冷技术在数据中心行业中并不新鲜，但在可插拔I/O设备的部署方面，它确实带来了一些固有挑战。为了实现液冷的自然途径，我们通常会使用单独的冷板来代替单个骑乘式散热器。然而，这样做会形成多达32个入口和出口。在有限的1RU/2RU系统空间中，这种数量级别的管道是无法管理的。因此，我们需要设计一块可以冷却多个I/O端口的冷板。然而，这种方法面临的挑战在于每个I/O端口都有不同的累积公差，具体取决于模块高度、模块在壳体内的位置和基座高度等因素。虽然我们可以确保冷板与一个端口的良好热接触，但每个端口的不同堆叠情况让我们无法保证冷板与每个端口的充分热接触。例如，在1x6端口的屏蔽罩配置中，基本上需要所有冷板基座以及与冷板接触的所有模块表面具有完美的共面性。这意味着我们需要一个兼容的基座，该基座能够可靠地解决每个端口的公差问题，同时提供足够的压力来进行足够的热接触。

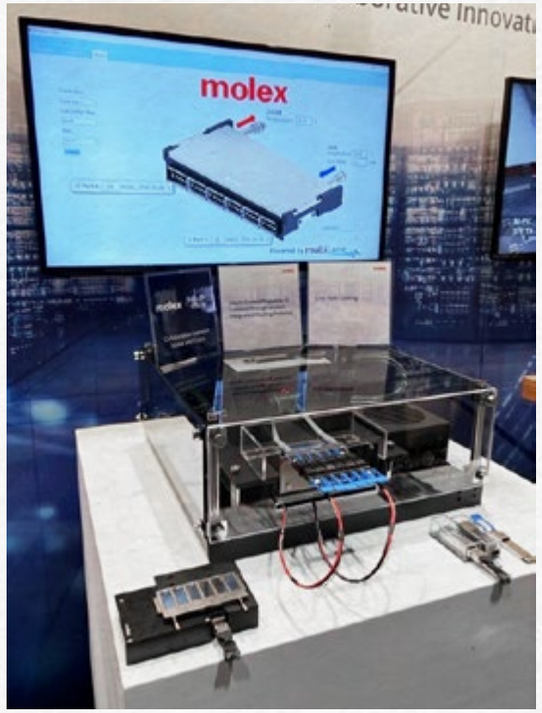
为了应对这些挑战，Molex莫仕开发了一种名为集成式浮动基座的液冷解决方案。在该解决方案中，每个与模块接触的基座都采用弹簧加压设计，可以独立移动，从而实现将单个冷板应用于不同的1xN单排和2xN堆叠屏蔽罩配置。这种独立移动的基座能够补偿每个端口的不同累积公差，同时提供所需的下压力以实现良好的热接触效果。

例如，下图展示的是1x6 QSFP-DD液冷解决方案。该方案配备了6个可独立移动的基座，能够适应每个端口的不同公差累积，同时确保良好的热接触（具备所需的下压力）。

Molex集成式浮动基座示例

通过这种集成式浮动基座，我们可以在无需使用热填隙剂或机械填隙剂的情况下实现I/O液冷。填隙剂会增加传导路径的热阻。在这个解决方案中，热量会直接从产生热量的模块流向基座，而基座则直接连接到流经冷板的液体。从理论上讲，这个路径是液冷解决方案可以实现的最短传导路径，有助于最大限度地降低热阻并提高传热效率。

虽然这种液冷解决方案很大程度上依赖于边界条件，但Molex已经证明，使用这种液冷解决方案，可以将功耗高达40瓦的模块冷却到规定温度范围内。



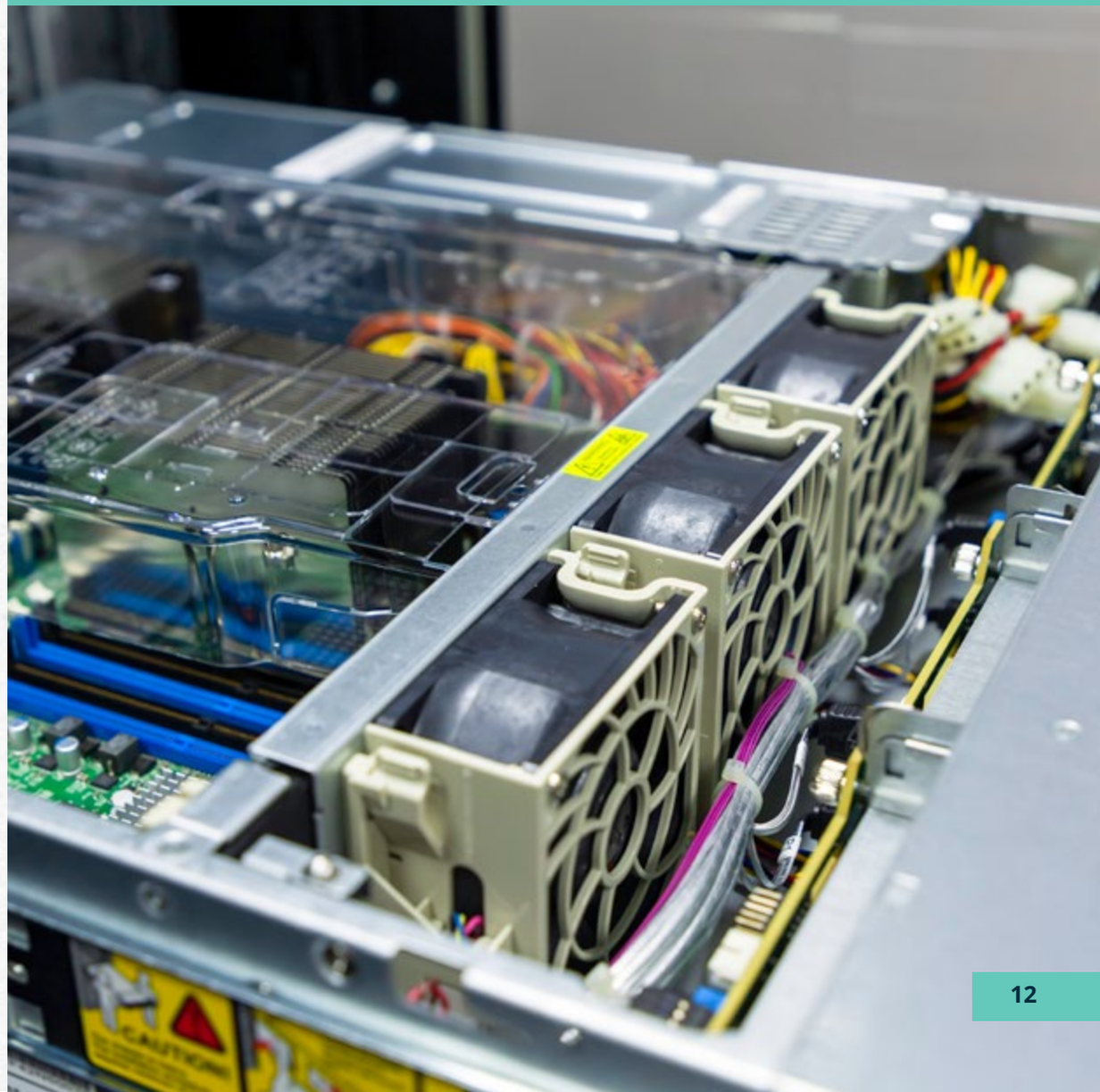
Molex液冷解决方案演示

下一代冷却策略的标准化和测试

在设计光模块冷却策略时，一个关键影响因素是在规范制定过程中以外壳温度作为模块温度的做法。然而，这些模块具有复杂的设计结构，因此外壳温度无法准确反映模块中关键组件的内在温度。实际上，模块温度是否过高取决于模块内部部件的温度是否超过限定温度。因此，以外壳温度为准的规范制定方法有待改进，这方面还有很多工作要做。

传统上，监测模块温度的方法是在模块外壳上选择一个监测点，通常位于散热器下方。系统冷却策略的设计旨在确保设备在运行过程中外壳温度（Tcase）不会超过规定温度，通常不超过75摄氏度。然而，由于散热器的存在，在设备运行过程中通常无法直接接触到该温度监测点，因此该点的温度并不能直接反映内部组件的实际温度。为了解决这个问题，内部传感器采用了数字光学监控（DOM）技术来报告Tcase值，该值可以通过软件管理界面（即CMIS）进行读取。

> 以外壳温度作为规范温度这种做法有待改进，有很多运行温度余量可以利用。



下表举例说明了使用外壳温度法所忽略的温度差。该表显示了典型模块的温度读数，这些模块位于堆叠屏蔽罩的下部端口处。通过将模块外壳温度限制值与确保模块正常运行和保持良好性能所需的关键内部部件的实际温度进行比较，在检查内部部件温度时，我们发现仍有额外的温度余量可以在设计时加以利用。

在这种情况下，可以重新设计冷却策略，以减少风扇负载，从而使机箱温度上升。这将允许系统利用一些额外的余量。使用模块外壳温度作为模块的限定温度仅有

模块	限定值	实际值	余量（ΔT）
Tcase外壳温度（数字信号处理器（DSP）上方）	75摄氏度	72.6摄氏度	2.4摄氏度
激光器	85摄氏度	76.4摄氏度	8.6摄氏度
TIA（跨阻放大器）/驱动器	105摄氏度	81.4摄氏度	23.6摄氏度
数字信号处理器（DSP）	105摄氏度	93.5摄氏度	11.5摄氏度

2.4摄氏度的温度余量。相反，如果将激光器温度限制值（激光器热裕度最小）作为模块的限制温度，我们会发现，在温升对激光器的任何性能影响被注意到之前，实际上存在8.6摄氏度的可用裕度。

因此，建议根据内部组件的最低温度裕度重新定义光模块的模块DOM读数，如下式所示。如前所述，在冷却系统设计中可以利用额外的裕量，同时保持与现有CMIS和系统软件的向后兼容性。DOM登记表中报告的值变为：

DOM = 75摄氏度 - (ΔT（激光）、ΔT（DSP）、ΔT（TIA）等）的最小值

这个提议的DOM定义有一个简单的解释：DOM值以及实际的温度裕度，应该基于模块操作环境中裕量最小的内部部件（例如，激光器件、光学器件、TIA、DSP芯片等）。报告DOM值过程中这种简单改进，有助于系统设计人员消除冷却系统架构中的多余余量，并为系统管理提供更好的模块控制。

为了实现数据中心冷却领域的未来而推动创新

Molex莫仕凭借在复杂数据中心环境热管理方面数十年的经验和广泛的专业知识，正在推出创新方式来应对更高数据速率所带来的日益增加的热挑战。尽管变革速度很快，但对系统设计和实施的制约因素仍然存在。为了适应标准化的设备外形尺寸，我们需要采用创造性解决方案，以克服空间限制难题，同时保持大量I/O端口不变。

Molex莫仕公司为I/O光模块开发了一种行业领先的冷却解决方案，该方案能够更可靠地散发高速运行系统产生的大量热量。与传统的热管理解决方案相比，这种特别设计的散热和接触方法在可插拔I/O光模块上表现出更高的可靠性和更低的复杂性。通过采用这种方法，我们不再需要使用繁琐的浸没式冷却方法。这为下一代数据中心互连架构的升级换代铺平了道路。

选择合适的供应商来解决数据中心热管理方面的复杂性问题，对于业界自信地向前发展至关重要。Molex为下一代数据中心架构带来了先进的功能，并奉行客户至上的协作方法，目的在于优化设备性能和效率。

creating connections for life

molex