|  |  |
| --- | --- |
| **分 数：** |  |
| **评卷人：** |  |

****

**研究生（数据中心技术）课程论文（报告）**

**题 目：数据中心技术能耗优化方法**

**学 号 M202073564**

**姓 名 岳跃盟**

**专 业 电子信息**

**课程指导教师 施展 童薇**

**院（系、所） 计算机科学与技术学院**

**2020年 12月25日**

数据中心技术能耗优化方法

岳跃盟1)

1)(华中科技大学计算机科学与技术学院, 武汉 430074)

1. 摘 要 数据中心是信息技术和云服务的基础设施，其高能耗问题一直受到广泛关注，并成为研究热点。而在其中冷却系统的能耗一直居高不下，故本文从冷却系统优化出发研究数据中心能耗优化。本文调研了三篇文献，Matt等人提出了虚拟熔化温度这个热感知的工作安排技术，以实现对数据中心热输出的更大控制，从而降低峰值冷却负载，达能节能的目的。Jiang等人针对温水冷却服务器架构，其能效会受到服务器之间热不平衡的影响，提出了一种混合冷却架构设计，它将热电冷却器结合到水冷却系统中，以细粒度的方式处理冷却不匹配。该混合水冷系统可以减少58.72%~78.43%的能耗来处理冷却不匹配。Zhu等人从混合水冷系统的服务器组件吸收的热量被直接喷射到水中，而没有被回收，导致能量浪费这一问题出发，提出了热发电这一经济且能源循环利用的温水冷却架构，热电发电机从“用过的”温水中获取热能，并发电供数据中心重复使用。以达到节能目的。

关键词 数据中心；能耗优化；虚拟熔化温度；水冷；热电致冷器；混合冷却架构；热电发电机

中图法分类号 TP302

A Survey on Energy Consumption Optimization Method

for Data Center Technology

Yue YueMeng1)

1)( School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

**Abstract** The data center is the infrastructure of information technology and cloud services, and its high energy consumption has been widely concerned and has become a research hotspot. Among them, the energy consumption of the cooling system has always been high, so this article starts from the optimization of the cooling system to study the energy consumption optimization of the data center. This paper investigated three documents. Matt et al. proposed a virtual melting temperature, a heat-sensing work arrangement technology, to achieve greater control over the heat output of the data center, thereby reducing the peak cooling load and achieving energy saving. Aiming at the warm water cooling server architecture, the energy efficiency of which will be affected by the thermal imbalance between the servers, proposed a hybrid cooling architecture design, which integrates the thermoelectric cooler into the water cooling system to handle cooling in a fine-grained manner Mismatch. The mixed water cooling system can reduce energy consumption by 58.72%~78.43% to deal with cooling mismatch. Zhu et al. The heat absorbed by the server components of the hybrid water cooling system is directly injected into the water without being recovered, resulting in energy waste. Starting from the problem of energy waste, they proposed thermal power generation, an economical and energy-recycling warm water cooling architecture, and thermal power generation. The machine obtains heat energy from the "used" warm water and generates electricity for reuse in the data center. In order to achieve energy saving purpose.

**Key words** Data center; Energy optimization; Virtual melting temperature；Water cooling；Thermoelectric cooler；Thermoelectric generator

# 引言

随着互联网规模的不断扩大和业务量的快速增长，云计算服务广泛应用于互联网相关领域。数据中心是云计算统一管理和资源调度平台，随着云计算的发展，数据中心的数量逐渐增多，规模也日趋庞大，但与此同时数据中心的能耗问题也日益显现。在美国，数据中心的年度能源消耗已达到910亿千瓦时，预计到2020年将增加到每年1400亿千瓦时。在全球范围内，能源消耗的快速增长已经成为数据中心管理和运营的关键。因此，在不改变数据中心服务质量的前提下降低能耗是一个新兴的研究和发展课题。

众所周知，数据中心不仅为信息技术设备，也为其他非信息技术设备(如不间断电源和冷却系统)消耗大量能源。例如，2014年美国数据中心的耗电量高达700亿千瓦时，占全国总能源成本的2%。然而，近48%的能源消耗是非信息技术设备，而79%的非信息技术能源消耗可归因于冷却系统。这一趋势推动了数据中心的革命，通过降低与冷却系统相关的电力消耗来提高其能效。文调研了三篇文献，第一篇文献是《Virtual Melting Temperature: Managing Server Load to Minimize Cooling Overhead with Phase Change Materials》，Matt等人提出了虚拟熔化温度（VMT）这一热感知的工作安排技术，以实现对数据中心热输出的更大控制，从而降低峰值冷却负载，达能节能的目的。第二篇文献是《Fine-grained Warm Water Cooling for Improving Datacenter Economy》，Jiang等人针对温水冷却可能会导致冷却失败的风险，并且由于缺乏精细的冷却控制，其能效会受到服务器之间热不平衡的影响这一问题，提出了一种混合冷却架构设计，它将热电冷却器结合到水冷却系统中，以细粒度的方式处理冷却不匹配，且开发了温水冷却策略，并根据工作负载的变化设计了自适应冷却控制框架，以使数据中心的水冷系统更加经济。针对第二篇文献提出的混合水冷系统的服务器组件吸收的热量被直接喷射到水中，而没有被回收，导致能量浪费这一问题，Zhu等人撰写了《Heat to Power: Thermal Energy Harvesting and Recycling for Warm Water-Cooled Datacenters》这一文献，在文献中提出了热电联产(H2P)，这是一种经济且能源循环利用的温水冷却架构，热电发电机从“用过的”温水中获取热能，并发电供数据中心重复使用。以达到节能目的。

# 虚拟熔化温度:管理服务器负载，利用相变材料将冷却开销降至最低

## 概述

### 2.1.1 研究背景

过去十年来，网络和云服务前所未有的增长刺激了对数据中心的巨大投资，数据中心也被称为“仓库级计算机”(WSCs) 。最大的数据中心设施每个消耗超过200兆瓦的电力，建造成本超过10亿美元，数据中心不仅在服务器设备方面，而且在电力、连接、设施和冷却基础设施方面都是巨大的投资。仅在美国，数据中心就消耗了2014年总发电量的2%以上。大量前期工作研究了如何构建更高能效的处理器，以及如何更好地从处理器和服务器中散热，但关于如何最大限度地提高利用率和效率，同时最大限度地降低从数据中心设施中散热的成本的研究相对较少。

即使是规模适中的数据中心，冷却系统的成本

每兆瓦临界功率可超过数十万美元。大型数据中心每年要花费数千万的资本成本和数百万的运营费用来为冷却系统供电和维护。仅在2015年，数据中心冷却资本支出总额就超过25.8亿美元，预计到2023年将超过60亿美元。研究服务器和数据中心冷却技术的前期工作证明了效率的提高，但无法解决日益增长的成本问题，因为有一个关键假设:功和热是耦合的，因此必须在做功的同时消除所有的热。

### 2.1.2 现有方法的不足之处

热时移(TTS) 通过将多余的热量储存在相变材料(PCM)中，并在稍后移除热量，将功和热分离。带动力系统控制模块的牵引力控制系统通过将一定数量的动力系统控制模块放置在机架式服务器的中央处理器插座的下风处来工作。在用户在线且负载高的高峰时段(中午至晚上)，变材料熔化吸收热量，以减少数据中心的热量输出。然后，在非工作时间(深夜和清晨)，当大多数用户都在睡觉，负载较低，并且有额外的冷却能力可用时，动力系统控制模块会重新制冷并释放储存的热能。

降低峰值冷却负载有两个主要优势:数据中心可以采用较小的冷却系统，同时仍能满足峰值负载的计算需求，或者同一数据中心可以在相同的冷却预算下运行更多和/或更热的服务器。这两种好处每年都可以节省数十万美元或数百万美元的资本支出，但这并不是一个通用的解决方案。

虽然用商业级石蜡实现的TTS既有热效率又有成本效益(每吨约1，000美元)，但它有关键的局限性。这些限制大多源于这样一个事实，即数据中心的最佳熔化温度取决于许多因素，从环境温度到工作负载，再到功率和交付限制。所有这些都可以从安装到安装，从一个季节到另一个季节，甚至从一天到一天。这是有问题的，因为:

（1）商业级石蜡只能在有限的熔化温度范围内购买，通常为40-60℃，但是，如果需要在此范围之外的熔化温度，分子级纯正链烷烃选项的成本超过每吨75，000美元。

（2）一旦安装，熔蜡温度无法调节。在负荷不导致蜡熔化的日子里，白天的冷却负荷没有变平，而在所有蜡熔化太快的日子里，峰值温度和冷却负荷没有降低。

（3）服务器的多年生命周期内，工作负载的功率和温度特性经常会发生变化。随着功率曲线的变化，理想(或所需)熔化温度也可能发生变化，从而需要新的蜡或完全脱离商业蜡熔化温度的范围。

在这三种情况下，在服务器中部署蜡几乎没有任何好处，TTS是一个无法适应的被动系统。

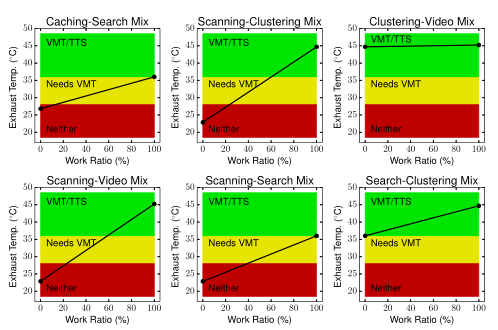


图1:热时移(TTS)可以在有限的温度范围内运行(绿色)，但是许多数据中心工作负载的混合不在此范围内。虚拟熔化温度(VMT)管理工作负载放置，以大大扩展部署相变材料的有用范围(绿色+黄色)。

### 2.1.3 作者提出的改进

这项工作中，作者提出了一种新的适应性技术，称为虚拟熔化温度(VMT)，以处理工作负荷功率混合物，TTS单独不能冷却。VMT这样做的方式是，它在(可配置的)不同于TTS的负载和平均温度水平下诱导相变材料熔化(从而导致热量再分配)，从而模拟熔点不同于部署蜡的物理熔点的蜡的操作。

这是通过重新平衡负载以将一些服务器的温度提高到相变材料的熔化温度以上，并将能量存储在选定的服务器中来实现的，具有降低冷却负载和降低功耗的好处。从战略上来说，采用VMT技术可以对蜡的熔化和冷却进行精细控制，使VMT能够在高温蒸汽发生器无法降低峰值冷却负荷的情况下降低峰值冷却负荷。

在这篇文章中，作者主要做出以下贡献:

（1）作者引入了VMT，这是一种通过控制工作负载布局来管理支持相变材料的数据中心的热特性的方法。我们介绍并讨论了两种工作负载放置算法来启用VMT。

（2）作者使用两种算法对VMT进行了横向扩展研究，使用之前验证过的模拟方法，在1000台支持PCM的服务器集群上执行VMT的设计空间探索。我们在运行五种不同工作负载的集群中研究了两种VMT算法，每种算法都具有独特的热特性。

（3）我们量化了VMT在集群和数据中心级别的影响，提供了关于如何在数据中心最佳利用VMT的有用讨论，并量化了VMT支持的冷却超额预订策略的潜在优势。

在集群级别，作者发现即使集群的平均热输出对于TTS来说太低，VMT也可以将峰值冷却负载降低12.8%。在数据中心级别，VMT将峰值冷却负载降低了3.2兆瓦，允许在相同的冷却预算下增加多达7339台服务器，或者在TTS无法提供可衡量的优势的情况下，数据中心以更小的冷却系统满负荷运行，节省了260万美元。

## 设计

VMT积极管理工作负载布局，以控制数据中心内的温度分布，提高部分服务器的温度以融化蜡(从而储存热量),同时降低其他服务器的温度以降低整个数据中心的峰值冷却负载。这就产生了一个“虚拟”熔化温度，虽然平均温度不能熔化蜡，但是我们可以从储存热量的一部分服务器子集中获取能量去融化它。VMT让系统或操作员主动控制数据中心蜡的熔化和冷却周期。

如果没有VMT，数据中心达到最佳物理熔化温度(服务器温度在材料熔化时保持稳定的点)的能力相对有限，最重要的是，在服务器的整个生命周期内保持不变，除非更换蜡(劳动密集型)。VMT是一种技术，允许数据中心改变数据中心的表观熔化温度来熔化蜡，即使蜡通常不会熔化。在不同的工作负载下，我们可以通过工作安排来产生热量不平衡。对于同类工作负载，我们可以通过负载不平衡来实现同样的目的；对于这项工作，我们假设前者。

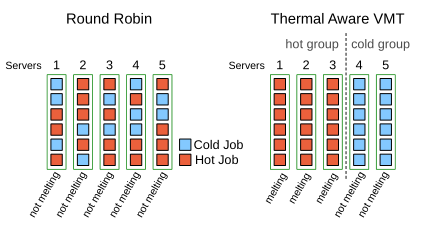
VMT还可以通过在已经融化的蜡的服务器子集中定位热作业来提高融化温度，保存蜡以应对即将到来的高温高峰。然而，这项工作的重点是降低熔点，而不是提高熔点。

作者介绍了两种调度算法来实现虚拟熔化温度:一种是热感知算法，它根据作业的热属性对作业进行分类和放置；另一种是蜡感知算法，它将作业从完全熔化的服务器上重新分配。

### 2.2.1热感知算法配置VMT

具有热感知作业安排的VMT(VMT-TA)建议将群集分为热服务器组和冷组，然后在热组中安排具有热配置文件的作业，而具有冷热配置文件的作业则放在冷组中(图2)。(请注意，热组和冷组服务器不需要物理群集:它们可以分布在整个数据中心，以保持相同的群集或DC级温度分布。)

根据工作负载所属的热配置文件，将作业放入热组或冷组:如果只有一个工作负载的服务器可以在峰值负载周期内融化大量的蜡，无论作业是否可以与其自身共置足够长的时间，只要它们可以与其他热作业并置，工作负载都被视为热的，VMT将尝试在热组中一起定位这些作业。否则，工作负载会被标记为冷，VMT会尝试将工作放在冷组中。

图 2热感知VMT调度

在这样的配置中，即使所有服务器内的平均温度或采用轮询/非热感知调度的平均温度不足以熔化蜡，热基团仍可以熔化蜡。

为了计算热组中放置的服务器数量，PMT-TA使用用户设置的分组值(GA)除以蜡的物理熔化温度的比率，公式如下:



其中，num\_servers是群集中的服务器数量，hot\_group\_size是热基团中的服务器数量。没有一个通用的解决方案将GV映射到虚拟熔化温度(VMT)，因为这种映射依赖于PMT以及工作负载功率分布和工作负载混合，但是对于给定的组合，可以通过实验导出映射。在计算热组大小后，冷组仅由剩余的服务器组成:



为了实施VMT-TA，首先根据热特性将工作负载类型分为热作业或冷作业。这可以使用封装的热传感器和/或功率传感器或型号(如英特尔RAPL)来完成。一旦部署，热作业将被放在热服务器组中，而冷作业将被放在冷组中。

在每个组中，作业均匀地分布在服务器中。这里必须注意确保每个组足够大，以支持其各自工作负载子集的峰值负载，否则必须丢弃或排队单个查询，导致服务质量下降。这可以通过动态调整VMT来修改组大小，或者如果一个组已满，则允许将作业调度到另一个组来处理。

### 2.2.2蜡感知算法配置VMT

在VMT-TA公司没有任何机制来处理热组中所有早期熔化的蜡的情况下，蜡感知算法配置VMT(VMT-WA)监控蜡的熔化状态，如果所有蜡在负荷峰值结束之前熔化，则自动增加热组的尺寸。

最简单地说，VMT-WA就像VMT-TA一样调度，直到热组中服务器上的蜡完全熔化。与VMT-TA不同，一旦服务器中的蜡完全熔化，VMTWA会将服务器从冷组移动到热组，在熔化的服务器上保持足够的负载以保持蜡熔化，并将额外的负载移动到新添加的服务器以继续熔化蜡(图3)。算法的详细描述如下。

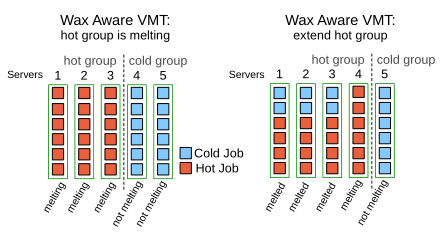


图 3蜡感知VMT调度

VMT-WT首先使用等式1计算热组和冷组的大小，与VMT-塔相同，但是组的大小随着蜡的熔化和冷却而动态更新。

群集计划程序通过扫描每台服务器上融化的蜡量，定期更新热组和冷组的大小。调度程序将每台服务器与蜡阈值进行比较，蜡阈值是服务器被认为完全熔化的熔化蜡的比例，并将阈值以上的每台服务器添加到完全熔化的服务器列表中。

计算完此列表中的服务器数量后，根据当前负载趋势，服务器从冷组中删除，并添加到热组中，在每次更新期间，调度程序从最小热组大小(等式1)重新启动，并按顺序添加服务器。在可能的范围内，我们不会在高峰期间将服务器从热组转移到冷组，因为冷却熔化的服务器会释放热量。

放置单个作业时，调度程序会考虑作业的热量分类(与VMT-TA相同)，但不会严格地将作业放置在相应的服务器组中。对于热作业，调度程序首先尝试通过考虑热组中当前低于特定蜡熔化量(蜡阈值)或低于蜡熔化温度的服务器子集来调度热组中的作业。在任一台这样的服务器上放置热作业将试图熔化更多的蜡或保持已经熔化的蜡熔化(两者对于减少冷却负荷都是有利的)。

如果没有符合这些特征的热组服务器(可能会突然出现负载高峰)，则服务器会从冷组依次添加到热组，直到热组中的服务器低于蜡阈值或熔化温度。如果不存在这样的服务器(所有服务器都被添加到热组的情况)，则作业将被调度到低于熔化阈值的任何服务器上，或者，除此之外，调度到任何剩余的服务器上。

要放置冷作业，计划程序会首先尝试将该作业放置在冷组中。如果作业无法放置在冷组中(热组增长时可能会出现这种情况)，调度程序会尝试将作业放置在热组中已经高于熔化阈值和熔化温度的服务器上，以最大限度地减少热影响。如果作业也不能放置在这些服务器中，则该作业将被放置在剩余的热组服务器之一中。

调度策略的这种排序只会在不受热约束的数据中心也将耗尽计算空间的情况下无法调度作业，因此作者不对这种情况建模。

VMT-WA需要了解集群中服务器的蜡的当前熔化状态，以适当调整热组的大小。蜡容器外部的单个温度传感器可以告诉我们蜡何时开始融化或冻结，然后我们在每个服务器上添加一个当前蜡状态的轻量级模型。该模型使用服务器中已有的中央处理器功耗和温度传感器，根据查找表来估计蜡的当前状态

实验

在图3中，作者绘制了使用VMT-TA和VMT-WA在100台服务器上扫描GV=10到GV=30的结果。两者都在GV=22时提供峰值降低，并且随着GV的增加，两者都紧密地向下趋势。对于这种特定的工作负载和PMT组合，这是最佳的GV，并且会因数据中心而异。然而，因为VMT赋予了控制全球变暖的能力，它提供了TTS所没有的必要程度的灵活性和适应性。

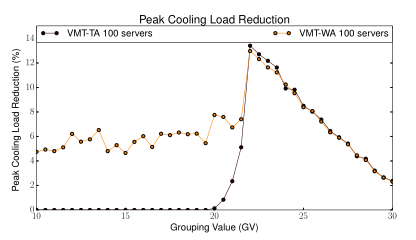


图4针对100台服务器，随着VMT-塔和VMT-瓦的GV调整，峰值冷却负载降低。两者都在GV=22时实现峰值冷却负荷降低。

为了评估VMT-TA相对于VMT-WA的优势，VMT-WA的优势在22以下最明显:当热组熔化太快且无法调整时，使用VMT-TA的峰值冷却负荷降低迅速降至零，而使用VMT-WA的降低立即降至6%左右，然后继续缓慢下降。

首先，两者在GV=22及以上时表现相似。这是因为在热组中的温度下降到熔化温度以下之前，有固定量的能量可以从空气中吸收，并且没有更多的热量可以吸收存储。即使VMT-WA中的蜡全部熔化，热团扩展，在理想的GV设置下，VMT-WA也不能比VMT-TA吸收更多的能量。理想的设置可能会随着工作负载组成或每日负载水平的变化而变化。

这说明VMT-WA的主要优势:稳健。在运营商可以每天准确预测负载的情况下，他们实际上可以每天将GV更改为最佳值。然而，对于VMT-TA，他们必须选择一个保守的值，因为选择一个太低的值的风险是极端的。有了VMT-WA，风险就更均衡了。

最后，作者使用先前工作中公布的方法量化了使用VMT降低峰值冷却负荷的潜在益处。峰值冷却负载降低带来的两个主要好处都来自冷却超额预订:要么该数据中心现在可以用更小的冷却系统实现相同的吞吐量，要么可以添加更多的服务器来提高相同冷却系统下的吞吐量。两者都大大节约了成本。

在1000台服务器的集群中，VMT-TA和VMT-WA都实现了12.8%的峰值冷却负载降低，相比之下，仅使用TTS的峰值冷却负载降低不到0.2%。考虑到第四部分中的25兆瓦数据中心，一个完全定制的冷却系统需要在峰值负载时从数据中心移除25兆瓦的热能。(以下成本节约包括在数据中心的每台服务器上部署蜡的成本估算。)

将峰值冷却负载降低12.8%可将数据中心的峰值冷却负载从25兆瓦降至21.8兆瓦，并实现12.8%的较小冷却系统。根据冷却系统成本估算，这在数据中心的整个生命周期内节省了2，690，000美元的成本[14]。

(请注意，为TTS部署熔化温度接近30℃的正石蜡，以实现相同的峰值冷却负载降低，成本约为1，000万美元，是VMT部署商用蜡成本的四倍多。)

对于更保守的方法，使用VMT-WA的数据中心可以选择仅减少6%的冷却系统，以考虑负载变化。冷却系统尺寸减少6%，仍可节省1，260，000美元的成本。

或者，在相同的冷却系统规模下，降低的峰值冷却负载可用于向数据中心添加更多服务器。VMT采用VMT-TA或VMT-WA最佳峰值冷却负载降低方案，使服务器数量增加14.6%，每个集群增加146台服务器，或者在25兆瓦的数据中心增加7339台服务器。VMT-WA保守的6%应用程序也提供了巨大的优势，使服务器增加了6.4%，每个群集增加64台服务器，或者数据中心增加3191台服务器，而不增加冷却资本支出。

如本文所评估的，减少冷却能力或更大的过度配置的好处来自于使用VMT来减少峰值(年)冷却负荷。可能还有其他每天控制熔化温度的能力所带来的好处，例如利用便宜的非高峰功率或绿色电源（当冷却能量可以暂时移动时）

# 细粒度温水冷却，提高数据中心的经济性

## 概述

### 3.1.1 研究背景

水冷作为一种有前途的冷却技术，已经成为许多数据中心降低冷却成本的主要选择。与空气冷却相比，水冷却实现了高密度的服务器实施和更高的冷却效率。图5显示了数据中心中的典型水冷系统。它主要由两个液体回路组成:(1)内部有冷却剂的内部回路将热量从服务器带到热交换器，(2)内部有设施水的外部回路将热量从热交换器传递到室外的水冷却塔。这座塔通过蒸发使水冷却下来。通常，数据中心使用冷水(例如7℃~10℃)来冷却服务器。在塔移除一部分热量后，冷却器继续从水中移除剩余的热量。这种水冷系统可以服务数百甚至数千台服务器，与空气冷却相比，它可以平均减少22%的冷却能量。

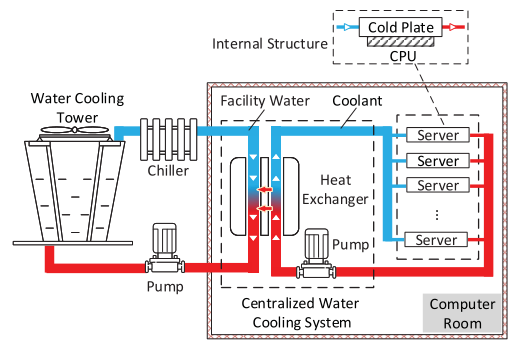


图5 数据中心的水冷架构

最先进的水冷却系统不能将冷却能量减少到零，因为冷却塔带走的热量非常有限，并且它需要冷却器(消耗大量能量)来帮助进一步冷却设施水，尤其是在热天。因此，通过在较高温度下运行中央处理器来提高设施水的温度可以显著降低冷却器的能耗。据报道，将设施水的温度从7℃~10℃提高到18℃~20℃可以导致冷却能量节省大概40%。由于当前数据中心的利用率较低，建议采用更积极的策略，即温水(例如40~45℃)冷却，以降低冷却成本。柯立特报告说，温水冷却平均可在六个月内实现快速投资回报。此外，ASHRAE热能指南“W5”建议使用超过45℃的水来冷却数据中心，以便出水足够热来加热建筑物，用于数据中心热量回收。然而，当一些服务器在温水冷却中获得高利用率时，它们将在几秒钟内超过安全操作温度，而冷却器需要相对较长的时间(例如几分钟)来冷却水并将其输送到过热的服务器，这导致冷却延迟/不匹配以及冷却失败的风险。尽管温水冷却很有吸引力，但实施起来有风险。

于成本和技术上的困难，在每台服务器的集中式水冷系统中安装水泵来实现细粒度的流量控制是不现实的。另一种架构是为每台服务器构建独立的水回路，以实现细粒度的流量控制(例如，IBM大型机)，这种架构成本很高。因此，当前水冷却系统的冷却控制是集中的(即，控制供应水的温度)。它不能为每台服务器分别提供细粒度的冷却，这就带来了温水冷却的热点问题，即一些服务器超过了安全运行温度，而另一些则没有。热点阻碍了温水冷却的能效，因为冷却器必须根据服务器的最高温度降低水温，而其他服务器可能不需要这种“冷”水。

### 3.1.2 现有方法存在的问题

许多基于软件的解决方案被认为可以优化数据中心的冷却效率。工作负载延迟可以推迟一些散热，以降低冷却峰值。但这种策略不能应用于期限紧迫的工作负载(如交互式工作负载)，它要求冷却高峰之后必须是冷却低谷，以便为延迟的工作负载提供冗余冷却能力。工作负载节流(即CPU频率缩放)可以缓解CPU的功耗及其冷却需求，但它不能应用于有性能保证的工作负载。例如，亚马逊EC2中的计算优化实例c5承诺提供至少3.0 GHz的CPU频率。服务器即使闲置也有静态功耗。因此，通过工作负载整合关闭利用率低的服务器可以消除服务器的空闲功耗，同时降低整体散热，但由于整合后服务器的利用率变高，这需要较低的水温，从而增加了冷却器的使用。服务器的电源优化不一定带来冷却优化。一些在线服务，如网络搜索，要求所有服务器保持运行，而不管流量强度如何。工作负载整合可能不适用于这种交互式工作负载，因为由于排队效应，响应时间通常会受到更高利用率的影响。工作负载平衡也可以考虑帮助缓解热点问题，而不一定带来热平衡。工作负载平衡通常旨在提高性能(例如，最小化作业的完工时间)。它的平衡通常意味着处理速度的平衡，而不是利用率的平衡(例如，在异构数据中心)。如稍后在Sec中从2011年到2017年的真实世界集群轨迹所示。2、集群中的利用不平衡是一种普遍现象。

尽管存在上述限制，但热感知工作负载管理策略可能与其他优化策略(例如，网络感知工作负载布局)相冲突，从而导致性能和冷却能量之间的权衡。数据中心的工作负载通常由更关心性能的软件工程师管理，而冷却系统由不同的部门或工程师管理，这使得热感知工作负载管理在实践中更加困难。

### 3.1.3 作者提出的改进

为了应对上述挑战，作者提出了一种混合水冷系统，该系统将每个CPU与热电冷却器(TEC)集成在一起，为单个服务器动态提供额外的冷却。具体而言，作者在该篇文章中做出以下贡献：

（1）作者提出了一种混合式水冷却系统，该系统将TEC整合到现有的水冷却系统中，并构建了一个原型来验证其可用性。TEC可以以精细的方式为水冷系统提供额外的冷却能力，从而允许安全有效地运行温水冷却策略。

（2）通过探索数据中心工作负载的空间和时间分布，设计了一种自适应冷却控制机制，该机制可以根据工作负载的变化调整冷却供应策略。具体来说，作者的自适应冷却控制机制使用冷却器来提供基本冷却能力，使用TEC来提供动态冷却能力来处理其余的冷却不匹配。

（3）作者基于硬件原型和阿里巴巴和谷歌的服务器跟踪来评估这个系统。结果表明，与传统的水冷却系统相比，作者的混合水冷却系统在处理冷却不匹配时可以减少58.72%~78.43%的能耗

## 设计

### TEC

热电冷却器(TEC)是基于半导体的电子元件，其通过珀耳帖效应工作，如图6所示。当向TEC施加直流电时，热量将从一侧转移到另一侧，产生两个相对的面:冷面和热面。冷端可用于冷却。

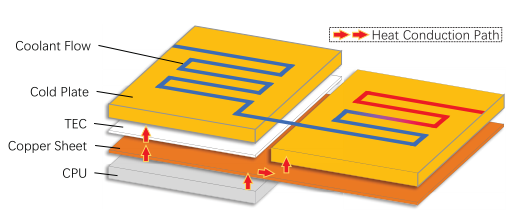


图6 服务器内部的TEC集成

图6示了作者的混合水冷设计。原来的水冷只包含一个冷板，直接连着CPU。我们的混合水冷系统还有三个组件:铜片、TEC和另一个冷板。CPU散发的热量传递到铜片上。当CPU温度小于安全工作温度时，TEC关闭，不消耗任何能量。然后热量从散热器铜传递到冷板(即，图6中的右导热路径)。请注意，TEC关闭时几乎不导热。

当CPU利用率高，冷却液太热而无法冷却CPU时，TEC开始工作，将热量传递给另一个冷板(即图7中的左导热路径)。这样就可以在TEC和水冷之间无缝切换冷却方式。

将TEC直接与铜片接触以实现水冷和TEC冷却之间的切换有一个缺点:右侧冷板中的热量可能会返回铜片，因为TEC工作时铜片会变得更冷。但是作为我们在原型上的测量，TEC仍然实现了较高的冷却速率。

### 3.2.1系统结构

如图7所示，在我们的混合水冷系统中，它有三个主要部分。第一部分是冷水机组和冷却塔，与常规水冷系统相同。第二个是额外的技术执行委员会模块，它可以在不降低水温的情况下实现细粒度的实时冷却。最后一部分是控制器，根据工作负载变化做出适当的冷却决策，并优化冷却能效。

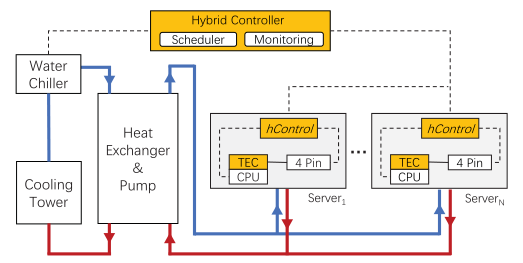


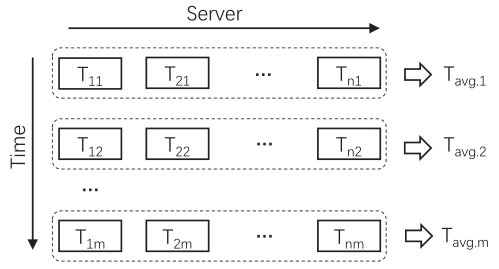
图7 混合水冷结构综述

显然，整个冷却系统的效率高度依赖于调度器的冷却策略。

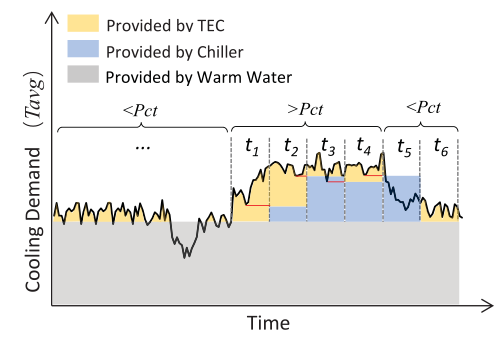
### 冷却管理

假设冷却塔可以持续提供TWarmWater温度的温水，且该温水可以将CPUi的温度在时间j内冷却到温服Ti,j。一旦服务器的利用率突然增加，并且其温度超过预定义的安全温度Tsafe，TEC可以实时响应突然的冷却不匹配，消除传统水冷的冷却不匹配。

为了解决冷却需求预测问题，尤其是在多个用户可能不定期提交工作的公共数据中心。作者提出了一种自适应冷却控制机制。对于每个时间实例j，我们可以得到CPU i2的温度Ti，j。然后我们有一个如图8(a)所示的中央处理器温度图。我们将超过预定安全温度的温度Ti，j称为热点，需要冷却。为了便于说明，我们使用图8(a)中的平均中央处理器温度来测量每个时间实例的总体冷却需求。我们仅在热点百分比大于预定义阈值Pct(例如80%)时才启动冷却器。如图8(b)所示，当热点的百分比大于Pct时，冷却不匹配可以如下处理:



1. 冷却控制间隔内服务器的温度图



1. 混合水冷系统中的自适应冷却控制。

图8 自适应混合冷却控制

步骤1.在第一个冷却控制间隔中，当热点的百分比大于Pct(例如t1)时，我们只使用TEC来冷却服务器，因为TEC可以实时冷却CPU。

步骤2 .在下一个间隔(例如，t2)的开始，如果在前一个冷却控制间隔(即，t1)中热点的百分比大于Pct，则冷却器根据在前一个间隔中的最低冷却需求来提供冷却能力。也就是说，将冷却器设置为将水温降低ΔT = min { Tavg,1，Tavg,2…Tavg,m}其余冷却不匹配由TEC处理；否则，请转到步骤3。

步骤3。如果前一间隔中的热点百分比小于Pct(例如t5)，则只有TEC处理该间隔中的冷却不匹配(例如t6)。

上述冷却控制基于前一个冷却控制间隔中的CPU温度。它可以在不预先了解工作负载的情况下动态安排冷却容量，并实现更智能的冷却控制。

## 实验

由于作者并没有配备混合水冷系统的真实数据中心，我们根据从原型获得的能源和热量模型进行实验，并通过使用阿里巴巴和谷歌的集群跟踪来模拟能源使用情况。

为了更好地评估我们的混合水冷却系统和冷却控制机制的性能，我们在评估中设置了三种类型的跟踪和三个基线。

激烈:这个痕迹来自阿里巴巴集群，里面包含了12小时内1313台服务器的CPU使用情况。它的整体CPU使用率有剧烈而频繁的波动。

不规则:Google trace提供一个月内12.5k服务器的CPU使用率。我们选择1000台服务器24小时的CPU使用情况来构建异常工作负载跟踪。它的整体CPU使用率比较常见，但偶尔会出现高峰。常见。

同样，我们从Google trace中选择另外1000台服务器，持续24小时，组成一个整体冷却需求，其波动程度小于Drastic。这三个痕迹将在后面的评估结果中演示。

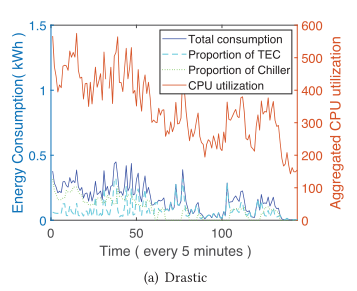
最佳基线(hybridOpt):如果我们预先知道即将到来的冷却需求或实施一些预测算法，那么我们可以提前找到最佳冷却控制解决方案，从而最大限度地减少我们的混合冷却能量。由于我们已经有迹可循，通过尝试不同的冷水机组设置(即尝试不同的冷水机组水温降ΔT从0)，我们可以得出以最小冷却能量为最佳基线的策略。

冷水机组基线(冷水机组):对于传统的水冷系统，只能用冷水机组来处理冷却不匹配的情况。同样，如果我们预先知道未来的冷却需求，我们会根据最高的冷却需求来设置冷却器输入冷却控制间隔，并将其冷却能耗作为另一个基准。

TEC基线(TEC):我们只能使用TEC来处理冷却不匹配。这将导致另一个能源消耗作为新的基线。

### 2.3.1测试结果

混合水冷却系统的能量模式:图13显示了我们的混合水冷却系统在三种不同工作负载下的工作负载轨迹(剧烈、不规则、常见)和相应的能量消耗。我们可以看到总能耗随着总CPU利用率的变化而变化。它避免了提供不必要的冷却能力和能源浪费。需要注意的是，聚合CPU利用率并不能完全描述CPU的散热需求，因为它遗漏了局部热点的信息。特别地，图13(a)中冷却器的能量比例高于图13(b)和图13(c)。这是因为它的整体CPU使用率高得多，导致触发冷水机的机会更多。在图13(b)和图13(c)中，我们可以看到冷水机组的能耗最多保持为0，只有在高峰到来时才会触发。这是因为它的整体CPU利用率在大多数时间相对较低(小于200)，并且热点的百分比在这些时期小于我们设置的80%。因此，冷却器不会启动。



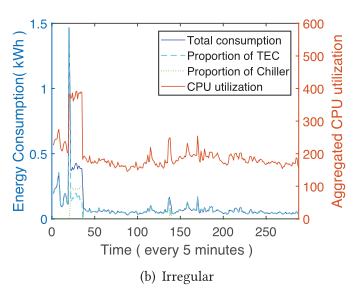
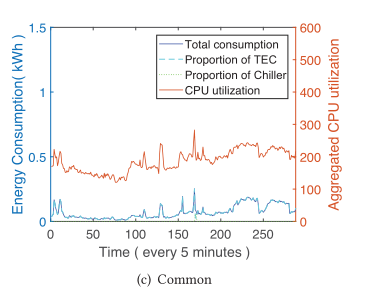
 

图9混合水冷却系统的能量使用模式(Twater = 500C，Pct = 80%，冷却控制间隔为15分钟)

与不同策略相比的节能:图10显示了不同策略处理冷却不匹配的能耗。如我们所见，与传统水冷(Echiller)相比，作者的混合水冷(Ehybrid)对于剧烈、不规则和常见的工作负载分别可以减少58.72%、74.48%和78.43%的冷却能量。

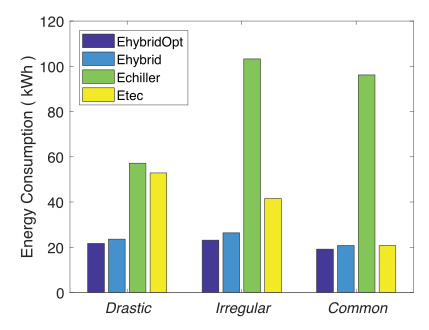


图10处理不同策略冷却不匹配的能耗(Twater = 500C，Pct = 80%，冷却控制间隔为15分钟)。

# 热能发电：温水冷却数据中心的热能收集和回收

## 概述

### 4.1.1研究背景

过去，工业界和学术界主要通过降低信息技术成本和冷却成本来提高电力使用效率(PUE) 。例如，谷歌将平均PUE值降低到接近1.1，代表了一种高效的设计。与此同时，现有技术的局限性没有给PUE减少留下太大的空间。因此，绿色电网提出能源再利用效率(ERE) 来衡量数据中心的能源再利用效益，其形式为:

最大限度地提高能源再利用，使这一比例低于1，促使数据中心从业人员更加注重再利用废弃能源，以提高能效。因此，来自数据中心的大量热量将不再被视为多余的副产品，它可以转化为能源再利用的机会。例如，废热可以通过以下三种方式收集和回收。

集中供热:集中供热是一种常见的余热再利用解决方案。数据中心需要与分布式哈希表(DHSs)合作，这是一个庞大的项目，既涉及硬件(如复杂的管道布置、城市供热系统重新设计和规划)又涉及软件(如数据中心与分布式哈希表之间的双赢合作机制)设计。这种补救措施不适用于中纬度和低纬度国家，因为这些国家全年都没有热需求，因此它们需要承担巨大的成本来建造一个成熟的城市供热系统。还有，热量的作用有限。如果能量是以电的形式存在，它在需要电而不是热的数据中心中可以发挥更大的作用。

热转化为电:以前的工作使用热电发电机(TEG)将热转化为电。小规模的TEG被放置在智能手机的芯片上或服务器的中央处理器上，以获取热量。然而，TEG是由绝热材料制成的。将TEG直接连接到中央处理器可能会降低中央处理器的散热效果。此外，TEG的冷侧直接暴露在空气中，会逐渐降低TEG热侧和冷侧的温差，从而降低其发电量。

热电冷联产(CCHP):这种解决方案指的是同时发电和有用的加热和冷却，主要来自气体的燃烧。在CCHP，废热通常被收集为蒸汽，既可用于加热，也可通过涡轮机将蒸汽输送到用于冷却的吸收式制冷机来发电。

然而，由于CCHP的建设和维护成本要高得多(例如，CCHP系统设计、设施部署和复杂的管道布置)，作者建议采用H2P的TEG集成系统，该系统易于以非常低的成本安装在现有的水冷数据中心。此外，大多数CCHP系统需要气体作为能源供应，并具有更严格的防火和防爆保护，而H2P不需要额外的能源用于现有数据中心的发电。因此，无论是在建筑/数据中心规模，还是在服务器/中央处理器规模集成TEG，H2P都拥有CCHP无法替代的独特优势。在H2P，作者将TEG置于服务器/中央处理器规模(温度最高)，以更好地重复利用废热。值得注意的是，CCHP和TEG集成解决方案可以相互结合，以进一步提高能效。

### 4.1.2 现有方法的不足之处

要进一步提高水冷数据中心的能效，关键在于降低冷水机组的能耗。据报道，通过简单地将设施水的温度从7℃~10℃提高到18℃~20℃，可以节省高达40%的冷却能耗。基于数据中心的服务器大部分时间利用率较低的事实，建议采用温水(例如40°C~45°C)冷却策略来降低冷却成本。然而，当一些服务器突然过载时，它们可能在几秒钟内超过安全运行温度，而冷却器需要相对较长的时间(例如几分钟)来冷却水并将其输送到过热的服务器，这可能导致冷却延迟/不匹配以及冷却故障的风险。

尽管有报道称高温对硬件可靠性的影响不是很大，但热水冷却系统中的热点问题仍然需要一个可行的解决方案。最近，Jiang等人提出了一种混合冷却架构，该架构将每个CPU与热电冷却器(TEC)集成在一起，以动态地为各个服务器提供额外的冷却。通过细粒度的自适应冷却控制，它可以根据工作负载的变化经济地处理冷却不匹配的情况。通过这种方式，温水冷却成为一种有前途的方法，它不仅可以减少对昂贵和耗能的冷却器的需求，还可以为收集和回收“用过的”温水中的热量提供更多的可能性。例如，ASHRAE热量指南“W5”建议使用45℃以上的水来冷却数据中心，以便出水足够热，用于热量回收和建筑物供暖。

然而，通过区域供热回收热量有其局限性。一方面，需要一个巨大的项目来将热量传递到所需的区域(例如，复杂的管道布置、城市供热系统的重新设计和规划)。另一方面，从一个季节到另一个季节，从一个地区到另一个地区，对热量的需求并不总是很大。事实上，区域供暖在冬季寒冷的高纬度地区更为常见，如北欧；而像新加坡这样热带地区一年四季都很热，因此对供暖的要求也很低。此外，大多数数据中心位于温暖地区，从春季到秋季(例如，华盛顿特区、旧金山和休斯顿的4月至10月)数据中心的高峰时段热容量超过住宅的热量需求，导致热量不匹配。但同样重要的是，热能很难储存。

### 4.1.3 作者提出的方案

面对这些问题，作者提出可以利用一个简单却强大的组件，即热电发电机(TEG)，它可以在两侧温差下发电。用来回收从“用过的”温水中收集的热能，将热量转化为电能，从而为数据中心本身提供部分电力。以前的一项工作考虑在智能手机的芯片上部署小规模TEG，以获取热量并将其转化为电能。受此启发，在这篇文章中，作者主张从服务器组件吸收的热量可以回收，以进一步提高温水冷却数据中心的能效。因此，作者提出了热电联产(H2P)，这是一种经济且能源循环利用的温水冷却架构，TEG从“用过的”温水中获取热能，并发电，用于温水冷却的数据中心。在这方面，作者的贡献总结如下:

（1）作者探索了TEG作为一种新的部件来回收温水冷却数据中心的废热，并介绍了TEG的特点。作者建议采用热电联产(H2P)，这是一种经济且能源循环利用的温水冷却架构，它将TEG纳入现有的水冷却系统。作者考虑了TEG的位置和规模，并仔细设计了热能收集和回收模块。

（2）作者建立了一个硬件原型来验证H2P的可行性。在样机上进行了经验测量，以证明在不同因素(如入口温度、流量和中央处理器利用率)影响下TEG的发电能力。

（3）由于发电受到测量结果的限制，我们提出了一些有效的优化方法，包括经济的水循环设计、精细调整冷却设置(即进水温度和流量)和动态工作负载调度(即工作负载平衡)，以增加TEG产生的电力。

（4）作者对H2P的总体拥有成本(TCO)进行了全面分析，总体拥有成本最高可降低0.57%(即，拥有100，000个中央处理器的数据中心每年可节省约410，000美元)。

## 设计

### TEG

热电发电机(TEG)是一种可以将热量转化为电能的装置。它由几对夹在两个陶瓷芯片之间的n型和p型半导体组成。当两个芯片之间存在温差时，TEG产生电压。温差越大，电压越高。这种现象被称为塞贝克效应。

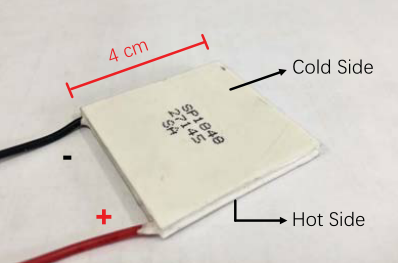


图11 热电发电机TEG

考虑到中央处理器的安全性和发电能力，作者将TEG模块放置在每个中央处理器的出口处，这是整个循环中最热的地方。

### 4.2.2 热电(H2P)系统架构

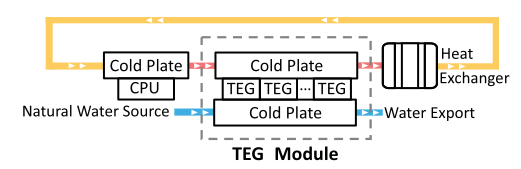


图12 热电(H2P)系统架构

图12描绘了H2P的系统架构。TEG模块(详见图13)由几个串联在每个服务器出口的TEG组成。为了利用TEG产生的电能，输出电压必须达到一定值。

TEG模块由两个冷板夹在中间，这两个冷板连接成两个独立的液体回路。特别是，在TEG的热

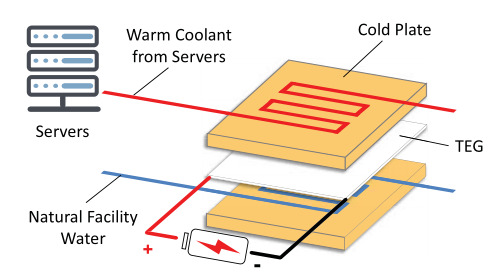


图13 一种热电发电（TEG）模块

端，冷板中的水是来自服务器的热冷却剂。这股水流已经被用来冷却服务器的CPU，带走CPU的热量后达到40°C以上。流经TEG模块后，流向热交换器，并将热量传递给FWS。在TEG的冷端，天然水源是生活用水或来自大自然的自来水，温度在20℃左右。

4.2.3经济的水循环设计

作者将H2P视为一个拥有1000台服务器的同构数据中心。通常，一个或几个机架中的服务器由一个CDU控制，共享同一个水循环，配备有一个冷却器和一个集中泵。在这个循环中，我们假设每个服务器的进水温度和流量是相同的。提高发电量的关键在于提高Twarm\_in，而Twarm\_in的上限由本次循环中最热的服务器决定。因此，应该考虑在水循环中布置多少个服务器。显然，垄断一个循环的每个服务器是最节能的，可以产生最多的电力，因为这样我们可以调整每个CPU的Twarm\_in。然而，考虑到技术限制和资本支出，为每台服务器配备一台冷却器和一台泵是不切实际的。

在实践中，我们假设一个水循环有n个服务器，所以1000个服务器被分成1000/n个水循环。在第i次水循环中，第j次CPU利用率为Tcpui,j，CPU的安全温度为Tsafe,可以预定义和调整(如CPU最高工作温度的80%)。请注意，在接近最高温度的情况下进行预期的操作可能会导致CPU性能下降并缩短CPU寿命。

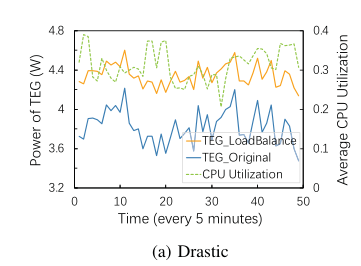
## 实验

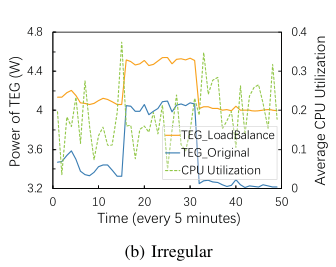
为了评估H2P的表现，我们设置了三种类型的跟踪:

剧烈:这个痕迹来自阿里巴巴集群，里面包含了12小时内1313台服务器的CPU使用情况。CPU利用率的变化有剧烈而频繁的波动。

不规则:Google 痕迹提供了12.5k服务器一个月的CPU使用情况。我们选择1000台服务器24小时的CPU利用率来构建异常工作负载跟踪。CPU利用率的变化比较常见，但偶尔会出现高峰。

常见:同样，我们从Google trace中选择另外1000台服务器，持续24小时。CPU利用率的变化比较常见，波动很小。





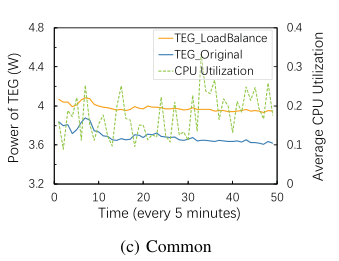


图14 三种CPU利用率下的发电量。

作者比较了H2P在两种方案下的发电能力:

原始TEG:基准数据中心配备TEG并调整冷却设置，但不考虑任何工作负载调度方法。

TEG负载平衡:通过增加工作负载平衡策略对原始TEG进行优化。

图14显示了在三种不同的工作负载和两种不同的工作负载调度方法(原始TEG和TEG负载平衡)下，H2P的工作负载轨迹(剧烈、不规则、常见)和相应的发电能力。如图14a所示，当中央处理器利用率高时，H2P的相应发电容量低。这是因为当一台服务器的CPU利用率高的时候，它的CPU温度也高，而进水温度要低到足以让服务器降温。在这种低进水温度下，发电能力因此较低。在图14b和图14c中，H2P的发电量显示出相同的趋势。特别是原始TEG在三种工作负荷轨迹(剧烈、不规则、常见)下的平均发电量分别为3.725瓦、3.772瓦和3.586瓦(平均为3.694瓦)。发电量的峰值分别为4.210瓦、3.935瓦和4.035瓦。工作负载均衡优化后，三条轨迹的TEG负载均衡平均发电量分别为4.349瓦、4.203瓦和3.979瓦(平均为4.177瓦)。峰值分别为4.595瓦、4.554瓦和4.082瓦。因此，我们可以得出结论，工作负载平衡策略可以有效地将平均发电量提高约13.08%(即从3.694瓦提高到4.177瓦)。

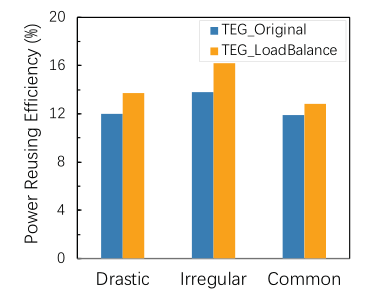


图15 TEG/CPU在三种CPU利用率下的功率复用效率

经过公式计算等，作者得出TEG/CPU在三种CPU利用率下的功率复用效率。如图15所示。对于剧烈、不规则和常见工作负载，原始TEG的电力再利用效率分别为12.0%、13.8%和11.9%。经过工作负载平衡优化后，TEG负载平衡的电力再利用效率分别为13.7%、16.2%和12.8%，平均电力再利用效率为14.23%。

# 总结

数据中心能耗上升的局面严重阻碍了可持续绿色数据中心的发展。本文从冷却系统优化角度讨论了数据中心技术的能耗优化方法，主要分三个层次进行了调研，一是调研了虚拟熔化温度这一热感知的工作安排技术，实现对数据中心热输出的更大控制，从而降低峰值冷却负载，达能节能的目的。二是针对温水冷却服务器架构，其能效会受到服务器之间热不平衡的影响，调研了一种混合冷却架构设计，它将热电冷却器结合到水冷却系统中，以细粒度的方式处理冷却不匹配。三是从混合水冷系统的服务器组件吸收的热量被直接喷射到水中，而没有被回收，导致能量浪费这一问题出发，调研了热发电这一经济且能源循环利用的温水冷却架构，热电发电机从“用过的”温水中获取热能，并发电供数据中心重复使用。以达到节能目的。

# 参考文献

[1] M. Skach, M. Arora, D. Tullsen, L. Tang and J. Mars. Virtual Melting Temperature: Managing Server Load to Minimize Cooling Overhead with Phase Change Materials. // Proceedings of the 45th Annual International Symposium on Computer Architecture (ISCA). Los Angeles,USA. 2018:15-28.

[2]Weixiang Jiang, Ziyang Jia, Sirui Feng, Fangming Liu, and Hai Jin. Fine-grained warm water cooling for improving datacenter economy. // In Proceedings of the 46th International Symposium on Computer Architecture (ISCA). New York, USA. 2019:474–486.

[3] X. Zhu et al. Heat to Power: Thermal Energy Harvesting and Recycling for Warm Water-Cooled Datacenters. // In Proceedings of the 47th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain. 2020:405-418.