计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2024/18(10)-2690-14
doi: 10.3778/j.issn.1673-9418.2310065

基于混合存储结构的分级协同节能方案

李大平[1],史庆宇[2,3+],唐忆滨[1],胡哲琨[1],高　毅[1]
1. 武汉数字工程研究所,武汉 430000
2. 湖南工商大学 计算机学院,长沙 410205
3. 湘江实验室,长沙 410083
+ 通信作者 E-mail: qingyushi@hutb.edu.cn

摘　要:新能源的应用已成为一种趋势,但具体到存储I/O领域,存储设备通常会根据存储负载的强度进行节点调度,而不规则波动的新能源和存储负载之间存在波峰波谷错位匹配的问题,增加了存储设备的调度难度。提出了一种针对NVM-SSD混合存储介质的分级协同节能方案MixSave,为每一层提供定制化节能方案并且协同调度,提升了新能源能效比。基于NVM和SSD存储介质的性能与能耗特点,MixSave采用NVM-SSD的高性价比分级存储架构,为用户提供高性能、大容量存储服务,并且通过副本保证数据可靠性。NVM作为缓存层采用负载驱动型方案,以保证性能为优先目标,即根据负载变化动态调整高功耗状态的NVM设备数量,满足负载需求;SSD作为数据层采用新能源驱动型节能方案,以提升新能源利用率为目标,根据新能源的变化动态调整SSD启用数量,进行缓存数据同步与预取。测试表明,与未采取节能措施的标准方案相比,MixSave性能下降不超过4%,在轻负载模式下,MixSave可节省73%~80%的传统能源,在重负载模式下,可节省55%~61%的传统能源。

关键词:存储系统;分级存储;非易失性存储;存储节能;新能源
文献标志码:A　　**中图分类号:**TP333

MixSave: Tiering-Cooperative Energy-Efficient Scheme for Hybrid Storage System

LI Daping[1], SHI Qingyu[2,3+], TANG Yibin[1], HU Zhekun[1], GAO Yi[1]
1. Wuhan Institute of Digital Engineering, Wuhan 430000, China
2. School of Computer Science, Hunan University of Technology and Business, Changsha 410205, China
3. Xiangjiang Laboratory, Changsha 410083, China

Abstract: Taking advantage of green energy is becoming increasingly popular. However, when it comes to the storage I/O, where storage devices usually schedule nodes based on the intensity of the storage workload, the irregular fluctuations of green energy lead to the mismatch of peaks and valleys between green energy and storage workloads, which increases the difficulty of scheduling storage resources. To address this issue, a tiering-cooperative energy-efficient scheme for hybrid NVM-SSD storage system, called MixSave, is proposed. MixSave customizes the energy-saving strategies for each layer and tunes them cooperatively between layers, reducing overall energy usage. It efficiently tiers NVM and SSD and replicates data among them based on their performance and energy characteristics,

基金项目:教育部人文社会科学项目(23YJCZH183);湖南省自然科学基金(2022JJ40129);湖南省教育厅科学研究项目(21B0572);湘江实验室开放基金(22XJ03014,23XJ01012)。
This work was supported by the Humanities and Social Science Project of Ministry of Education of China (23YJCZH183), the Natural Science Foundation of Hunan Province (2022JJ40129), the Research Foundation of Hunan Provincial Department of Education (21B0572), and the Open Project of Xiangjiang Laboratory (22XJ03014, 23XJ01012).
收稿日期:2023-10-19　　**修回日期:**2024-03-12

achieving both high performance and large capacity. The workload-driven strategy is used by the NVM layer, which is used for caching, to ensure the best performance, and the number of active NVM devices depends on the storage workload. Meanwhile, the SSD layer employs a green energy-oriented strategy to conserve traditional energy, and the number of active SSD devices will be proportionally to the amount of green energy supplied. Evaluation results show that, when compared with energy-save-unaware approaches, MixSave degrades performance by less than 4% and saves traditional energy by 73% to 80% and 55% to 61% under light and heavy workloads, respectively.

Key words: storage system; hierarchical storage; non-volatile memory; storage energy saving; green energy

信息技术的发展为各行业数字化转型提供关键技术保障，随着5G、人工智能、工业互联网等新一代信息技术的试点应用落地，信息世界飞速扩张。全球数据总量将从2018年的33 ZB增至2025年的175 ZB，而中国的数据总量将以领先全球的年平均增长速度在2025年增至48.6 ZB，成为世界上数据量最大的国家。全部数据中能被长久存储的数据占比很小，2018年国内的全部存储容量仅占总数据量的12%。

为挖掘海量数据中蕴藏的价值，存储至关重要，而现阶段数据存储的主力是分布在世界各地的数据中心。随着数据中心的发展，能耗问题越发显著。信息通信技术（information and communications technology，ICT）电量消耗占全球总用电量的比例从2022年开始将呈现指数上升态势，到2030年ICT的电量消耗将占全球总电量的20.9%，其中数据中心电量消耗占据全球总电量的7%[1]。巨大的能耗不仅会导致大量二氧化碳排放，加剧恶劣气候出现的范围和频率，还会增加数据中心的成本，降低数据中心能耗至关重要[2-3]。

数据中心的能耗中，信息技术（information technology，IT）设备能耗占比约为45%，IT设备对外提供计算和存储服务，在计算密集型服务器中，存储设备部分占据超过20%的能耗；在数据密集型服务器中，存储设备部分占据超过60%的能耗；一些针对低成本大规模的数据密集型服务器中，存储设备的外存部分能耗占比能达到20%左右[4]。整体来看，存储设备当前的能耗比例不容忽视，数据总量的爆炸式增长使得存储节能势在必行。

追求极致性能的存储系统通常采用动态随机存取存储器（dynamic random access memory，DRAM）构建大型缓存系统。但是DRAM存在容量有限、数据易失和功耗高（尤其是刷新功耗[5]）等问题，应用规模受限。存储硬件技术的发展带来新的转机，新型非易失性内存（non-volatile memory，NVM）既能够提供接近DRAM的访问性能，又能提供大容量的持久化存储空间，对解决高性能存储问题意义重大。

组建全NVM的存储系统追求极致存储性能的成本过高。而且数据被访问的概率往往会随着时间的流逝而衰减，被访问概率小的数据没有必要继续存储在NVM中。因此，可以构建一个以NVM为缓存层的分级存储架构，目的是通过底层设备存储被访问概率较小的数据，增大整体存储容量，分摊存储成本。常见的存储介质中，性能仅次于NVM且容量较大的介质是固态硬盘（solid state drive，SSD），因此计划采用上层NVM、下层SSD的分级结构构建了高性价比的存储系统。本文提到的NVM特指比SSD性能更好，更接近DRAM的非易失性内存。

为缓解数据中心的能耗压力，除从存储设备架构优化层面考虑外，还可以加大对新能源的使用。学术领域涌现出很多针对新能源应用的研究[6-7]，对光伏发电和风电的大规模部署也展开了论述[8-9]。苹果、脸书和谷歌等公司率先做出要实现数据中心百分之百新能源供能的承诺，并且已经在各自的数据中心中展开新能源的部署。

新能源应用中也存在着很多挑战，新能源的波动性和间歇性供能问题是其在数据中心全面部署的一大阻碍，当遇到同样具有波动性和不确定性的存储负载时，这一问题也必将更加突出[10]。由于新能源本身的间歇供能和频繁波动问题，新能源和存储负载的波峰波谷不能很好地匹配。在负载最为繁重时，本来配套建设的新能源不能保证处在波峰位置，往往难以依靠新能源供应开启足够的存储设备来保证服务的性能；而在负载较轻时，新能源若不在波谷，则会存在明显的新能源浪费问题。目前在存储系统中的新能源应用方案主要可以分为负载驱动型

和新能源驱动型两类,但是两类方案出发点都比较单一,无法做到负载和新能源的有机融合。因此需要基于实际存储结构对现有的两类节能调度策略取长补短,寻找更加高效的新能源应用方案。

为解决新能源和负载难以匹配的问题,基于NVM和SSD的分层存储架构提出一种分级协同全局节能存储系统MixSave,NVM缓存层和SSD数据层分别采用负载驱动型和新能源驱动型节能策略,并且根据副本架构调整设备启用顺序。最后的测试表明,与基准方案standard相比,MixSave方案基本没有性能损失,并且在轻负载模式下,能够节省约73%~80%的传统能源,在重负载模式下,能够节省55%~61%传统能源,并且有着高达96%的新能源利用率。

本文的主要贡献如下:

(1)新能源应用时的节能方案主要可以分为负载驱动型和新能源驱动型两类,两类方案分别依据负载和新能源来控制启用设备的数量进行节能调度,各有优缺点。基于分级存储结构可以把两者的优点结合起来:负载驱动型方案能保证存储性能,因此在性能更好的NVM缓存层实施,根据负载的轻重来调整需要保持在高功耗状态的NVM数量;新能源驱动型方案能保证新能源利用率,因此可以在规模较大、能效较低的SSD数据层实施,根据新能源的数值来确定可以启用的SSD设备数量。两层存储结构协同互补,能够充分发挥两种节能调度方案的优点。

(2)在因特尔服务器上实现了MixSave以及对比方案,并将所有方案进行了详细的对比和分析,从轻重负载下的性能和能耗等方面验证了MixSave的作用。

1 背景及相关工作

面对数据中心日益严重的能耗压力,使用更加环保且便宜的新能源意义重大。罗格斯大学和惠普实验室已经建立由新能源承担一部分供能任务的数据中心,用来验证自己的策略[11-12],也有一些研究人员通过仿真的方式来对节能减排进行研究[13]。

对于新能源在存储领域的运用,学术界的研究总体上可以分为三类:负载导向性策略、新能源驱动型策略以及电池类策略。三类方案都是通过控制开启设备的数量来调控能耗,减少对传统能源的使用,最大化新能源的利用率,并且尽量保证稳定的性能,但是调控的依据并不相同。

1.1 负载驱动型节能

负载驱动型策略更倾向于根据负载强度来决定开启设备的数量,负载越大,则开启设备越多,若新能源不足,则会采用传统能源供能以保证所需求的性能。这类方案能够获得较好的性能,适用于需要实时交互的应用场景,但是如果只采用这一基础的思想,不仅会在负载较重且新能源供应量低时消耗掉大量的传统能源,还会在负载较轻时造成新能源的浪费。针对这一问题,多种方案从不同的角度对负载驱动型策略做出了优化调整。

代表性方案GreenSlot[14]、GreenHadoop[15]以及GreenSwitch[16]的主要思想都是实时处理需要在线响应的任务,保证服务性能,不过出于降低传统能源消耗并提高对新能源利用率的目的,会延迟处理不需要在线响应的任务,把这部分请求推迟到新能源供应量高的时候再集中处理,以此来降低对传统能源的消耗并提高对新能源的利用率。其中,GreenSlot和GreenHadoop通过预测未来时刻的太阳能信息来对计算量进行延迟调整以迎合太阳能的变化,如果必须使用传统能源,则会选择价格相对便宜的时间段来降低开销。GreenSwitch主要是通过控制负载执行时间和对能源的选择来达到节能的目的。

GreenCassandra[17]通过一种简单的分布式多副本架构对新能源进行使用,不管新能源供应量的多少,一直保持主副本的开启,从副本的开关可以去追逐新能源的变化。也有方案[18]通过动态负载预测来减轻服务器交换机的负面影响,并根据每日负载波动情况在服务器内以细粒度的方式进行资源配置。

1.2 新能源驱动型节能

新能源驱动型策略通常会根据新能源的供应量来决定开启设备的数量,新能源供应量越大,则可以开启的设备越多。这类方案目标是得到较高的新能源利用率,但是并不保证性能,当新能源供应量低而负载重时,会出现明显的性能下降。

为缓解这一问题,SolarCore[19]在所使用的太阳能供应量较低时,通过暂时性地关闭设备来降低服务器的能耗。Blink[20]采用交错调度的方式来缓解风能和太阳能的变化性不可控特点,增进新能源的利用效率。当新能源充足可用时,GreenPar[21]增加活动作业的资源分配,保证服务性能,当新能源不足需要使用传统能源时,GreenPar在有限约束内减少资源分配,以节省能源。

1.3 电池方案

HEB(hybrid energy buffering)[22]方案考虑采用电池来存储富余的新能源,然后在新能源不足时释放,

这样可以让新能源供应变得相对平稳可控。HHEB（hybrid and hierarchical energy buffering）[23]方案提出了一种混合分级电源管理方案，能够解决数据中心多种电池供应功率与用户负载需求功率不匹配的问题。也有方案[24]开发了一种具有智能工作负载迁移的新型电池分配框架。该框架同时保护交互式工作负载不被中断，并最大限度地减少批量工作负载的等待时间。

但是电池方案缺点也较为明显：(1)电池循环充电次数有限，大概在2 000到3 000次，频繁充放电会进一步减少电池的使用寿命；(2)电池放电功率有限，因为大功率放电会导致电池可使用容量减少；(3)为防止充电时电池过热，无法通过大电流给电池快速充电；(4)电能通过电池循环一次后会造成15%~20%的能量损失；(5)电池储能设备前期投入巨大，还需要定期维护，会加大数据中心的建造成本。

1.4 综合考虑方案

GreenGear[25]根据服务器的性能和能耗表现，把服务器划分为强弱两种，根据负载需求和新能源供应量来动态调整强弱服务器的工作状态，能达到不错的节能效果。GreenHetero[26]针对异构服务器，做了更进一步的细粒度划分方案和调度策略。但是这类方案对电池有不小的依赖，而大规模使用电池不仅需要较大的前期投入开销，而且还会因为能源转换的损失影响运行效率。

综上所述，大部分方案都是着重地考虑负载或者新能源中一方的影响。负载驱动型节能方案更为关注性能，但是忽略了对新能源的高效利用，存在着巨大的浪费；新能源驱动型节能方案更为关注新能源利用率，但是会忽略掉负载的波动，造成性能下降。

2 问题分析

2.1 新能源和负载的不匹配问题

图1和图2分别展示了夏季和冬季的太阳能、风能和混合新能源的波动情况，混合新能源由太阳能和风能组成。新能源数据来自于美国国家可再生能源实验室，选取的是美国道奇市2月份和6月份一周的数据。由于太阳能在夜间断供，风能也不够稳定，采用太阳能和风能构成的混合新能源可以减少新能源供应时的空窗期。从图1和图2中可以看出，单独的太阳能和风能都存在明显波动，两者组成的混合新能源也不例外，波峰、波谷明显。

新能源波动明显，而存储负载也充满变化，因此

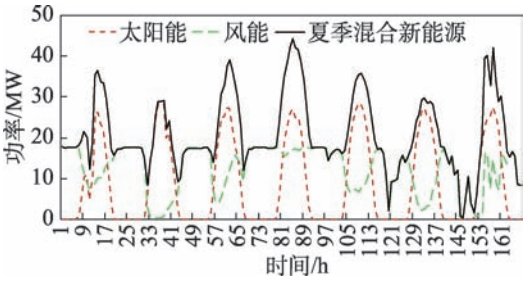

图1 夏季的太阳能、风能和混合新能源的波动情况

Fig.1 Fluctuation of solar, wind and hybrid new energy sources in summer

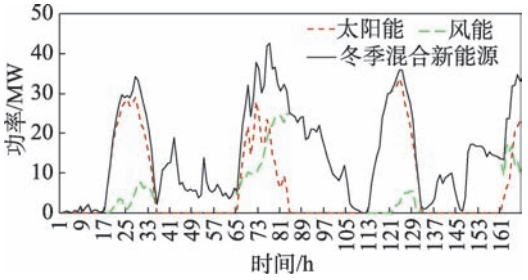

图2 冬季的太阳能、风能和混合新能源的波动情况

Fig.2 Fluctuation of solar, wind, and hybrid new energy sources in winter

负载和新能源的波峰和波谷难以重合匹配。在负载较为繁重时，若新能源供应量不足，则难以开启足够的设备来保证服务质量；在负载较轻时，若新能源供应过剩，则会造成明显的资源浪费。

若不借用其他硬件设施，目前数据中心的新能源应用节能方案主要可以分为负载驱动型和新能源驱动型两类。两类方案分别依靠负载和新能源信息来控制开启设备数量，出发点都比较单一，无法做到负载和新能源的有机融合。

若采用电池存储富余新能源，可以让新能源供应变得相对平稳可控。但是电池循环充电次数有限，通常有2 000多次，频繁充放电时使用寿命会更短[23]；放电功率受限，因为大功率放电会导致电池可使用容量减少；为防止充电时电池过热，无法通过大电流给电池快速充电；电能通过电池循环一次后会造成15%~20%的能量损失；电池储能设备前期投入巨大，还需要定期维护，会加大数据中心的建造成本。

因此需要寻找更加高效的新能源应用方案，综合考虑负载和新能源的波动情况，解决负载和新能源的不匹配问题。

2.2 不同存储设备的能效对比分析

组建全NVM的存储系统追求极致存储性能的

成本过高,而且数据被访问的概率往往会随着时间的流逝而衰减,被访问概率小的数据没有必要继续存储在NVM中。因此,可以构建一个以NVM为缓存层的分级存储架构,目的是通过底层设备存储被访问概率较小的数据,增大整体存储容量,分摊存储成本。

常见的存储介质中,性能仅次于NVM且容量较大的介质是SSD,常用的SSD有SATA SSD和PCIe SSD。因此,可从设备的能效和价格方面对NVM和两种SSD进行对比分析,确认是否适合搭建NVM-SSD的分级存储架构。

图3展示了三种存储设备在不同读写类型下的能效情况。能效是指存储设备在1 J能耗下能处理的请求数量,该数值越大则能效越高。其中NVM设备是128 GB的Intel商用持久化缓存存储设备,PCIe SSD是1 TB的三星970 EVO Plus NVMe™ M.2设备,SATA SSD是960 GB的希捷Nytro 1551 SATA固态硬盘,测试所用的请求大小为4 KB。

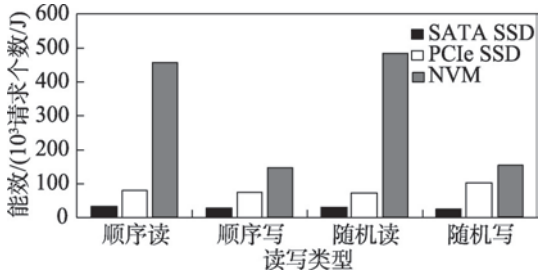

图3 三种设备在不同读写类型下的能效对比

Fig.3 Comparison of energy efficiency of three devices under different read and write types

从图3中可以看出NVM能效最高,NVM读能效分别是PCIe SSD和SATA SSD的3.8倍和9倍,NVM写能效分别是PCIe SSD和SATA SSD的1.2倍和3.9倍。

表1展示了三种设备的单位容量价格,可以看出NVM单位容量价格最高,是两种SSD数值20倍左右。

表1 三种设备的单位容量价格

Table 1 Unit capacity price of three types of equipment

设备类型	单位容量价格/(元/GB)
PCIe SSD	1.278
SATA SSD	0.998
NVM	22.650

综上来看,适合搭建NVM-SSD的分级存储架构。

2.3 问题分析小结和设计思路

(1)构建以NVM为缓存层的分级存储架构

NVM能效最高,同时单位容量价格也是SSD的20倍左右,因此适合作为缓存层,对外提供高性能服务;存储能效和单位容量价格都很低的SSD适合作为底层存储设备,在后台提供大量存储空间。因此最终决定采用NVM-SSD的混合架构,既能对外提供高性能服务,又能够保证较大的存储容量。

(2)融合负载驱动型和新能源驱动型方案

在存储领域,新能源应用时的节能方案主要可以分为负载驱动型和新能源驱动型两类,两类方案分别依据负载和新能源来控制启用设备的数量进行节能调度,各有优缺点。基于分级存储结构可以把两者的优点结合起来:负载驱动型方案能保证存储性能,因此在性能更好的NVM缓存层实施,根据负载的轻重来调整需要保持在高功耗状态的NVM数量;新能源驱动型方案能保证新能源利用率,因此可以在规模较大、能效较低的SSD数据层实施,根据新能源的数值来确定可以启用的SSD设备数量。两层存储结构协同互补,能够充分发挥两种节能调度方案的优点。

3 MixSave设计

针对新能源和负载的难以匹配问题,提出一种基于混合存储架构的分级协同节能方案MixSave。如图4所示,整个MixSave的系统架构主要由用户负载、供能模块和存储架构组成。

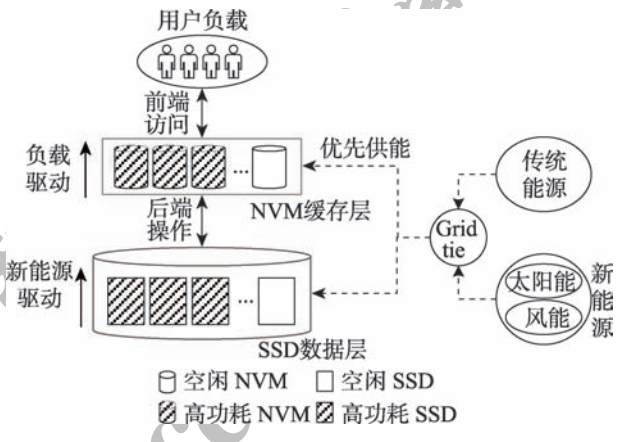

图4 MixSave系统架构

Fig.4 System architecture of MixSave

用户负载部分代表整个系统接受的负载信息,出于简化考虑,主要分为轻重两种负载。轻负载指

带有请求间隔的稀疏数据流，重负载指没有请求间隔的密集数据流。

供能模块由新能源、传统能源和Grid tie组成。新能源采用太阳能和风能结合的混合能源，目的是减小新能源的波动和无法供能的时间比例。采用Grid tie技术[27]对新能源和传统能源进行同步管理，保证多种能源的平滑衔接。MixSave会优先对NVM缓存层进行供能，新能源不足时，使用传统能源补充，保证正常服务。

存储架构部分是NVM-SSD分级存储架构。NVM缓存层为前端访问提供性能保障，通过推迟写数据进入SSD数据层的方式处理所有写请求，通过预取操作处理大部分的读请求；SSD数据层主要处理由数据同步倒盘和数据预取构成的后端操作，执行时间较为灵活，在提供充足存储空间的同时还能够降低硬件成本。

根据NVM缓存层和SSD数据层的不同作用，采用负载驱动和新能源驱动相结合的分级协同节能方案。NVM缓存层采用负载驱动型节能方案，根据负载轻重调整需要保持在高功耗的NVM数量；而SSD数据层采用新能源驱动型节能方案，根据新能源供应量的多少来调整可以启用的设备数量。设定所有设备启用前都处于低功耗空闲状态，启用后都处于高功耗状态。

3.1 分级协同节能思想概述

分级协同节能方案的主要目的是融合负载驱动型和新能源驱动型节能方案的优点，以在最大化新能源利用率、最小化传统能源消耗量的同时保证充足的存储性能。本节对分级协同全局节能思想进行简要概述。

(1)NVM缓存层调度-负载驱动型为主

图5为NVM缓存层调度的简要示意图。NVM缓存层采用负载驱动型节能策略，用户请求越多，则高功耗状态的NVM越多。用户请求整体较少，即轻负载时，小部分高功耗NVM能够处理掉大部分的请求；剩余的NVM因为需要处理的请求较少，大部分时间处于低功耗空闲状态，因此节能效果明显。用户请求整体较多，也就是重负载时，少部分高功耗NVM已经无法保证整体性能，因此需要大部分NVM参与进来；但是还会有小部分NVM因为需要处理的请求较少，可以大部分时间处于低功耗空闲状态，不过整体节能效果较弱。整个过程中，指向空闲状态NVM的访问大部分都会被重定向到部分高功耗NVM上。

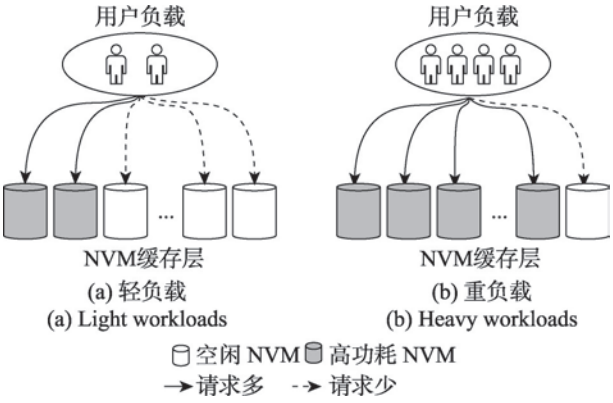

图5 NVM缓存层调度的简要示意图

Fig.5 Brief schematic diagram of NVM cache layer scheduling

(2)SSD数据层调度-新能源驱动型为主

图6展示了SSD数据层调度的简要示意图，灰底方框表示高功耗SSD，由新能源启动；空白方框表示空闲SSD，没有访问，处于低功耗空闲状态。整个SSD数据层采用新能源驱动型节能策略，因此新能源越多，高功耗SSD越多，NVM缓存层只会与高功耗SSD之间进行后端操作。

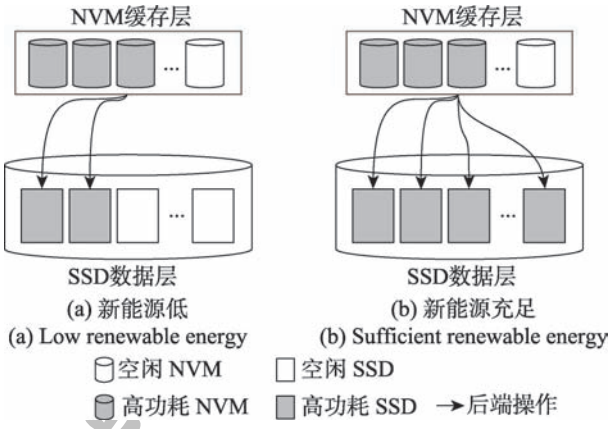

图6 SSD数据层调度的简要示意图

Fig.6 Brief schematic diagram of SSD data layer scheduling

需要注意的是能源会被优先供应给NVM缓存层，用来保证性能，在满足NVM缓存层之后如果新能源还有剩余，才会用来供应SSD数据层，直到新能源用尽或者启用全部设备为止。可以看出新能源越多，则启用的SSD设备越多。若NVM空间已满，需要向SSD强制同步数据，则会优先使用新能源供能，新能源不足则用传统能源替代。

3.2 分级存储架构

如图7所示,分级存储架构部分由客户端、NVM缓存层和SSD数据层组成。

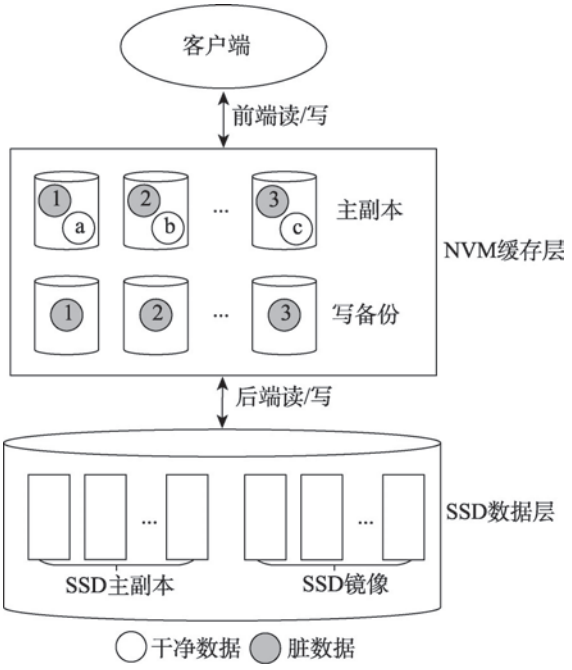

图7 分级存储架构

Fig.7 Hierarchical storage architecture

NVM缓存层全部由Intel公司的新型持久化内存存储器件Optane DC PM构成。为保证缓存层数据的可靠性,设置了双副本结构,分为主副本和写备份。提供读写服务的是NVM主副本,所有的写请求数据都会在NVM写备份盘上存在一个备份,以保证数据的可靠性。因为要保证重负载时的备份性能,所以采用同等数量的写备份NVM。

SSD数据层旨在为上层提供大容量且低成本的数据存储空间。为保证数据的可靠性,SSD数据层也采用双副本结构,其中主副本负责处理所有的读取操作,SSD镜像为SSD主副本提供备份。

3.3 NVM缓存层数据流

以8个NVM的缓存层为例进行说明,图8展示了NVM0和NVM0′处于高功耗状态时的读写数据流情况。NVM0和NVM0′被称为主NVM,负责处理到来的请求,其他NVM大部分时间处于空闲状态。主NVM的数量会根据负载轻重情况进行调整。NVM写备份中NVM(i)′会为对应的NVM(i)上的脏数据提供一个备份,保证写入数据的可靠性。

NVM0和NVM0′被分为写缓冲区和数据区两个部分。写缓冲区用来缓存被分配到其他空闲NVM

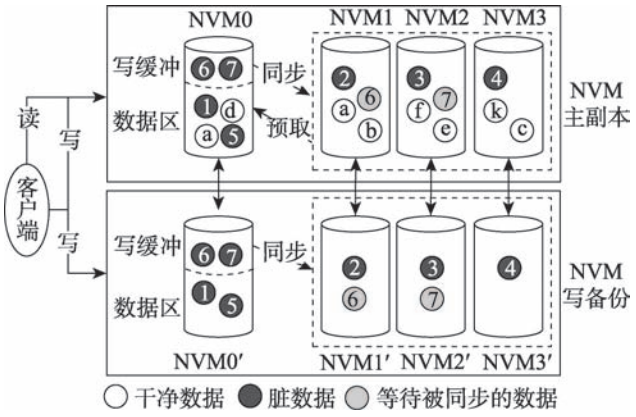

图8 NVM高功耗状态时读写数据流示例

Fig.8 Example of read/write data flows with NVM in high power state

的写请求数据,NVM0对应管理NVM1、NVM2和NVM3这3个空闲NVM;NVM0′对应管理NVM1′、NVM2′和NVM3′这3个空闲NVM。数据区用来存储分散到各个NVM上的读写数据。NVM0和NVM0′中的脏数据6和7会通过后台同步操作转移到NVM1、NVM2、NVM1′、NVM2′上,目前状态表示同步操作还未执行。

3.4 数据组织结构

为有效地管理缓存层和SSD存储层的数据,采用了如图9所示的索引结构。图9是以4个NVM主副本设备和4个NVM写备份设备为例进行说明,其中NVM0和NVM0′作为主NVM,处于高功耗状态,其他NVM大部分时间会处于空闲状态。

开始为一个简单的哈希算法,用来判断到来的请求被分配到哪个NVM。如果对应的NVM是空闲状态,则会把请求发送给对应的主NVM。如果是读请求,则直接交给NVM主副本进行处理,如果是写请求,则会同时发送一份给NVM写备份进行备份。

接下来是NVM主副本和NVM写备份,两者中都会保证至少启用一个NVM,例如此处NVM0与NVM0′处于启用后的高功耗状态。每个NVM设备中都会有一个B+树来管理自身存储的数据,这个区域叫作数据区。

NVM0和NVM0′中的写缓冲区都是用来缓存被分配到其他空闲NVM的写请求数据,这些数据会被以日志形式暂时缓存起来,并且每个空闲NVM的数据日志都会采用一个B+树进行管理。NVM0对应管理NVM1、NVM2和NVM3这3个空闲NVM,NVM0′对应管理NVM1′、NVM2′和NVM3′这3个空闲NVM,因此NVM0和NVM0'对应管理的NVM中有几个处

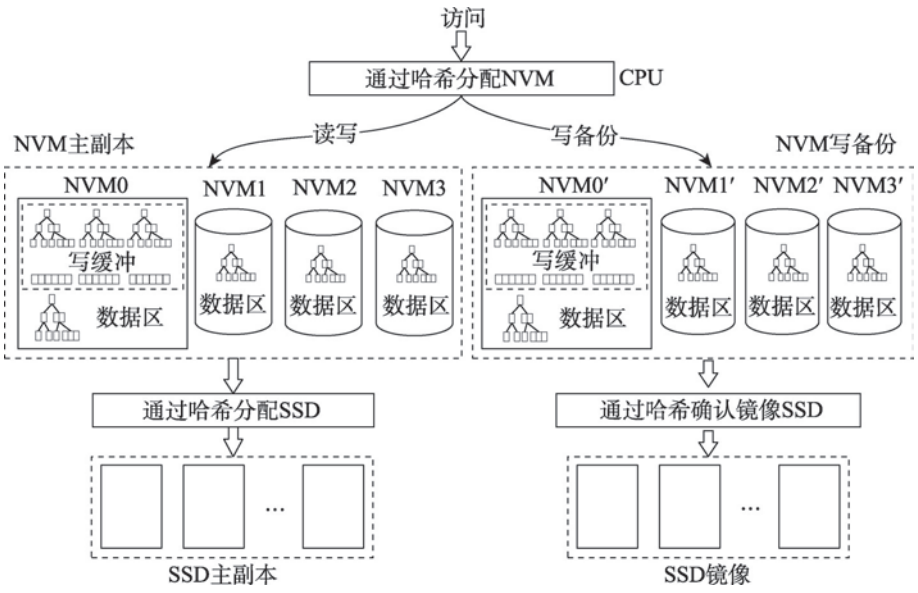

图9 分级存储系统的索引结构

Fig.9 Index structure of hierarchical storage system

于空闲状态，其写缓冲区内就会有几个数据日志和B+树。

最后是SSD数据层，包含了SSD主副本和SSD镜像。为保证数据的访问效率、减少倒盘等因素对性能的影响，限定NVM主副本的数据向SSD主副本进行持久化，NVM写备份的数据向SSD镜像进行持久化，并且主要由SSD主副本提供读服务。除此之外，这两对副本分别通过设置哈希表的方式来确认数据的对应关系。

当有请求到来时，会先使用哈希算法进行计算，确认该请求将会被分配到哪个NVM，如果对应的NVM没有被启用，那么会先把该请求重定向到第一对NVM主从副本上进行处理，如果缓存层不能处理该请求，会接着把该请求进行新一轮的哈希计算后发送到对应的SSD上进行处理。

4 实验与分析

4.1 测试环境

测试平台：使用包含20个SSD（希捷Nytro 1551 SATA固态硬盘960 GB）和8个NVM（Intel® Optane™ DC Persistent Memory 128 GB）的真实服务器进行测试。服务器的CPU为Intel® Xeon® CPU E5506@2.13 GHz，Linux版本为2.6.35.6-45.fc14.x86 64。

为评估MixSave方案在性能和能耗上的表现，一共设置了4种对比方案：Standard（无节能方案）、负载驱动型方案（workload-driven scheme，WDS）、新能源驱动型方案（green-driven scheme，GDS）和MixSave。

Standard：不使用任何的节能方案，保持设备全开，新能源不足时采用传统能源供电。

WDS：负载驱动型的节能调度方案，为保证性能，轻重两种负载时WDS会一直保持NVM缓存层和SSD主副本和部分SSD从副本处于被启用后的高功耗状态。

GDS：新能源驱动型的节能调度方案，根据新能源供应量来依次启用缓存层、SSD主副本和SSD镜像。轻负载时保证启用2个NVM，重负载时保证启用2个NVM和4个SSD，之后依据新能源调度。

MixSave：本文方案的实现。轻负载时保证启用2个NVM，重负载时保证启用6个NVM和4个SSD。

测试工具：测试时采用C++版本的YCSB（Yahoo! cloud serving benchmark）[28]来进行负载仿真。如果没有另外说明，后续测试中都是4 KB的请求，数据分布方式为zipfian，使用20个线程来分发数据，每种负载都是20 GB的总数据量。（1）YCSB A，50%读操作、50%写操作、0 μs时间间隔；（2）YCSB B，95%读操作、5%写操作、0 μs时间间隔；（3）新生成的YCSB 3，33%读操作、67%写操作、0 μs时间间隔；（4）在前边3种负载的请求之间加上一个约50 μs的间隔来模拟分散的轻负载，分别命名为YCSB A-T、YCSB B-T、YCSB 3-T。

测试中使用RAPL（running average power level）[29]来间接监测NVM的实时功耗，RAPL能够获取Intel

主板上关键区域的功耗信息，每个处理器对应的DIMM（dual-inline-memory-modules）区域功耗会被单独显示。NVM各个状态的能耗数据事先通过单独测试得到，SSD功耗采用的是官方给定数据。

新能源信息：新能源供应量数据是从美国国家可再生能源实验室[30]获取的，分别选取了美国道奇市夏天和冬天同一周的太阳能和风能数据，包含了太阳能和风能每10 min的平均功率信息。实验测试时，为了减少新能源波动对测试的影响，并且保证合适的供应数值，最后采用了太阳能和风能结合的混合新能源，具体有冬天（winter）和夏天（summer）两种。两个混合新能源是根据8个NVM和20个SSD的存储规模对相加后的风能和太阳能数值扩大3倍后所得。

4.2 轻负载测试对比

（1）轻负载下的能耗分析

图10展示了不同轻负载和新能源组合下多种方案的总能耗。柱状图的下半段表示新能源使用量，上半段表示传统能源使用量。

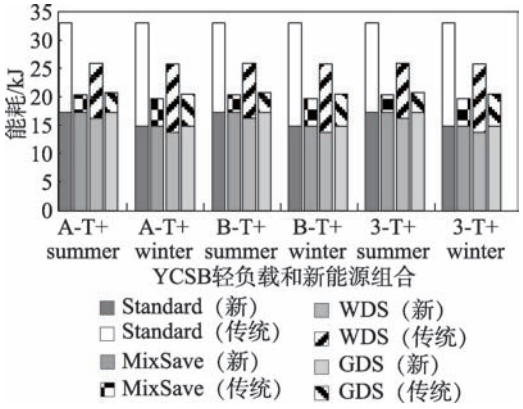

图10 不同轻负载和新能源组合下多种方案的总能耗

Fig.10 Total energy consumption of various solutions under different light workloads and new energy combinations

4种方案的能耗情况在各个轻负载下基本一致，因为在处理轻负载时存储系统的大部分时间都被浪费在了50 μs的时间间隔上，运行时长和功耗基本一致，掩盖了各个轻负载的特性。

从图10中可以看出，与Standard方案相比，其他3种方案都能够大幅度降低对传统能源的消耗。与WDS方案相比，MixSave和GDS方案节能效果更好，同时WDS方案的总能耗也比MixSave方案和GDS方案更高。因为负载轻时，WDS方案也需要启用缓存层、SSD主盘和少部分SSD镜像，而后两种方案是依据新能源供应量来进行启用设备，高功耗设备相对较少。

图11展示了轻负载A-T下多种方案对传统能源的节能比例，其中MixSave方案节能效果最好，GDS次之，WDS最差。在新能源summer和winter下，WDS方案分别能节省37%和33%的传统能源，MixSave方案分别能节省80%和73%的传统能源，GDS方案分别能节省78%和68.8%的传统能源。3种方案都在新能源summer时具有更高的节能比例，因为summer比winter平均功率更高，且更加平稳。MixSave方案比GDS更好一些，因为GDS方案启用设备的顺序没有进行优化，盲目启用NVM设备并不能达到最优节能效果，浪费了一些能源。

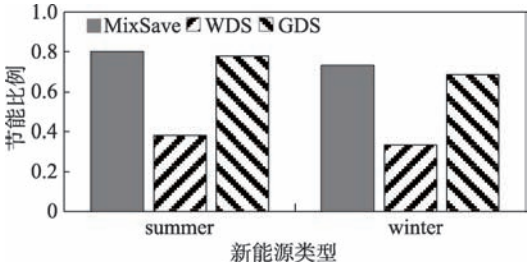

图11 轻负载A-T下各方案对传统能源的节能比例

Fig.11 Energy-saving ratio of each scheme to traditional energy under light workload A-T

（2）轻负载下的性能分析

图12和图13分别展示了轻负载下summer新能源供能时多种方案的平均响应时间（归一化后结果）和IOPS（input/output operations per second）。可以看出WDS方案和Standard方案有着基本相同的平均响应时间和IOPS。因为与Standard方案相比，WDS方案只是没有启用全部SSD镜像设备，并不会对整体性能造成明显影响。GDS与MixSave方案都会至少

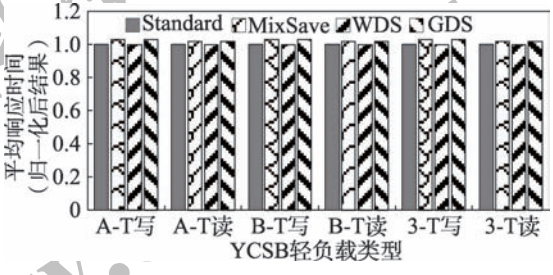

图12 轻负载下summer供能时各方案的平均响应时间

Fig.12 Average response time of each scheme with summer new energy used under light workloads

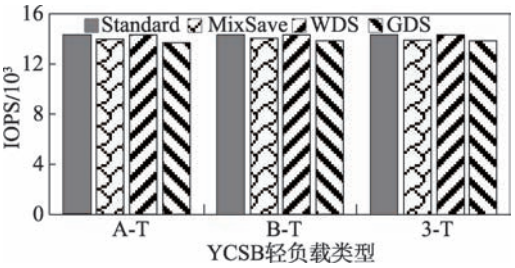

图13 轻负载下summer供能时各方案的IOPS

Fig.13 IOPS of each scheme with summer new energy used under light workloads

维持两个高功耗状态的NVM，但是因为GDS的设备启用顺序没有优化，所以GDS性能会比MixSave稍差。

在多种轻负载中，Standard方案都比MixSave方案表现好，但是差异只有2%~3%。因为轻负载的带宽与缓存层中单个NVM的最大带宽相差甚远，MixSave方案能够从容地通过数据预取把需要的热数据提前转移到缓存层中的高功耗NVM中。新能源供应量不充足时通过一对高功耗NVM处理掉所有的写请求和大部分的读请求，少量对底层SSD的读访问并不会带来明显的性能波动。

新能源的种类不会影响Standard和WDS方案，但是会对MixSave和GDS方案造成影响。图14和图15展示了轻负载时MixSave和GDS方案在两种新能源下的平均响应时间（归一化后结果）和IOPS。可以看出，在轻负载测试中，新能源对MixSave方案和GDS方案的影响基本可以忽略。但是，两种方案在summer时能获取更好的性能。因为相比winter，summer平均供应功率更大，能够启用更多的高功耗状态设备来提升性能。

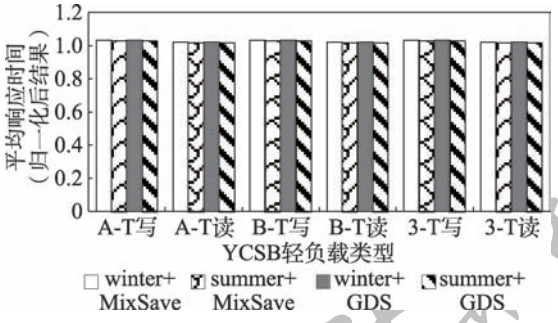

图14 轻负载时两种新能源下MixSave和GDS方案的平均响应时间

Fig.14 Average response time of MixSave and GDS schemes with two types of new energy under light workloads

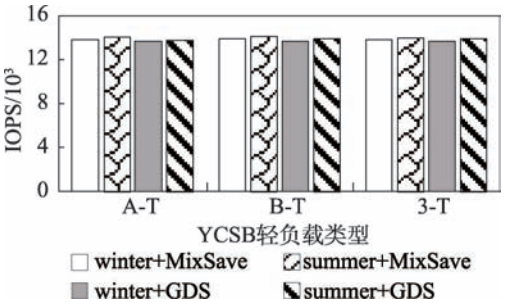

图15 轻负载时两种新能源下MixSave和GDS方案的IOPS

Fig.15 IOPS of MixSave and GDS schemes with two types of new energy under light workloads

此外，与能耗结果类似，两种方案在多种轻负载下基本一致，因为50 μs时间间隔掩盖了各个负载的特性。

4.3 重负载测试对比

（1）重负载下的能耗分析

图16展示了不同重负载和新能源组合下多种方案的总能耗。柱状图的下半段表示新能源使用量，上半段表示传统能源使用量。从图16中可以看出，Standard和WDS方案总能耗最高，且对新能源和传统能源的消耗量相同，因为要保证所有设备处于高功耗状态。与Standard和WDS方案相比，MixSave和GDS方案都能够大幅度降低对传统能源的消耗。MixSave方案可节省55%~61%的传统能源，GDS方案可节省60%和67%的传统能源。两种方案都在新能源summer时具有更高的节能比例，因为summer比winter平均功率更高，且更加平稳。MixSave比GDS节能效果稍差一些，有5%左右的能耗差距。因为重

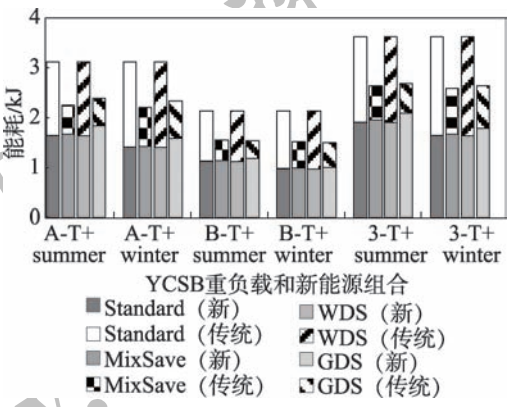

图16 不同重负载和新能源组合下多种方案的总能耗

Fig.16 Total energy consumption of various solutions under different heavy workloads and new energy combinations

负载时,为了保证性能,MixSave要启用更多的存储设备,所以能耗更大。但是在性能上,MixSave比GDS高2%~30%。因此综合考虑时,MixSave要优于GDS。

相同新能源供应下,4种方案在不同重负载下的能耗情况有明显区别。因为各个重负载的读写比例不同,而缓存层的读性能明显好于写性能,所以在保持相同数据总量的情况下,读比例最高的负载执行时间最短。YCSB A负载、YCSB B负载和YCSB 3负载的读比例分别为50%、95%和33%,因此YCSB B负载下的总能耗最低,而YCSB 3负载下的总能耗最高。

当新能源和负载相同时,在新能源使用方面,Standard和WDS方案新能源使用总量最低,因为这两种方案性能最好,处理同样的数据集执行时间最短。此外,GDS方案的新能源使用量比MixSave方案多,MixSave方案的新能源使用量比Standard方案多。因为GDS方案性能最差,需要的执行时间更长,MixSave方案性能稍好,Standard方案性能最好。

对于MixSave方案来说,新能源影响的主要是SSD的启用数量。为直观展示MixSave方案中设备数量随新能源的变化情况,图17引入了重负载A下采用summer新能源供能时MixSave方案的实时高功耗NVM和SSD数量。从图17中可以看出,NVM数量变化很小,重负载时会至少保持6个高功耗NVM,剩余2个根据新能源进行变动。此时节能的主力是SSD,可以看出,高功耗SSD的数量会随着新能源发生较大的波动。

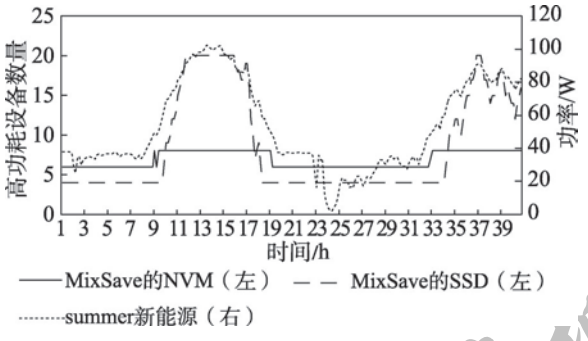

图17 重负载A下summer供能时MixSave实时高功耗NVM和SSD数量

Fig.17 Real-time high-power-consuming NVM and SSD number of MixSave scheme with summer new energy used under heavy workload A

(2)重负载下的性能分析

重负载下,通过读写平均响应时间来展示各个方案的性能。图18和图19分别展示了重负载下多种方案的读写平均响应时间(归一化结果)和IOPS,采用新能源summer供能。在多种重负载测试中,Standard方案和WDS方案的读写性能比较稳定,两者性能基本一致并且处于第一梯队,因为两种方案都会保证启用NVM缓存层和SSD主副本。

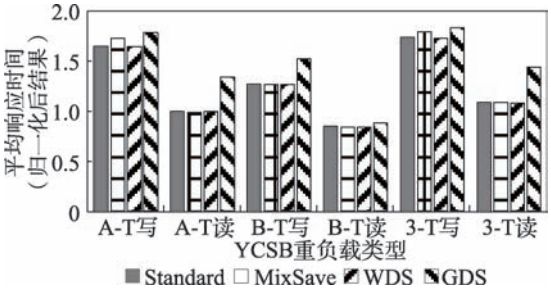

图18 重负载下summer供能时各方案的平均响应时间

Fig.18 Average response time of each scheme with summer new energy used under heavy workloads

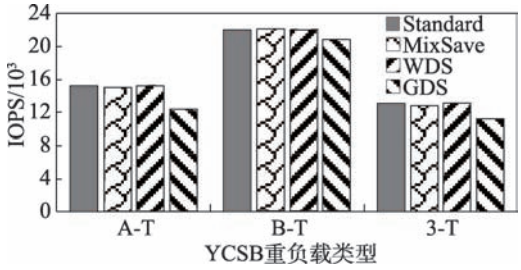

图19 重负载下summer供能时各方案的IOPS

Fig.19 IOPS of each scheme with summer new energy used under heavy workloads

相比Standard方案,GDS方案在读写性能上都有5%~30%的性能下降,且每种负载下读写总有一种出现大幅度下降;MixSave方案在写性能上下降约4%,在读性能上基本保持一致。

在YCSB B负载下MixSave方案写性能下降程度最小,因为YCSB B负载的写请求比例只有5%,影响较小。但此时,GDS方案写性能下降20%,因为GDS方案只有在新能源供应充足时才会启用更多的存储设备,才能取得优异的性能,而占比较小的写请求遇上新能源波谷时会出现明显的性能下降。

相比Standard方案,虽然少开一对NVM时读性能会有所下降,但是MixSave方案的预取操作能提高读性能,因此MixSave方案读性能不仅没有出现明显下降,甚至在读请求比例较高的YCSB A负载下读性

能最高。

新能源的种类不会影响Standard和WDS方案，但是会对MixSave和GDS方案造成影响。图20和图21分别展示了重负载时MixSave和GDS方案在两种新能源下的平均响应时间（归一化结果）和IOPS。可以看出，在重负载测试中，新能源对MixSave方案的影响很小，但是对GDS的影响相对稍大。整体上，在采用平均功率更高的summer新能源时性能略好。

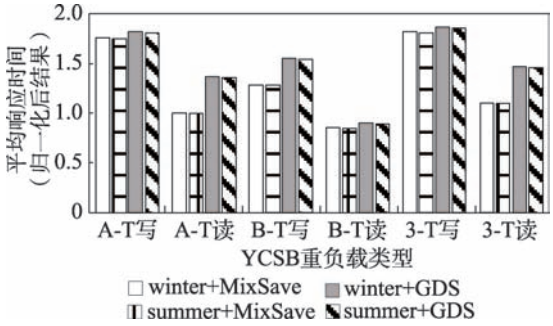

图20 重负载时不同新能源下MixSave和GDS方案的读写平均响应时间

Fig.20 Average response time of read/write of MixSave and GDS with different types of new energy used under heavy workloads

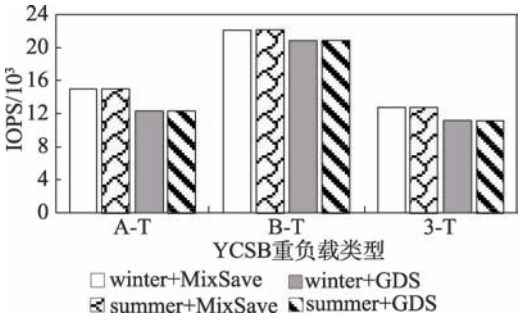

图21 重负载时不同新能源下MixSave和GDS方案的IOPS

Fig.21 IOPS of MixSave and GDS with different types of new energy used under heavy workloads

5 结论

在高性能应用场景中，采用全NVM存储系统确实能够取得优异的性能，但是成本太高，因此本文采用上层NVM、下层SSD的分级结构构建了高性价比的存储系统。为进一步降低功耗成本，引入了新能源，但是新能源的波动性和间歇性供能问题也对其在数据中心的应用带来了很大的困扰。为解决新能源和负载难以匹配的问题，提出一种分级协同全局节能存储系统MixSave，NVM缓存层和SSD数据层分别采用负载驱动型和新能源驱动型节能策略，并且根据副本架构调整设备启用顺序。最后的测试表明，与Standard方案相比，MixSave方案基本没有性能损失，并且在轻负载模式下，能够节省约73%~80%的传统能源，在重负载模式下，能够节省55%~61%传统能源，并且有着高达96%的新能源利用率。

参考文献：

[1] JONES N. How to stop data centres from gobbling up the world's electricity[J]. Nature, 2018, 561(7722): 163-166.

[2] AUER H, CRESPO DEL GRANADO P, OEI P, et al. Erratum to: development and modelling of different decarbonization scenarios of the European energy system until 2050 as a contribution to achieving the ambitious 1.5℃ climate target- establishment of open source/data modelling in the European H2020 project openENTRANCE[J]. Elektrotechnik und Informationstechnik, 2021, 138(3): 256.

[3] MASANET E, SHEHABI A, LEI N, et al. Recalibrating global data center energy use estimates[J]. Science, 2020, 367(6481): 984-986.

[4] VERMA A, KOLLER R, USECHE L, et al. Srcmap: energy proportional storage using dynamic consolidation[C]//Proceedings of the 8th USENIX Conference on File and Storage Technologies, San Jose, Feb 23-26, 2010: 267-280.

[5] OH B, ABEYRATNE N, AHN J, et al. Enhancing DRAM self-refresh for idle power reduction[C]//Proceedings of the 2016 International Symposium on Low Power Electronics and Design, San Francisco, Aug 8-10, 2016. New York: ACM, 2016: 254-259.

[6] JIN T D, SHI T Q, PARK T. The quest for carbon-neutral industrial operations: renewable power purchase versus distributed generation[J]. International Journal of Production Research, 2018, 56(17): 5723-5735.

[7] 蔡浩然. 面向绿色数据中心的计算和能源高效协同调度方法研究[D]. 武汉：华中科技大学, 2020.
CAI H R. Research on computing resources and power sources collaborative management approaches in green data centers[D]. Wuhan: Huazhong University of Science and Technology, 2020.

[8] KASSEM Y, CAMUR H, AATEG R A F. Exploring solar and wind energy as a power generation source for solving the electricity crisis in Libya[J]. Energies, 2020, 13(14): 3708-3719.

[9] LI C. Evaluation of the viability potential of four grid-con-

nected solar photovoltaic power stations in Jiangsu province [J]. Clean Technologies and Environmental Policy, 2021, 23: 2117-2131.

[10] QU X Y, WAN J G, WANG J, et al. GreenMatch: renewable-aware workload scheduling for massive storage systems [C]//Proceedings of the 2016 IEEE International Parallel and Distributed Processing Symposium, Chicago, May 23-27, 2016. Washington: IEEE Computer Society, 2016: 403-412.

[11] ROSTIROLLA G, GRANGE L, MINH-THUYEN T, et al. A survey of challenges and solutions for the integration of renewable energy in datacenters[J]. Renewable and Sustainable Energy Reviews, 2022, 155: 111787.

[12] LIU Z H, CHEN Y, BASH C, et al. Renewable and cooling aware workload management for sustainable data centers [C]//Proceedings of the 2012 ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, London, Jun 11-15, 2012. New York: ACM, 2012: 175-186.

[13] BROWN M, RENAU J. ReRack: power simulation for data centers with renewable energy generation[J]. SIGMETRICS Performance Evaluation Review, 2011, 39(3): 77-81.

[14] GOIRI I, LE K, HAQUE M E, et al. GreenSlot: scheduling energy consumption in green datacenters[C]//Proceedings of the 2011 Conference on High Performance Computing Networking, Storage and Analysis, Seattle, Nov 12-18, 2011. New York: ACM, 2011.

[15] GOIRI I, LE K, NGUYEN T D, et al. GreenHadoop: leveraging green energy in data-processing frameworks[C]//Proceedings of the 7th EuroSys Conference 2012, Bern, Apr 10-13, 2012. New York: ACM, 2012: 57-70.

[16] GOIRI I, KATSAK W, LE K, et al. Parasol and GreenSwitch: managing datacenters powered by renewable energy [C]//Proceedings of the 2013 Architectural Support for Programming Languages and Operating Systems, Houston, Mar 16-20, 2013. New York: ACM, 2013: 51-64.

[17] KATSAK W, GOIRI I, BIANCHINI R, et al. GreenCassandra: using renewable energy in distributed structured storage systems[C]//Proceedings of the 6th International Green and Sustainable Computing Conference, Las Vegas, Dec 14-16, 2015. Washington: IEEE Computer Society, 2015: 1-8.

[18] WANG Q, CAI H R, CAO Q, et al. An energy-efficient power management for heterogeneous servers in data centers[J]. Computing, 2020, 102(7): 1717-1741.

[19] LI C, ZHANG W Y, CHO C B, et al. SolarCore: solar energy driven multi-core architecture power management[C]//Proceedings of the 17th International Conference on High-Performance Computer Architecture, San Antonio, Feb 12-16, 2011. Washington: IEEE Computer Society, 2011: 205-216.

[20] SHARMA N, BARKER S, IRWIN D, et al. Blink: managing server clusters on intermittent power[C]//Proceedings of the 16th International Conference on Architectural Support for Programming Languages and Operating Systems, Newport Beach, Mar 5-11, 2011. New York: ACM, 2011: 185-198.

[21] HAQUE M E, GOIRI I, BIANCHINI R, et al. GreenPar: scheduling parallel high performance applications in green datacenters[C]//Proceedings of the 2015 International Conference on Supercomputing, Newport Beach, Jun 8-11, 2015. New York: ACM, 2015: 217-227.

[22] LIU L J, LI C, SUN H B, et al. HEB: deploying and managing hybrid energy buffers for improving datacenter efficiency and economy[C]//Proceedings of the 42nd Annual International Symposium on Computer Architecture, Portland, Jun 13-17, 2015. New York: ACM, 2015: 463-475.

[23] LIU L J, SUN H B, LI C, et al. Exploring highly dependable and efficient datacenter power system using hybrid and hierarchical energy buffers[J]. IEEE Transactions on Sustainable Computing, 2021, 6(3): 412-426.

[24] SHEN L F, WANG F X, WANG F, et al. Backup battery allocation and workload migration against electrical load shedding at edge[J]. IEEE Internet of Things Journal, 2023, 10(19): 16804-16815.

[25] ZHOU X, CAI H R, CAO Q, et al. GreenGear: leveraging and managing server heterogeneity for improving energy efficiency in green data centers[C]//Proceedings of the 2016 International Conference on Supercomputing, Istanbul, Jun 1-3, 2016. New York: ACM, 2016.

[26] CAI H R, CAO Q, JIANG H, et al. GreenHetero: adaptive power allocation for heterogeneous green datacenters[C]//Proceedings of the 2021 IEEE 41st International Conference on Distributed Computing Systems, Washington, Jul 7-10, 2021. Piscataway: IEEE, 2021: 160-170.

[27] DENG N, STEWART C, LI J. Concentrating renewable energy in grid-tied datacenters[C]//Proceedings of the 2011 IEEE International Symposium on Sustainable Systems and Technology, Chicago, May 16-18, 2011. Piscataway: IEEE, 2011: 1-6.

[28] YAHOO! Inc. YAHOO! cloud serving benchmark in C++ [EB/OL]. (2014-12-07) [2023-10-10]. https://github.com/basicthinker/YCSB-C.

[29] KHAN K N, HIRKI M, NIEMI T, et al. RAPL in action: experiences in using RAPL for power measurements[J]. ACM Transactions on Modeling and Performance Evaluation of Computing Systems, 2018, 3(2).

[30] National Renewable Energy Laboratory. Solar and wind datasets[EB/OL]. (2015-05-15) [2023-10-10]. https://www.nrel.gov/grid.

李大平（1992—），男，河南南阳人，博士，工程师，主要研究方向为计算机系统结构、分布式存储系统。

LI Daping, born in 1992, Ph.D., engineer. His research interests include computer system structure and distributed storage system.

史庆宇（1992—），男，湖南长沙人，博士，讲师，硕士生导师，CCF会员，主要研究方向为数据中心网络、分布式存储系统。

SHI Qingyu, born in 1992, Ph.D., lecturer, M.S. supervisor, CCF member. His research interests include data center network and distributed storage systems.

唐忆滨（1989—），男，湖北武汉人，博士，高级工程师，CCF会员，主要研究方向为智能计算系统、智能存储系统。

TANG Yibin, born in 1989, Ph.D., senior engineer, CCF member. His research interests include intelligent computing system and intelligent storage system.

胡哲琨（1986—），男，湖南常德人，博士，研究员，硕士生导师，CCF会员，主要研究方向为异构分布式计算、云原生系统。

HU Zhekun, born in 1986, Ph.D., professor, M.S. supervisor, CCF member. His research interests include heterogeneous distributed computing and cloud native systems.

高毅（1984—），男，湖北鄂州人，博士，研究员，硕士生导师，CCF会员，主要研究方向为计算机系统结构、云计算。

GAO Yi, born in 1984, Ph.D., professor, M.S. supervisor, CCF member. His research interests include computer system architecture and cloud computing.