转载：NVMe SSD如何用之应用端缓存加速

时间：2021-03-04 来源：本站 点击：4277次 分享到：

[摘要] 本文将介绍iCAS软件加速方案，以及企事录实验室利用iCAS在加速数据库方面的测试分享。

iCAS，即Intel Cache Acceleration Software（英特尔缓存加速软件），是Intel公司推出的一款轻量级缓存加速软件，其安装在应用服务器之上，利用应用服务器上的SSD，对本地存储、外置SAN存储或者直连JBOD等进行加速。可运行于Windows和Linux两大类别的操作系统平台，在缓存加速方面支持包括Write-Through和Write-Back在内共4种模式。

在对这4种模式进行简单介绍之前， 先给出名词解释：

Cache 设备——在这里指相对性能更高、容量更小、价格更高的设备。

核心设备（core device）——在这里指相对性能更低、容量更大、价格更低的设备——用于数据的持久化存储。

Cache设备和核心设备是相对的概念。譬如，以SATA SSD为cache设备， 则机械硬盘可以是核心设备；如果NVMe SSD为cache设备， 则SATA SSD或者机械硬盘都可以是核心设备；如果以Optane SSD为cache 设备， 则NAND的NVMe SSD、SATA SSD、机械硬盘都可以是核心设备。

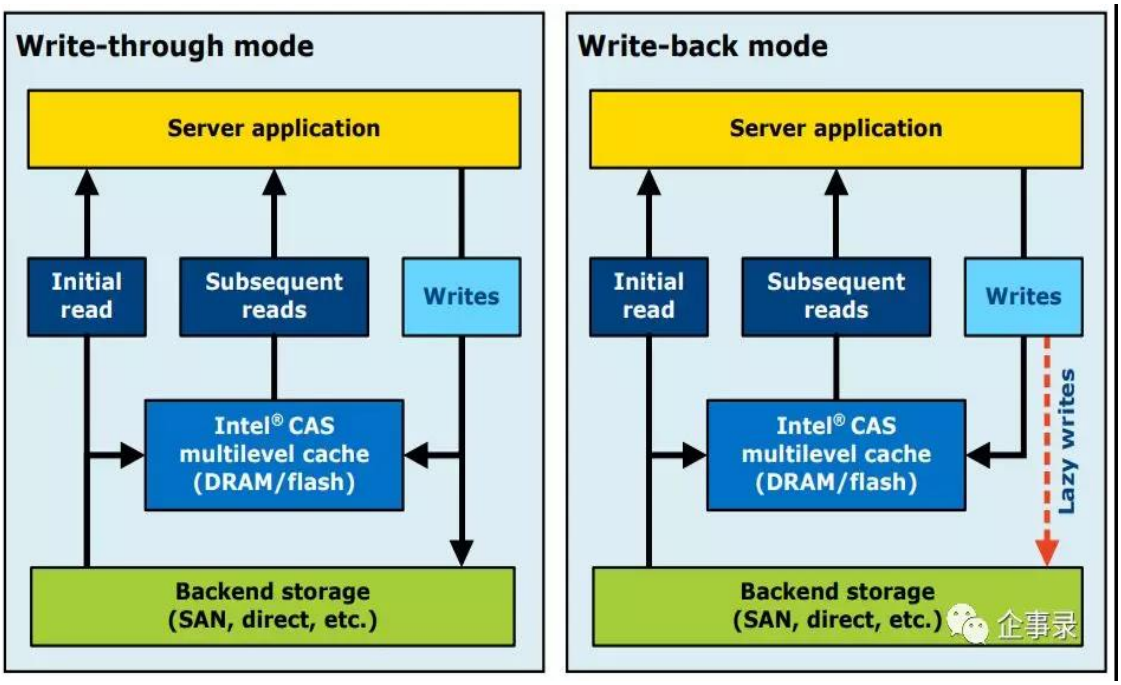
CAS软件可以工作在如下模式：

Write-through模式。在该模式中，iCAS在往cache设备写数据的同时也向core设备同步写。该模式保障cache里的数据和core里的数据100%同步的，对读密集型操作更有效。Core里的数据对共用的其他服务器都有效。

Write-back模式。iCAS先往cache里写数据。一旦写入cache 成功， 便对应用返回确认写成功。这个确认发生在数据被写入core设备之前。 然后，周期性地，cache里的数据会在既定的机会里写入core设备。该模式对读密集型操作和写密集型操作都同时提升。

Write-around模式。只有当cache里已经有了数据块（block），iCAS才会将数据同时写入cache和core设备。和write-through类似，cache里的数据100%同步于core设备。然而， write-around进一步优化了cache, 避免“cache污染”。比如，应用要写入的数据在后续确定不会被经常重读，数据便不用写入cache而是仅仅直接写入core设备。这个模式对读密集型操作更有效。

Pass-through模式。这个模式下，iCAS略过cache设备。当用户在真正启用缓存设备前，可以利用这个模式把所有准备被缓存的core device关联起来。一旦这些core设备被关联好以后， 用户可以动态地切换到他们所要的cache模式。



Intel缓存加速软件提供的两种加速模式：Write-through和Write-back，均可对本地存储、SAN存储和直连存储进行缓存加速

从上图可以看出，iCAS对数据读写操作都有加速作用：读缓存（Read Cache）和写缓冲（Write Buffer）。顾名思义，Write-through和Write-back两种模式的区别体现在写入操作上，读缓存方面工作原理没有区别：在接收到应用的读请求后，先在SSD缓存中查找，如果缓存命中，即读取iCAS中的缓存数据；如果缓存未命中，则从后端读取数据返回给应用，并将数据缓存到SSD中。

在写缓冲中区别就体现出来了：如果使用Write-through模式，iCAS会将应用数据同时写入到SSD缓存和后端数据存储之后，再返回写操作成功。这种模式实际并不能加速写操作，因为其写延迟取决于最慢的返回操作（即后端存储）；Write-back模式则能够加速写操作，如同前者一样，数据会同时写入到SSD缓存和HDD磁盘存储中，但SSD缓存写完即返回操作成功，磁盘存储将在后台继续写入，直到完毕。

iCAS直接安装在应用服务器上，针对存储卷（Volume）进行加速，所以可以在多种应用场景下使用，比如以数据库为代表的块存储场景，文件存储场景，以及虚拟化环境。iCAS在虚拟化场景下的使用方式跟物理机上的使用方式没有不同，其并非安装在Hypervisor层，而是应用虚拟机之上，所以能够针对应用数据进行加速。

目前iCAS支持Windows平台和Linux平台，包括主流使用的Red Hat Enterprise Linux（RHEL）、CentOS、SUSE Linux Enterprise Server（SLES）以及Ubuntu Server等等，可以应用在绝大多数的企业环境之中。同时其安装也很简单，稍有Linux基础的用户通过一两个命令行即可安装使用。

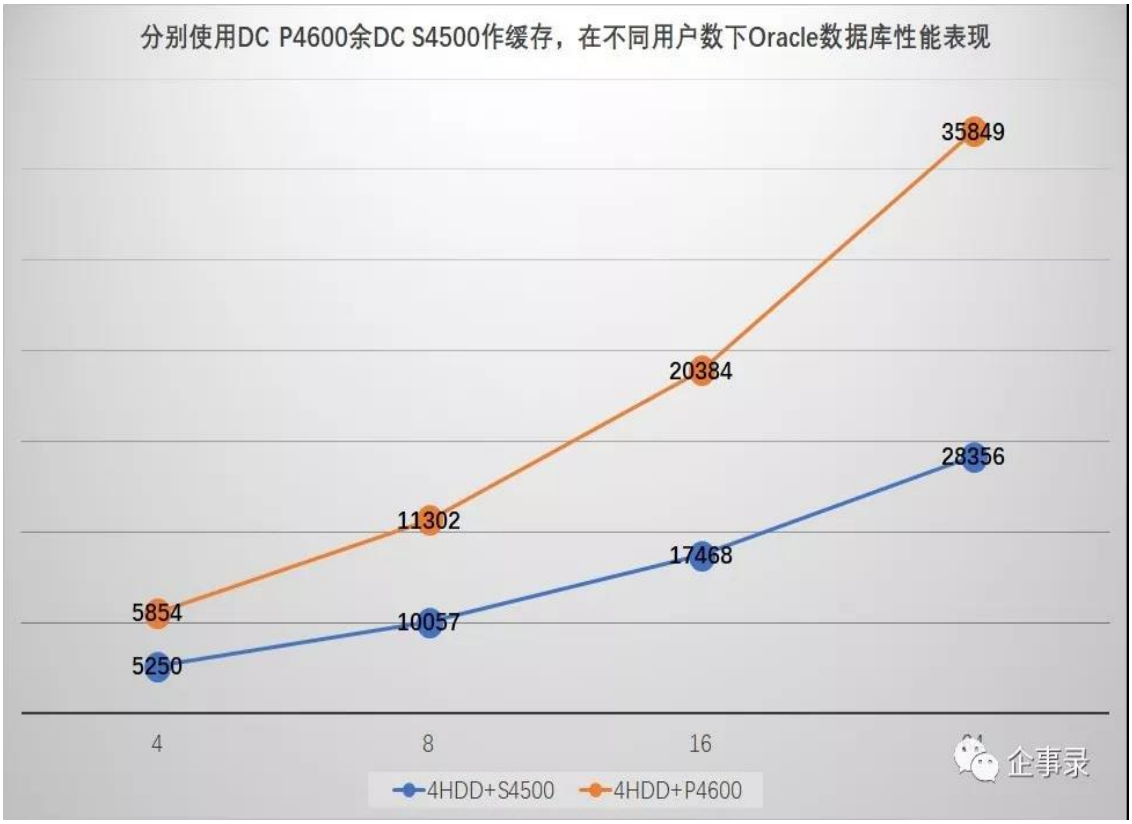
企事录实验室验证测试

在iCAS缓存加速软件的验证测试中，企事录实验室使用与之前评估Intel DC P4500/4600SSD的同台服务器（型号为Intel R2208WFTZSX），配备双路Xeon Gold 6146处理器和256GB内存。将原来的DC P4500/4600替换为希捷（Seagate）公司的Exos 7E8系列大容量硬盘驱动器（5400 RPM，4TB容量），并加装SAS/SATA RAID卡将4片4TB磁盘组建RAID 5；分别保留一块DC P4600和DC S4500作为iCAS缓存：

企事录实验室用于评估iCAS缓存加速性能的测试方案，分别使用Intel DC P4600和Intel DC S4500作为缓存，希捷磁盘作为数据实际存储。在Oracle数据库环境下，评估其在不同用户数并发下的性能表现

在Oracle数据库服务器上安装Intel iCAS软件，将DC P4600和DC S4500分别划出300GB大小分区，用作缓存，启用Write-Back模式，以加速写性能。构建测试数据库后写入120GB数据，分别测试其在4、8、16、64等不同并发数下，Oracle数据库的性能表现。

在开启iCAS缓存加速测试功能之前，企事录实验室原本希望基于HDD存储进行测试，将测试结果作为参照组，以与开启iCAS缓存加速功能后的性能结果进行对比。但由于所采用的大容量磁盘存储在随机读写方面性能较弱，即使是4块HDD（RAID 5）并发的情况下，仍不能为Orace数据库提供满足其所需的性能，导致测试无法正常进行。虽然经过多次调试，仍无法获得结果，企事录实验室遂跳过这一阶段。将DC P4600和DC S4500作为缓存，考量其在Oracle数据库下的应用加速表现：



在DC P4600作为缓存的情况下，Oracle数据库在64用户数并发下获得3.5万TPS性能；而使用DC S4500作为缓存则获得了2.8万TPS性能。在4/8/16/64逐步增加用户数的情况下，Oracle数据库性能随之增长，提升幅度可达20%

需要注意的是，这一测试方案经过优化，iCAS在Write-back下对所有的读写数据都有完全的加速作用，并且测试方案的构建基于中小型应用在数据量不大的情况下，热点数据全部缓存SSD中，所以取得了最理想的加速效果。

同时，Oracle数据库作为了一个完整应用，考量的是整个硬件平台/组件的综合性能，不能也无法排除内存对于Oracle数据库的加速影响。为了尽可能发挥Intel新一代硬件平台中六通道内存的性能，测试用数据库SGA人为设置为64GB大小，其对数据库性能有着不小的性能增益。

并且，结合企事录实验室以往的测试结果发现，高主频的Xeon Gold 6146处理器（3.2GHz，可睿频到4.2GHz）在高性能的NVMe SSD环境下，其**对数据库等强计算性能应用利好。**

企事录实验室建议

在实验室条件下验证了Intel公司的iCAS缓存加速软件确实能够大幅提升应用服务器的性能，能够接近或者达到全闪存的性能，但这都是严格控制实条件的情况下获得的。如果要将iCAS缓存加速软件用于实际应用环境，并获得较好的加速效果，那么用户需要注意：

首先，用于iCAS的SSD应该是耐写型的，即写入性能较高、写入寿命较长。因为作为读缓存和写缓冲的设备，不仅要承受所有的写入数据量，缓存热点数据（加速读取）也要根据时间的推移更新——这也意味着数据的写入。如果写入性能较差，则写缓冲对写入操作的加速不明显；如果写入寿不够，则可能会提前耗尽（损坏）。好在，写入性能好的SSD，写入寿命通常也会比较长。

在我们测试所用的SSD中，P4600是所在系列中相对耐写型的（与P4500相比）——论写入还比不了天赋异禀的P4800X，但胜在容量大、价格有优势。S4500不以写入见长，但限于条件我们手里没有写入特性更好的S4600。如果用户需要iCAS搭配SATA SSD使用，建议选择DC S4600。

其次，计算不能成为瓶颈。较高性能的CPU，并配备适量的内存容量。企事录实验室认为，较高的计算性能是iCAS缓存加速软件发挥作用的前提条件，较多的内核（例如12核及以上）或更高主频的CPU都对数据库性能利好，相较而言，更高主频的CPU在数据库性能提升方面更为直接简单。同时，更大内存容量能够给数据库提供更多高速缓存，能够大幅提升性能。

再者，应当注意SSD（缓存）和HDD（数据实际存储）的容量比例，更大容量的SSD缓存能够明显提升缓存命中率，尽可能减少对磁盘的读写操作，或者尽可能让磁盘处于顺序读写状态，都对性能利好。一般而言，SSD缓存与HDD实际存储的容量比例保持在1:10是一个较好的状态，如果SSD缓存的比例更高，则显著提升iCAS的性能表现。

最后，SSD缓存的容量配比要结合应用实际，即分析应用产生的数据增长情况，与热点数据的生命周期。应用数据的增长情况将会直接影响写加速（Write Buffer）的效率，Write Buffer能够完全容纳增长的数据容量，将具有最佳的加速效果。同时，还要与热点数据的生命周期相结合，及频繁访问数据的时间范围，比如应用数据在一个月内有较高的访问频率，超过这个时间范围，其访问频率迅速下降，甚至不再访问，那么读缓存的容量至少要能完全容量一个月以上的数据存储空间需求。

如果热点数据的生命周期为一个月的话，那么SSD缓存的容量应该为一个月的热点数据容量（Read Cache），还要给每天应用所产生的新数据留出足够的写入空间（Write Buffer）。考虑到预估与实际情况会存在一定的偏差，SSD缓存在满足上述条件的同时，再留有一定的剩余容量，在实际应用中将更有成效。

上一篇：东芝等大厂陆续宣布扩增NAND Flash产能，2019年市场恐供过于求

下一篇：今年上半年服务器品牌出货排名出炉，Inspur跃升至第三、Lenovo退居第五