

存储区域网络在联机事务处理下的性能研究

苟彦李华

(重庆大学计算机学院 重庆 400044)

摘要 存储区域网络是近年来流行的一种基于“块”数据访问的存储解决方案,应用于很多领域。联机事务处理也是目前应用很多的业务,将联机事务处理业务运行在存储区域网络中,能够提高业务的质量。如何保证业务能够快速稳定地运行,或一旦出现问题能够在最短的时间内解决问题都是大家关注的事情。这就需要我们研究其性能,进行性能优化,保障系统高质量运行。

关键词 存储区域网络,联机事务处理,性能因素,性能瓶颈,性能调优

Performance Research of Storage Area Network under the Online Transaction Processing

GOU Yan LI Hua

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract Storage area network is a storage solution based on "block" data access, it is popular in recent years and used in many areas. Online transaction processing is popular business. Online transaction processing business is running in a storage area network, to improve the quality of business. How to ensure the business can run fast and stable, or how to resolve problems in the shortest time is something that we are concerned about. This requires us to research their performance and optimize it, in order to support system running with a high quality.

Keywords Storage area network, Online transaction processing, Performance factor, Performance bottleneck, Performance optimization

1 引言

由于互联网的高速发展、信息量剧增,传统的存储系统已无法满足需求,因此出现了网络存储技术。存储区域网络(Storage Area Network, SAN)是一个高性能网络,它的基本目的是使存储设备与计算机系统或存储设备与存储设备之间通信^[1]。SAN网络特别适合于有大量数据交换、对可靠性要求很高的存储系统,其应用于很多领域。因为目前大业务量用户数量多,一旦不能满足要求,轻则造成用户长时间等待服务器响应,重则导致服务器资源耗尽,服务不可用。因此,在业务推广之前,必须经过严格的性能测试,对其进行性能优化,使其性能达到最优^[2]。

联机事务处理(On-Line Transaction Processing, OLTP)也称为面向交易的处理系统,是传统的关系型数据库的主要应用,主要用于基本的、日常的事务处理,在金融、证券、期货以及信息服务等系统得到广泛应用。SAN网络在不同业务下所表现的性能不一样,在OLTP下如何组建网络、如何配置各项参数使性能最优是很重要的。

由于网络环境的复杂性,通过网络传输数据,每一个环节都可能导致数据堵塞或者响应过慢,这样,就会影响业务的正常使用。只要我们根据具体情况分析出问题所在,正确找出结症,就能够很快地将问题解决。特别是现在的网上事务发展迅速,在业务能够运行的情况下,提升性能将会提升整个系统的服务级别,更能保障系统在特殊环境(高/低负载)下顺利

运行。

2 影响存储区域网络性能的关键因素

在SAN网络中有3个主要的部件:服务器系统、存储设备和互联设备。改善SAN性能的方法主要都集中在这3个部件中。

2.1 存储区域网络的性能指标

关于SAN网络性能的研究,首先需要弄清楚评价性能优劣的指标和影响性能的主要因素,这样才能根据现象快速找出性能问题所在。

衡量存储性能的最重要指标如下:

IOPS(Input/Output Operations Per Second)即每秒读写I/O操作的次数,指的是系统在单位时间内能处理的最大I/O频度。IOPS随着数据库的增大而不断减少,当达到了存储处理能力的极限后,再减少数据块的大小对增大IOPS并没有任何帮助。一般情况下,OLTP应用涉及更多地频繁读写,更多地考虑IOPS。

带宽(Throughput)即为吞吐量,存储单位时间内能处理的数据的总量,衡量标识了最大的数据传输量。带宽值随着数据库的增大而增大,只是增大的趋势逐渐放缓,直到趋于平稳(达到带宽的瓶颈)。

I/O响应时间是从发出一个IO请求到接收到IO响应的的时间。一般地,随IOPS增加,因为IO出现等待,IO响应时间也会随之增加。对一个OLTP系统,10ms以内的响应时间是

比较合理的。当然,还有一些指标可以与以上指标结合起来综合衡量一个系统性能的优劣,如波动率。

2.2 影响存储区域网络的关键因素

根据 SAN 网络的特点,找出影响性能的关键点,主要有以下几个因素:raid 组级别、成员盘个数、读预取算法、负载均衡算法、IO 调度算法等。

RAID(Redundant Array of Independent Disks)组级别: Raid 是由多个独立的高性能磁盘驱动器组成的磁盘子系统。Raid 组级别其实是一种算法,是将 N 个硬盘结合成虚拟单台大容量的硬盘使用的方法。不同级别采用的工作模式和算法不同,因此具有不同的可靠性和性能。选择 raid 时,应根据应用的需要选择相应的级别^[3]。目前较流行的是 raid10 和 raid5 级别组。

成员盘个数:将多少个磁盘划分成一个 raid 组(一个虚拟大容量磁盘),不同 raid 组级别,成员盘个数不同,性能也不尽相同。

分条深度:在划分 raid 组时,需选择分条深度。不同的业务由于数据块的大小不同,其选择的大小也不尽相同,一般情况下,对于 raid10 来说,分条深度选择 512kB 较为合适。

读预取和写惩罚:在对数据进行读操作时,为了提升其性能,往往需要将数据从磁盘中预取到内存中。不同的读预取策略适合于不同的场景,例如,固定预取就适合于业务单一,IO 顺序性强的场景。如多为随机小 IO 的场景,预取一般不起什么作用。写惩罚是为确保数据的一致性,在数据写入的过程中,必须先读取特定区域的某些数据,然后与需写入的数据一起写入磁盘。

负载均衡:根据业务的 IO 特性,保证所有的资源得到公平的利用,提高并行处理能力,消除资源热点,在硬件环境和软件代码无法优化或已经优化了的情况下,负载均衡是性能调优的精髓。在 san 网络中,可以从前端主机通道负载均衡、控制器负载均衡、后端通道负载均衡、硬盘单元负载均衡几个方面给予优化,这就涉及到相关的负载均衡算法。

3 存储区域网络在联机事务处理下的性能调优

3.1 SAN 网络配置优化

OLTP 基本特征是顾客的原始数据可以立即传送到计算中心进行处理,并在很短的时间内给出处理结果。OLTP 可以及时地处理输入的数据,及时地回答。衡量联机事务处理的一个重要指标是系统性能,具体体现为实时响应时间(Response Time),即用户在终端上送入数据后,到计算机对这个请求给出答复所需要的时间。

根据联机事务处理的特点,我们可以从以下几个方面的设置来提升存储区域网络的整体性能。

Raid 组级别选择:持续顺序读性能方面,raid5 性能略高于 raid10;持续顺序写性能方面,在前端(服务器)压力足够的情况下,raid5 性能明显高于 raid10,且 raid10 写带宽会略高于 raid5 的写带宽值的一半,在前端压力较小的情况下,raid5 的带宽小幅优于 raid10;在持续随机读性能方面,raid5 和 raid10 性能基本相当;在持续随机写性能方面,raid5 性能明显低于 raid10。就空间利用率来说,raid5 明显优于 raid10。安全性方面来说,raid5 低于 raid10。总结:总体性能方面,大

文件读写时,raid5 的性能会明显好于 raid10;在随机小数据块读写为主(如联机事务处理)时,raid10 是最优的选择。

成员盘个数选择:在 raid 组级别为 raid10 时,成员盘的个数越多,则下发或读取数据越快,但当达到一定值时,性能会趋于一个稳定值,一般情况下,取 12 盘为一个 raid 组。

分条深度选择:在划分 raid 组时,需选择分条深度。不同的业务由于数据块的大小不同,其选择的大小也不尽相同,一般情况下,对于 raid10 来说,分条深度选择 512KB 较为合适。

读预取与写惩罚:预取策略大致分为固定预取、可变预取、只能预取和不预取几种。固定预取按系统设置的固定预取窗口大小进行预取,主要适用于业务单一,IO 顺序性强的场景,如 DSS 和数据备份等业务,建议预取窗口与 LUN 分条对齐。可变预取按系统设置的倍数预取,读 IO 长度 * 倍数,主要适用于顺序性强,但 IO 大小呈规律性变化的业务。智能预取按 IO 顺序程度,动态启停预取,动态调整预取窗口大小,主要适用于 IO 特性比较复杂的场景,如 OLTP 应用中数据文件的访问。不预取适用于可明确的 IO 随机场景。写惩罚是随机写的业务中影响性能的关键因素,这也是我们在 OLTP 业务下选择 raid10 而不选择 raid5 的原因之一。

负载均衡:前端主机(服务器)通道负载均衡要求组网时保证主机与存储可控制器间有多条路径连接,通过软件控制。

控制器负载均衡要求在 raid 组和 lun 配置时,保证将业务的负载平均分配到现有的多个控制器上,保证控制器的 CPU、cache 和通道资源得到有效利用。

后端通道负载均衡要求均衡控制器到磁盘间数据传输通道的负载,如后端包含多个环路,需保证各环路压力的负载均衡。因此在大压力情况下,有必要按跨环路的方式创建 raid 组。

硬盘单元负载均衡:由于硬盘单元在随机应用中最容易出现瓶颈点的模块,其 IO 处理速度远小于 CPU 和内存,因此有必要在划分 raid 和 lun 时仔细考虑,尽量为高负载的区域分配更多的硬盘单元,避免在热点区域出现导致部分硬盘出现瓶颈,而其他盘却很闲的情况。

3.2 分析和解决性能瓶颈

系统出现性能问题时,如性能数据波动大,稳定性不够,性能数据不能满足所要求的值,IO 延迟时间长,用户明显感觉业务响应慢等情况,怎样快速准确地找到和解决瓶颈问题是现实中最直接的问题。IO 从服务器的用户缓冲区开始一直将数据传送到存储端的物理磁盘上,这个过程中,服务器和存储的各层各模块都有可能发生阻塞,我们需要查看各层的响应时间和 IO 特性来确定其瓶颈。

1. 确认主机是否正常把压力传递给存储。若正常传递,说明存储端出现问题,若没有正常传递,则说明主机端出现问题。

2. 在找出瓶颈出现的位置后,则需要分析出现瓶颈的原因。一般情况下分为硬件和软件的原因。检查设备各主要模块或器件有无故障,如确定不是故障问题,则定位为某硬件已不能满足该业务的需求或者该层软件模块的设计问题。

3. 分析主机端或存储端的 CPU、内存、硬盘和通道等占用情况。

(1)CPU 占用

CPU 占用现象	潜在问题
硬中断处理>系统调用	系统可能存在硬件问题,导致硬件中断频繁
上下文切换值过高	系统线程设计上可能存在问题,线程数过多导致上下文切换频繁
IOWAIT 过高,其他应用 CPU 利用率较低	存储系统可能存在问题,导致 IO 延时过大,系统阻塞等待 IO 返回
CPU 平均队列深度>2	CPU 无法快速处理请求,导致请求堆积

根据 CPU 资源的使用情况判断是 IO 瓶颈还是 CPU 运算能力或数量不够,如 CPU 已无法快速处理请求,则需要更换更快的 CPU 或增加 CPU 的数量。

(2)内存

内存资源主要为 CPU 提供高速缓存,保证数据快速处理,分为物理内存和虚拟内存。

内存占用现象	潜在问题
内存使用 80% 以上 (Linux 系统下重点关注 free page 指标)	内存资源被大量数据占用,即将成为瓶颈
虚拟内存占用>75%	内存资源基本耗尽,系统被迫采用磁盘空间作为缓存,内存资源成为瓶颈
页面请求延时过大	系统内存资源争用严重,导致申请页面请求出现等待
内存命中率低	应用系统设计不恰当或 OS 预取策略存在问题

查看 SWAP 区的利用率,若利用率高,表明 SWAP 区存在较频繁的页面换入换出,说明该内存偏少无法满足应用需求。若利用率低,说明内存容量可以满足。

(3)硬盘

硬盘使用现象	潜在问题
硬盘利用率>70%	硬盘资源已被大量占用,硬盘出现性能瓶颈
IO 延时过大	对于 OLTP,延时一般在 10ms 以内,不能超过 30ms,如延时过大,则说明硬盘已成为瓶颈
单盘 IO 队列深度>IO	硬盘 IO 堆积严重,已成为瓶颈
业务实际下发 IO 请求数>存储侧接收 IO 请求数	IO 出现拆分,主机的卷管理设置或 HBA 卡设置可能出现问题

若检查出个别盘是慢盘,则更换该磁盘。

若每块物理盘都过载,说明磁盘阵列当前配置下的 IO 能

力无法满足应用对 IO 的需求,产生 IO 瓶颈,导致应用响应缓慢,处理时间过长。

若只有某个或 N 个逻辑卷片区存在与服务器 CPU 的“IO 等待”相对应的明显高的读写操作时段,说明该系统的 IO 请求超出了该卷的 IO 吞吐能力,使该磁盘组中的大部分磁盘过载,从而造成系统的服务器 CPU 进程时间都处于等待 IO 的闲置状态。这时需查看该逻辑卷片区上哪些数据文件被应用频繁访问,造成了访问过载,将访问过载的数据文件从该磁盘阵列上迁移至高性能的存储设备(SSD)上存储,或者分配更多的磁盘单元。

(4)通道

物理通道主要关注两个指标。

通道带宽利用率——利用率高达 90% 以上时,说明该传输通道已出现性能瓶颈,需扩大通道传输带宽或减少传输数据量。

通道每秒出现的数据传输错误次数——每秒出现 3 次以上的传输错误或误码时,说明传输通道受到干扰。

(5)san 网络模块较多,每层都可能是影响性能的短板,随着业务的不同,读写的 IO 大小和个数都不相同,那么每层设备的参数(如并发数)也会不同。若在某层出现 IO 堵塞情况,也有可能是该层的并发数设置较小,可适当增大。

结束语 本文分析了存储区域网络在联机事务处理下的性能研究,并提出了分析和解决性能瓶颈的方法,使读者可以快速准确地搭建一个完善的存储区域网络并对其进行性能调优,使整个系统能够高速稳定地运行。

参考文献

- [1] Barker R, Massiglia P. 存储区域网络精华-深入理解 SAN[M]. 舒继武,等译. 北京:电子工业出版社,2004:4
- [2] 熊军军. 架构应用软件性能测试问题研究[EB/OL]. <http://se.csai.cn/testnews/201006081348351467.htm>, 2010-06
- [3] 赵文辉,徐俊,周加林,等. 网络存储技术[M]. 北京:清华大学出版社,2005:14-20

(上接第 171 页)

在图 5 中,图 5(a)为加入幅度为 0.01 的白噪声的 600 个随机数据,即为加噪样本,图 5(b)为函数的真实图像,图 5(c)为 RBF 神经网络通过对加噪样本进行训练得到的函数曲面,图 5(d)为加噪样本经过免疫网络模型处理后再进行 RBF 神经网络训练得到的函数曲面图。

表 2 实验 2 结果的性能比较

逼近方法	输入样本数	是否加噪	测试数据数	平均误差
RBF 网络	600	是	600	0.0549
MFAIN 算法的 RBF 网	600	是	600	0.0404

通过实验 2 可以进一步说明,基于 MFAIN 算法的 RBF 网不仅对噪声有一定的平滑作用,而且存在误差率低、泛化能力强、速度快等特点。

结束语 针对目前人工免疫网络研究和应用的不足,本文给出了一种基于 MFAIN 算法的 RBF 网模型,在改进现有的人工免疫网络的基础上结合 RBF 网并加以应用。实验表明,MFAIN 算法具有减少数据冗余、平滑噪声、提取特征数据的功能,且应用中可以有效提高问题求解的速度和性能。

下一步工作将尝试将人工免疫网络模型算法与其他算法的性能进行比较,考虑与其他方法结合,并将其用于解决模式分类、系统辨识、函数逼近、模型预测等领域的一些工程问题。

参考文献

- [1] de Castro L N, Timmis J. Artificial immune systems as a novel soft computing paradigm[J]. Soft Computing, 2003, 7(8)
- [2] Timmis J, Knight T, De Castro L N. An overview of artificial immune systems[M]. Natural Computation Sries, Springer, 2004
- [3] Timmis J. Artificial immune systems—today and tomorrow[J]. Sringer Science Business Media B. V, 2006
- [4] 岳训,迟忠先,莫宏伟,等. 人工免疫网络模型的数据特征提取性能评价技术[J]. 小型微型计算机系统, 2007(5)
- [5] Chen Yun-fang, Wang Ru-chun. A Classification Algorithm Based on Artificial Immune. Chinese[J]. Journal of Electronics, 2008
- [6] 谢景新,程春田,全磊光. 改进的免疫算法在函数优化中的应用[J]. 哈尔滨工业大学学报, 2009(7)
- [7] 韩立群. 人工神经网络教程[M]. 北京:北京邮电大学出版社, 2006
- [8] 耿生玲,康宝生. 基于 SOFM 与 RBF 神经网络的自由曲面重建[J]. 计算机工程与设计, 2007(6)
- [9] 焦李成,杜海峰,等. 免疫优化计算、学习与识别[M]. 北京:科学出版社, 2007:17-170
- [10] 莫宏伟. 人工免疫系统原理与应用[M]. 哈尔滨:哈尔滨工业大学出版社, 2003
- [11] 阎平凡,张长水. 人工神经网络与模拟进化计算[M]. 北京:清华大学出版社, 2005