RS/6000小型机系统性能优化

RS/6000 Minicomputer-system Performance Optimization

朱光远 （青岛市农村信用联社 266071）

摘要：RS/6000小型机的性能优化是提高主机系统利用率的最有效手段，通过合理的内存分配提高文件的访问效率，调整缓冲区减少文件使用时的磁盘访问，设置调页空间加快系统的反应速度，利用逻辑卷管理的数据条块化功能减少磁盘资源的竞争，优化I/O Pacing避免大规模的I/O请求耗尽CPU时间等等，可极大提高主机的运行性能。

关键词：RS/6000小型机　性能优化　提高性能

随着计算机技术的飞速发展，RS/6000小型机的应用越来越普及。在应用过程中经常会遇到诸如小型机的使用效率低下，性能未充分发挥等棘手问题，造成了应用系统瓶颈，妨碍了正常的生产运行，必须采取措施加以解决，这便涉及到了小型机的性能调整问题。本文基于长期的RS/6000小型机系统性能调试经验，通过系统管理的几个方面对小型机的性能优化进行了描述。文章后的具体实例为某银行应用系统瓶颈的分析和优化，相信会在此方面给读者一定的启发。

1 系统内存管理

AIX操作系统中当很多进程都要求很多的内存时，彼此之间就会发生竞争，一旦出现这种情况，系统会自动把暂时不用的内存页面调换到磁盘上，即发生调页。很多数据库系统使用文件系统来存放数据，再AIX中，往往采用大量的缓冲区保存最近读写的文件块来提高文件访问的效率，而数据库系统本身也会把大量的数据缓冲区驻留在内存中以提供数据库系统的工作效率，这就造成了内存使用上的冗余，在内存不足时，由此产生的频繁调页行为将会严重的影响系统的性能。另外存放调页空间的磁盘也必将成为I/O访问的瓶颈，最终整个系统吞吐能力会降到及其可怜的低点。AIX中对缓冲区页面的调度策略是可以通过参数来控制的，需要强调的是，对操作系统的工作方式的任何修改必须加倍小心，否则很可能导致系统运行环境更趋于恶劣，甚至崩溃！有四个与缓冲区页面的调度策略相关的AIX系统参数：

(1) minfree：最小的空闲页链表尺寸，一旦低于改值，系统开始换出一部分最近不访问的内存页（称之为偷页），以填充空闲页链表。

(2) maxfree：最大空闲页链表尺寸，一旦高于该值，系统停止偷页。

(3) minperm：用于文件I/O访问的最少缓冲区页数。

(4) maxperm：用于文件I/O访问的最多缓冲区页数。

修改这些参数，必须以root用户身份执行命令vmtune。注意：不同操作系统版本吓得vmtune命令有可能不同，在使用前最好仔细阅读该命令的帮助，否则可能导致整个操作系统的失败。

Minfree值得设置可依据有响应时间要求的程序代码段尺寸，至少空闲页链表中有足够的空闲内存页，以免装入该程序代码到内存时，不需要先偷页腾出空间。maxfree值必须大于max（8，maxpagahead）（意思是取maxpgahead和8之间的大者）。例如：你得minfree值设置为128，而maxpagahead值为16，那么你可以执行命令：/usr/lpp/bos/samples/vmtune-f 128 -F 144把minfree设为128，而maxfree设为144。

2 文件缓冲区管理

AIX缓冲区的作用在于减少使用文件时的磁盘访问。如果设置得太小，会导致磁盘过于繁忙，甚至有个别磁盘达到饱和的状态；如果设置得太大，则会浪费宝贵的内存资源。

缓冲区的大小可以通过设置minperm和maxperm参数值来调整。一般来说，如果缓冲区的命中率太低（低于90%，可用sar -b命令参看），可以增加minperm值；如果缓冲区的命中率对系统并不是很关键，那么可以通过减少minperm值来增加可用的物理内存。

AIX提供一种松散的方式来控制内存中不同类型页面的比率。主要有两种内存页：存放文件数据的内存页（通常称之为文件型的内存页）和存放程序代码及所使用内存空间的内存页（通常称之为计算性的内存页），这两种内存页的比率控制也是通过设置minperm和maxperm参数值来实现，具体说明如下：

(1) 如果文件型内存页在实存中所占的比例低于minperm值时，调页算法在进行偷页时，将不考虑（某类型内存页）重新调入的频率，偷页的类型包括文件型和计算型。

(2) 如果文件型内存页在实存中所占的比例高于maxperm值时，调页算法在进行偷页时，只

选择文件型的内存页。

(3) 如果文件型内存页在实存中所占的比例介于minperm和maxperm两值之间时，通常内存管理系统偷页时只选择文件型的内存页，但如果文件型内存页的重新调入(即文件中的内容再次被访问)频率高与计算型时，偷页也会包括计算型的内存页。minperm和maxperm的缺省值分别为20和80，表示20%和80%的比例值，用命令：vmtune -p 5 -p 20，把minperm值改为5%，maxperm值改为20%。

(4) 如果数据库文件使用裸设备，由于AIX的文件缓冲区仅仅对文件系统有用，因此可以把minperm和maxperm两个参数设得很低(如分别为5%和20%)，这样数据库系统可以分配到更多的内存空间。

3 调页空间管理

系统调页空间的不足，会导致系统反应令人无法忍受的慢以至于完全挂起，AIX可以动态的增加调页空间，但合理的设置调页空间的大小仍然十分重要，这要根据物理内存大小和应用系统需求来综合考虑。

(1) Lsps命令可以用来查看调页空间的使用情况，vmstat则可以用来监测系统调页的活动情况。

(2) 大多数情况下，调页空间的大小设为物理内存2-3倍比较合适，对于内存大于1GB的高配置系统，调页空间设为物理内存的1.5倍就基本满足需求了。

4 AIX逻辑卷管理

AIX的逻辑卷管理(LVM)提供数据条块化功能，可以把数据均匀的分布在多块硬盘上，减少磁盘资源的竞争，分散磁盘I/O访问，从而提高系统的整体性能。

(1) 设计条块化的逻辑卷，定义一个条块化的逻辑卷之前，需要确定有关的属性。

① 参与的硬盘：至少两块硬盘，硬盘上的I/O访问要尽可能的少，条件许可的话，最好由不同的适配卡驱动。

② 条块单元大小：可以设成从2KB到128KB之间的值(2的n次方)，大多数应用环境下设为32KB或64KB比较合适。

③ 逻辑卷的大小：物理分区(PP)数必须是参与硬盘的整数倍。

(2) 建议的条块化逻辑卷参数，下面的建议参数通常能最大化串行读写的性能。

① 块单元大小：64KB。

② minpgahead值：2(最少的预读页数)。

③ maxpgahead值：参与硬盘个数的16倍(最多的预读页数)，这样最多的预读数据值就等于条块单元大小(64KB)和硬盘个数的乘积(这是因为AIX中存储管理操作的页大小为4KB)，结果就是在理想的情况下，预读将涉及每块硬盘上的一个条块单元。maxfree值：根据maxpgahead值变化(参见前面的内容)。

(3) 其他的考虑事项：虽然调整LVM所能起的性能提升程度与应用系统的类型有关，一般来，对于决策支持系统(DSS)效果明显，对于联机事务处理(OLTP)系统或混合型往往也能有很大的提高。

(4) 避免把不同的数据文件放在同一块硬盘上：如果硬盘上的分区有可能同时被访，那么应该避免把不同的数据文件放在同一块硬盘。例如：由于数据文件和日志文件总是同时存取，所以做条块化时他们应该放在不同的硬盘组中，否则会导致磁盘磁头频繁移动，如果同时访问到物理分区相距很远，情况将更加恶劣，系统性能会受到严重影响。

5 I/O Pacing

I/O Pacing时AIX提供的一项技术，用以限制对一个文件同时执行得I/O请求的个数，这可以避免大规模的I/O请求耗尽CPU时间，使得交互式或CPU需求大的应用的响应时间有所保障。I/P Pacing是通过high-water mark和low-water两个参数来实现控制的，但对一个文件执行得I/O请求达到high-water mark时，所有对该文件写操作的进程将被挂起，只到执行的I/O请求回落至low-water mark时，才重新被唤醒执行。

High-water和low-water marks两个参数可通过smit命令来修改，遵循下面的SMIT路径，可以进入修改界面：

Smit

->System Environment

->Change/Show/Characteristics of Operating System

具体值的设置可以边试边改，一般high-water设为33，low-water设为24作为开始比较合适，如果想关闭I/O Pacing功能，把这两项参数都设为0。

6 某银行系统调试实例

(1) 现象：主机系统经常性自动重起，影响银行联机业务的正常开展。

(2) 小型机系统报错信息：

FE9E9357 0508200001 P H ssa0 DISK OPERATION ERROR

F7863CFE 0508190001 P H pdisk5 DISK OPERATIO

具体信息：

Errpt -a -j fe9e9357

Description

DISK OPERATION ERROR

Probable Causes

DASD DEVICE

Failure Causes

DASD DRIVE

Errpt -a -j f7863cfe

Description

DISK OPERATION ERROR

Probable Causes

DASD DEVICE

Failure Causes

DASD DRIVE

这两个错误号每隔整时报一次。

(3) 原因分析：查看了系统的ERROR REPORT输出，发现在死机自动重启动都有一条LABEL为KERNEL_PANIC的错误记录，详细摘要如下：

LABEL: KERNEL_PANIC

IDENTIFIER: 225E3B63

Date/Time: Mon May 14 22:08:51

Sequence Number: 679284

Machine Id: 000109704C00

Node Id: 000109704C00

Class: S

Type: TEMP

Resource Name: PANIC

Description

SOFTWARE PROGRAM ABNORMALLY TERMINATED

Recommended Actions

PERFORM PROBLEM DETERMINATION PROCEDURES

Detail Data

ASSERT STRING

PANIC STRING

HACMP for AIX dms timeout - halting hung node)

dms为DeadMan Switch的缩写，表示HACMP的一个节点在特定时间间隔内一直处于挂起状态就会自动停止该节点，这使得后备节点能够获得该节点所占有的资源，避免竞争资源的问题出现。

导致节点上的HACMP软件无法复位其内部计数器（HACMP通过其判断系统是否处于挂起状态）的可能原因有很多，例如：某些程序运行的级别很高，不把CPU让出给HACMP进程；或者系统进行大规模的写操作；或者网络忙等都有可能。

(4) 调试优化方法，如果频繁出现这种情况，一般按下面顺序进行调整：

① 设置I/O pacing，I/O pacing用于解决大量磁盘写操作而造成的系统挂起（比如compress写入大文件的情况），用smit chgsys进入设置其中的high-water marks和low-water marks（缺省值都为0，关闭I/O pacing），当一个进程在系统处于high-water marks时写文件，将进入等待状态，直到系统进入low-water marks，一般调整从high-water marks=33、low-water marks=24开始。

② 提高syncd的频率，缺省为60（见/etc/inittab中的syncd项，单位为秒），一般不修改。

③ 增加用于通讯的内存，如果AIX错误日志中指示LOW_MBUFS，则可以增加thewall值，缺省是系统内存的25%，最多可以加到50%，用no -o thewall=xxxxx(单位KB)。

④ 修改HACMP错误监测的时间间隔，在1、2步骤都无法解决问题时，才去修改，菜单路径大致为：

SMIT HACMP--->Cluster Topology--->Configuring Network Modules--->Change/Show a Cluster Network Module，选择网络模块（应该是ether），把failure detection rate改为"Slow"。

另外：I/O、内存、CPU的状态观察可以用下面命令：

vmstat <时间间隔> <次数，省略则一直执行>

其中最后4列，前两列user+sys为CPU繁忙的百分比，free为CPU空闲的百分比，wait为I/O等待的百分比。

通过以上方法对系统性能进行了调整，再未出现系统重起现象。