五．输入输出的性能

1．物理卷和逻辑卷

1）性能有关的基本概念

AIX 通过 LVM 逻辑卷管理器对物理卷（硬盘）和逻辑卷进行管理，在 LVM 使用中涉及到下列和性能有关的概念：

Intra-Policy：数据在单个硬盘上的内部分布原则。硬盘的园柱面可以分为 Inner Edge、Inner Middle、Center、Outer Middle 和 Outer Edge 五个区，根据磁盘臂找到数据所需的平均移动时间，在 Center 区的数据具有最好的性能，Middle 区次之，而 Edge 区最差。

Inter-Policy：数据在多个硬盘上的外部分布原则。是尽量跨越多个硬盘，还是集中在尽量少的硬盘上。前者性能更佳。

Mirroring：每个逻辑分区有几个镜像。

镜像写：镜像数据的写操作是并行写，还是串行写。前者性能好，后者可靠性高。

Strictness：每个镜像拷贝是否分布在不同的硬盘上，提高读性能和可靠性。

镜像写一致性：用于意外发生后，保证镜像数据的一致性。

写效验：写操作结束后，将数据读出，以保证正确写入。

Striping：数据在多个硬盘上的分布条带化，用于提高性能。

希望最好的 LVM 性能，应采取这些措施：不做镜像，如必须镜像，也要镜像在不同的硬盘上，采用并行读写；不做写效验；内部分布按 Center、Middle 和 Edge 的顺序；外部分布尽量跨越多个硬盘。

希望最高的 LVM 可用性，应采取这些措施：每个分区做 2 个镜像（3 个拷贝）；镜像在不同的硬盘上，不同的 SCSI 总线、卡及电源上；采用串行读写；做写效验；外部分布集中在尽量少的硬盘上；每个卷组中至少 3 块硬盘。

2）常用工具

iostat：

该命令可以监视逻辑卷和物理卷的使用情况。当 iowait 太高，如超过 70%，原因可能是存在不平衡的磁盘读写，碎片太多，或因内存不足导致的频繁的 paging。iostat 只反映了物理卷的问题，逻辑卷和文件系统的 I/O 情况，可以用下面的工具来监视。

filemon：

该命令在后台运行，通过 trace 工具采集文件、虚存、逻辑卷和物理卷的活动情况，直到 trcstop 运行后，停止采集生成统计报告。可以从物理卷的读写统计中，反映出数据的连续性，从文件的读写统计中，反映出文件的繁忙程度。

语法：

filemon [-i infile][-o outfile][-d][-Tn][-v][-O opt]

-i：数据输入文件，默认是从实时 trace 采集的数据中输入。

-o：统计输出文件，默认是标准输出。

-d：延迟到 trcon 运行后，trace 才开始采集数据。

-Tn：设 trace 缓存大小，默认是 32000 字节。

-v：不带 v 参数，只看前 20 名。

-O：opt 的可选值是 lf（文件）,vm（虚存）,lv（逻辑卷）,pv（物理卷）,all，用于采集指定类型的数据。默认只有 vm,lv,pv。

lslv 和 lspv：

查看分布原则：	lslv lvname
查看逻辑卷离散情况：	lslv –l lvname
查看逻辑卷的分布位置：	lslv –p hdisk# lvname
查看物理卷的划分：	lspv –p hdisk#

3)优化技术

SCSI 卡和硬盘的关系

多个 SCSI 卡连接不同的硬盘，对提升连续串行的读写性能有帮助，但对随机的读写帮助不大。何时需要增加 SCSI 卡，可以先用 lsdev –Cc disk 观察在同一个 SCSI 卡上的硬盘有哪些，再用 iostat 观察，如果 iowait 较高，而上述硬盘的 Kbps 之和接近该 SCSI 卡速率的 70%，则需要增加 SCSI 卡。当有阵列时，SCSI 卡队列长度应由默认的 40 加大，最大可到 128，可用命令 chdev –l scsi3 –a num_cmd_elems=70 来修改。另外，硬盘的 SCSI ID（0-15）的值越小，其响应的优先级越高。

物理卷的配置

一般性原则：对随机访问较多的应用，采用大数量的小容量硬盘较佳；对串行访问较多的应用，采用小数量的大容量硬盘较佳；

RAID 技术：RAID0 可以提高性能，RAID1 可以提高可用性和读响应时间，RAID2 较少采用，RAID3 适用与 CAD 和 CAM 需要连续访问大文件的领域，RAID4 效验数据集中在一块盘上，存在瓶颈，RAID5 效验数据分散在多块盘上，具高可用性，写比读更快。SCSI 设备会将多个小请求集中成一个大请求，再提交，这个请求默认是 64KB 大小，对于连接采用 RAID 技术盘组的 SCSI 卡，应改大 ODM 数据库 PdAt 中的 max_coalesce 值。

卷组的配置建议

rootvg 只放置操作系统，用户数据和应用数据放置在其它卷组中，将较忙的文件系统放在单独的硬盘上，它的日志逻辑卷放在另一个硬盘上。

逻辑卷的组织

将访问较多的几个逻辑卷分别放在不同的硬盘上；尽量将访问较多的逻辑卷跨越多个硬盘；将访问最多的逻辑卷放在硬盘的 Center 区，将访问最少的逻辑卷放在硬盘的 Edge 区；同一个硬盘上，尽量让访问较多的几个逻辑卷分布的近一些；让逻辑卷在硬盘上的分布保持连续；用 reorgvg 命令重新组织逻辑卷时，将需要较高优先级的逻辑卷放在该命令所带参数的前列。

另外，还可以通过建额外的日志逻辑卷和交换区来提升性能。

2．文件系统

1）文件系统的几个概念

AIX 中普通文件系统的数据块大小是 4K，每个文件系统由许多 Allocating group 组成，文件系统创建时要指定 NBPI（每个 I 节点可指向多少字节的数据），用 Allocating group 的大小除以 NBPI，就是该 Allocating group 中的 I 节点数，每个 I 节点可以有 8 个数据块链接指针，即支持 32K 以内的文件，超出 32K 时，分配一个 4K 块，放链接指针，可放 1024 个指针，即支持 4M 以内的文件，超出 4M 时，用两层链接指针，但第二层只能放 512 个指针，所以在 AIX4 中，最大的文件可支持到 2G（512*1024*4K）。而大文件系统的数据块大小可达 128K，因此对于大文件可以减少系统用于指针块的数目。

由于数据块分配给逻辑文件和逻辑卷是动态增长的，所以会产生影响性能的离散分布；建文件系统时采用压缩选项，也将降低文件系统的可用性和影响性能；另外，用于文件缓冲的内存量和远程文件的访问都会影响性能。

为了解决一个 5K 的文件要占用 2 个 4K 的数据块，会导致 3K 的浪费，AIX 引入碎片的概念，用户可以定义碎片的大小为 512、1024、2048 或 4096 字节，这样上述的 3K 空间可以按相应的大小分割为其它文件所利用，当然这也是以牺牲性能为前提的，如果碎片定为 4096，和数据块一样大，则实际上碎片的作用没有发生，统计表明可能有 45%的空间被浪费。

2）文件系统性能测量

除了上面介绍的 filemon 命令，可以反映哪些文件的读写量大外，再介绍一个命令可以观察文件的位置和碎片。

fileplace

语法：fileplace [-l|-p][-i][-v] filename

-l：看逻辑块

-p：看物理块

-i：看间接块

-v：详细信息

下面列举常见问题的使用命令：

繁忙的文件系统在物理卷上的分布合理吗？------lslv

繁忙的文件系统是否跨越多个物理卷吗？------lslv

繁忙的文件系统是本地还是远程的？------filemon

交换区的活动影响了磁盘的操作吗？------filemon

是否有足够的内存交换运行进程的文件页？------svmon

繁忙的文件有大量的碎片吗？------fileplace

是硬盘的类型较差引起的瓶颈吗？------filemon

是 SCSI 卡的类型较差引起的瓶颈吗？------iostat

3）优化技术

对应用 I/O 的优化，应多采用异步 I/O，少用同步 I/O，两者的区别是，后者提交请求，等待完成，再继续，而前者提交请求，马上返回，有利于数据库和文件服务器类的应用。

文件系统分布较离散时，应用下述方法重组织该文件系统，但 paging、sysdump、boot、log、/tmp 和/usr 不适用这个方法：

init m

cd 到该文件系统的 mount 点

find . –print|backup –ivf/dev/device

umount 该文件系统，重建该文件系统并 mount

restore –xvf/dev/device

init 2

当文件系统有采用小于 4KB 的碎片定义时，一段时间要用 defragfs 命令来消除碎片，提升性能。

对条带化（striped）逻辑卷的串行 I/O，可以优化串行读的预读，通过 vmtune 加大 minpgahead 和 maxpgahead，第一次读时，只读一块，但第二次读的是连续的块时，将读入 minpgahead 块，如果一直是连续读，预读的快数将一直增加到 maxpgahead 块。

为了防止有大量 I/O 的进程在系统中运行，影响其它进程的 I/O 响应，可以通过设置系统的高低水平线来控制，用 smit 或 chdev 修改 maxpout 和 minpout 值，当一个进程的 I/O 请求达到高水平线时，该进程进入睡眠，随着 I/O 完成一些后，请求下降到低水平线，再唤醒该进程，这样其它进程才能得到 I/O 服务。

六．网络性能

1．性能有关的概念

thewall：AIX 提供一部分实存作为网络通讯的缓存，这些缓存依数据大小分为 256 字节的 mbufs 和 4096 字节的 clusters，总量有多少则通过系统参数 thewall 来控制。

MTU：最大传输单元，该参数可调，对应于一个给定的网络接口，网络上所有主机的 MTU 值应相同，这样，会有较好的性能。MTU 不同时，当发包主机 MTU 尺寸大于转发主机 MTU 时，将分割成两个包发送，第二个包的有效数据可能很少，影响性能。

2．分析性能和优化工具

和网络有关的影响性能的原因有：客户端网络接口、网络带宽、服务器网络接口、服务器的 CPU 处理能力、服务器的内存和效率不高的配置等等。可以用以下工具来分析性能问题发生在哪个环节。

netstat

该命令可以查看接口状态、路由表、路由统计、每个协议的统计和特定的接口信息，还可以查看活动的套接字、设备驱动信息和网络数据结构。有以下常用参数：

-i：可以观察每个接口的 MTU、输入输出包和输入输出错误包。

如果 Oerrs/Opkts>1%，则需要增加发送队列尺寸，可以用 lsattr 命令观察当前发送队列尺寸 xmt_que_size：

#lsattr -l ent0 -E

```
bus_intr_lvl    12      Bus interrupt level              False
intr_priority    3      Interrupt priority               False
xmt_que_size   512      TRANSMIT queue size              True
.........
```

修改则使用下述命令：

#chdev -l ent0 -a xmt_que_size=new value

如果 Ierrs/Ipkts>1%，则可能是内存分配的问题，如何处理在下文介绍。

如果 MTU 值不合适，可以通过 ifconfig、chdev 或 smit 来修改，注意满足前面介绍的要求，如：ifconfig en0 mtu 1500。

-m：观察 mbufs 和 clusters 是否够用。

如果 failed 的次数较多，可以用 no –o thewall=new value 来加大用于通讯的内存。

netstat 的参数还有-v、-n 等等，可以通过系统帮助进一步了解。

netpmon

该命令可以观察哪些进程和中断正在处理，多少和网络有关，是什么引起 CPU 空闲时间，网络设备的 I/O 驱动通过哪个接口，队列长度是多少，还能监视套接字的调用和网络文件系统的输入输出请求情况。该命令也要用到 trace 工具采集数据，所以运行 trcstop 后才停止采集数据。有以下常用参数：

-o：指定输出文件，默认是标准输出。

-d：延迟到 trcon 命令执行才开始采集，默认是立即采集。

-T n：设置 trace 缓存的尺寸，默认是 64000。

-v：更多的输出结果，默认只有前 20 位。

-O stats：指定报告的类型，stats 可以是 cpu、dd（网络设备驱动 I/O）、so（套接字）、nfs、all。

no

no 是 Network Option 的简写，用于配置网络相关的属性参数。有以下常用参数：

-a：查看当前值

-d：恢复默认值

-o：修改新值

例如：

no –o thewall=3072 修改立刻生效，重启后修改无效。

no –d thewall 改回默认值。

chdev –l sys0 –a maxmbuf=3072 这条命令可以让 thewall 值的修改永久生效。

no 还可以修改 tcp/udp_sendspace、tcp/udp_recvspace 等参数，但要注意同时要将 sb_max 加大，最好加大到前面这些值的和。

3．NFS 优化

网络文件系统的优化涉及到增加 biod 和 nfsd 的数量，增加可以分配给文件页的虚存数量，修改发送和接收队列的尺寸等方面。

nfsstat

该命令用于反映和 NFS（网络文件系统）、RPC（远程进程调用）有关的通讯情况。

-c：客户机数据

-s：服务器数据

-n：NFS（Network File System）信息

-r：RPC（Remote Procedure Call）信息

-z：清除统计数据

例如：

nfsstat –rc 的输出中，如果 retrans(重传包)>5%*calls（总调用数），或者 timeout(超时)较多，或者 badxid(服务器来不及响应)较高，那么需要分担该服务器的负荷，或者增加 timeout 值，或者升级服务器的处理能力。

nfsstat –rs 的输出中，如果 nullrecv 太高，说明 nfsd 数目太多，需要减少。

用 netstat –p udp 或 netstat –s 观察，如果 socket overflows 太高，则需要增加 udp_recvspace 和 nfsd 的数目，分别用 no 和 chnfs –n #来修改。

如果没有使用 ACL 访问控制功能，在客户机的/etc/filesystems 文件中，将远程文件系统节中加入 options=noacl 可以提高性能。

后记

RS/6000 的性能调整还有 trace 命令和图形工具的使用等等内容，由于使用较少，我们将在以后的文章中介绍。在您看完性能调整的各种方法后，最后希望您了解：性能优化并不是万能的，并非总能解决问题，有时解决一个瓶颈又会导致另一个瓶颈的产生，在试过上面这些方法后，如果您的系统依然低效，您可能真的要卖台新的 RS/6000 了。