存储管理、LVM 和文件系统（一）

本章将论述存储管理、逻辑卷管理(LVM)和文件系统。

§1 逻辑卷存储概念

关于 AIX 逻辑存储的 5 个基本概念是：物理卷、卷组、物理分区、逻辑卷和逻辑分区。在这些概念之间的关系参见图 1：

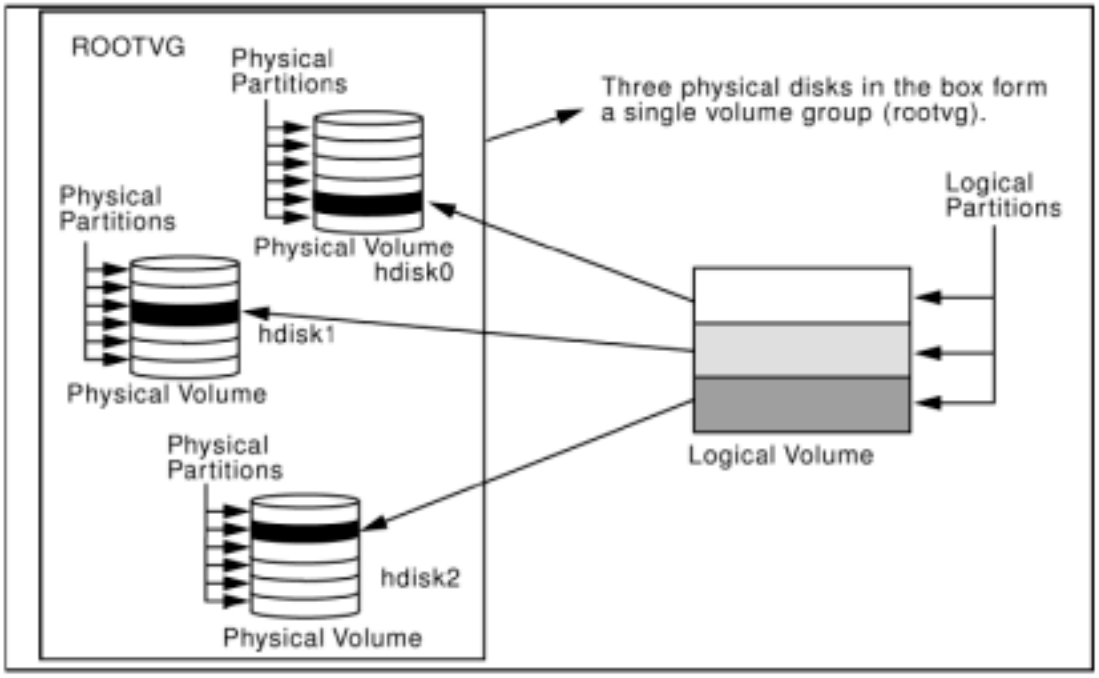

图 1 逻辑存储各组成部分之间的关系

图 1 的有关说明：

- 每个独立的物理硬盘称为一个物理卷 (PV)，附有一个名字(例如：hdisk0、hdisk1、或 hdisk2)。
- 所有的物理卷属于一个称为 rootvg 的卷组 (VG)。
- 在卷组中的所有物理卷被划分成大小一样的物理分区(PPs)。
- 在每个卷组内，定义了一个或多个的逻辑卷(LVs)。逻辑卷是位于物理卷上的信息集合。在逻辑卷上的数据对用户来说是连续的，但在实际物理卷上的分布可能是不连续的。
- 每个逻辑卷由一个或多个的逻辑分区(LPs)组成。每个逻辑分区至少对应于一个物理分区。如果逻辑卷作了镜像，那么就要分配附加的物理分区用于存储每个逻辑分区的附加拷贝。
- 逻辑卷能用于很多系统目的(例如：交换区), 但在存放普通系统数据、用户数据或程序的逻辑卷上, 包含一个日志文件系统(JFS) 。每个 JFS 由大小为 4KB 的块组成。在 AIX 4.1，每个文件系统能定义大小不到 4KB 的碎片(如 512 字节、1KB、2KB) 。

系统安装完成后，会生成一个卷组(rootvg)，该卷组由系统启动所要求的基本逻辑卷集合组成，还包括你在安装脚本中所指定的任何逻辑卷。

§2 逻辑卷管理

用于建立和控制逻辑卷存储的操作系统命令、库函数和其它工具的集合称为逻辑卷管理(LVM)。LVM 把实际的物理磁盘数据映射到一个简单而灵活的逻辑存储视图，藉以控制磁盘资源。

2.1 LVM 配置数据

用以描述 LVM 配置信息的数据不仅仅存放在一个地方。实际上，重要的卷组、逻辑卷和物理卷描述数据存放在操作系统中的若干个地方。

2.1.1 对象数据管理(ODM)数据库

AIX 版本 3 上的大多数系统配置数据均存放在 ODM 数据库中。ODM 数据库中包含所有已配置的物理卷、卷组和逻辑卷信息。这些信息是 VGDA 信息的拷贝。例如，导入一份 VGDA 的过程就包含把所导入卷组中的 VGDA 数据拷贝进 ODM。当导出一个卷组时，关于该卷组在 ODM 中所保存的信息将从 ODM 数据库中删除。

ODM 数据同样复制了在逻辑卷控制块中保存的信息。

2.1.2 卷组描述符块(VGDA)

位于每个物理卷开始部分的 VGDA 中，包含了该物理卷所在卷组的所有逻辑卷和物理卷的描述信息。几乎所有的 LVM 命令都会更新 VGDA。VGDA 让每个卷组都能够自我描述。AIX 系统能从一个磁盘上读取 VGDA，并藉以得知这个卷组是由那些物理卷和逻辑卷组成的。

每个磁盘至少包含一份 VGDA。这在激活时显得特别重要。通过读取 VGDAs 上的时间戳，能确定哪份 VGDAs 正确反映了卷组的状态。VGDAs 有可能失去同步状态，例如，在 4 个磁盘组成的一个卷组中有一个磁盘失败的时候。由于此时磁盘不可操作，所以不能更新在该磁盘上的 VGDA。因此，当磁盘重新可用时，你需要有一个方法来更新该磁盘上的 VGDA，这就是在激活过程中所做的操作。

分配磁盘作为物理卷的同时分配 VGDA (使用 mkdev 命令) 。该操作仅仅在磁盘开始部分保留一段空间用于存放 VGDA。当物理卷加入一个卷组时(使用 mkvg 或 extendvg 命令)，实际的卷组信息才放入 VGDA 中。当从卷组中删除物理卷时(使用 reducevg 命令), 同时删除 VGDA 中的卷组信息。

2.1.3 卷组状态块(VGSA)

VGSA 包含物理分区和物理卷的状态信息。例如, VGSA 知道在卷组中的一个物理卷是否不可用。

卷组描述符块和卷组状态块中有很重要的开始和结束时间戳。这些时间戳使 LVM 在激活卷组时能够认明最近的 VGDA 和 VGSA 拷贝。

LVM 要求所选择的 VGDA 和 VGSA 具有相同的时间戳。

2.1.4 逻辑卷控制块(LVCB)

LVCB 位于每个逻辑卷的开始部分。它包含逻辑卷的信息并占用数百字节。

下列例子使用 getlvcb 命令显示在逻辑卷 hd2 的 LVCB 中保存的信息：

```
# getlvcb -TA hd2
AIX LVCB
intrapolicy = c
copies = 1
interpolicy = m
lvid = 00011187ca9acd3a.7
lvname = hd2
label = /usr
machine id = 111873000
number lps = 72
relocatable = y
strict = y
type = jfs
upperbound = 32
fs = log=/dev/hd8:mount=automatic:type=bootfs:vol=/usr:free=false
time created = Tue Jul 27 13:38:45 1993
time modified = Tue Jul 27 10:58:14 1993
```

2.2 磁盘临界值

在卷组的每个物理磁盘上至少有一份 VGDA/VGSA。从在下列例子可看出，在单个磁盘上的 VGDAs 份数与其所在卷组包含磁盘的个数有关：

卷组中所含的磁盘数	VGDA 的分布
只有一块磁盘	每块磁盘上有 2 份 VGDAs
有两块磁盘	第一块磁盘上有 2 份 VGDAs，第二块磁盘上有 1 份 VGDA
有三块以上磁盘	每块磁盘上有 1 份 VGDAs

临界值是指在一个卷组中 51%或更多的物理卷可正常存取的一种状态。一个临界值相当于当前正常的卷组描述符块和卷组状态块(VGDA/VGSA)的份数占总共份数相对的大比率。一个临界值用于在磁盘失败时保证数据完整。

在一个单个磁盘上创建卷组时，在磁盘上有初始的 2 份 VGDA/VGSA。如果卷组由 2 块磁盘组成，一块磁盘仍然有 2 份 VGDA/VGSA，但另外的磁盘上只有一份 VGDA/VGSA。当卷组由 3 块或更多磁盘组成时，每个磁盘上分配一块 VGDA/VGSA。

图 2 中演示的是：由于太多的磁盘及它们的 VGDA/VGSA 不能访问，而导致可用的 VGDA/VGSA 块不再保持 51% 多数，即失去临界值。

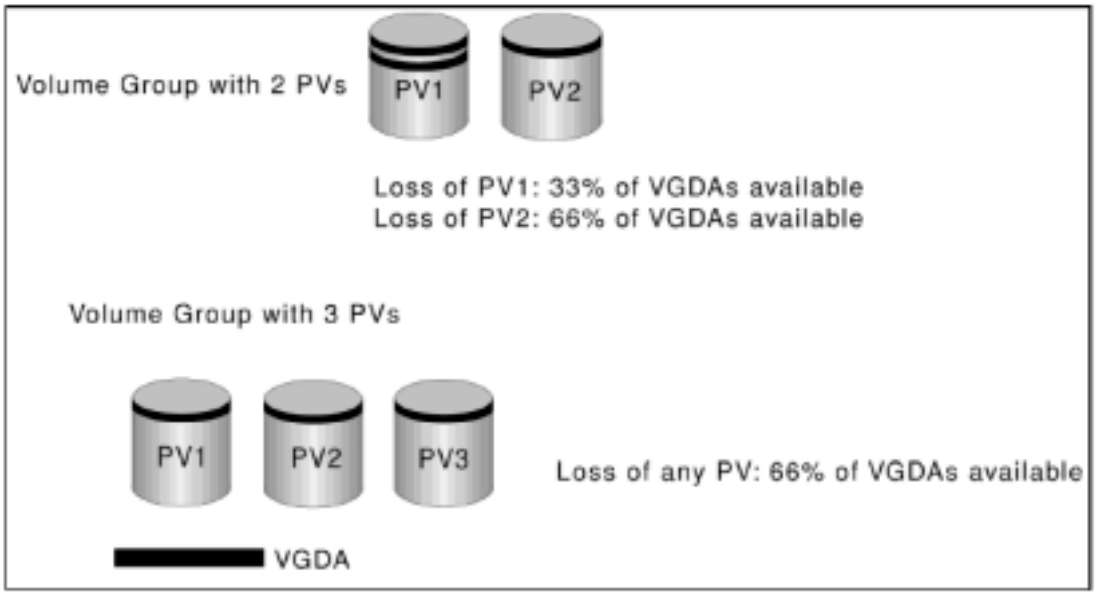

图 2：磁盘临界值

当失去临界值时，卷组把自己变为非激活状态而使逻辑卷管理(LVM)无法存取磁盘。这阻止对该卷组进一步的磁盘 I/O 以避免丢失数据，即阻止在物理故障发生时进行可能的写操作。另外, 由于变为非激活状态，错误日志中将通知用户发生了硬件故障，必须采取相应措施。

如果你想要通过镜像来保证磁盘的高可用性时，有一个隐含的注意事项。在一个采用 2 块磁盘镜像的系统中, 如果第一块磁盘失败，你将失去 66%的 VGDAs，而导致整个卷组不可用，所以无法达到镜像的目的。由于这个原因，3 个或更多个(并且通常为奇数) 磁盘单位能提供更高的可用性，并强烈推荐在需要镜像时使用这种配置。

注意：*在任何卷组上都能关闭磁盘临界值保护功能。关闭临界值保护允许一个卷组在达到临界值或多数的 VGDAs 不正常时，仍然保持激活状态。这将允许卷组在先前所描述的镜像状况下仍然工作。这种能力提供了一个较为廉价的镜像解决方案，但是确实带来数据损失的风险，尤其在一个磁盘失败以后, 数据虽然还可存取，但已没有镜像了。*

2.3 磁盘镜像

磁盘镜像是在逻辑卷内为每个逻辑分区建立同时到 2 个或 3 个物理分区的关联。在逻辑卷内写数据时，它同时写入与逻辑分区相关联的所有物理分区。因此, 数据镜像增加数据的可用性。

AIX 和逻辑卷管理提供了一系列在逻辑卷水平上建立磁盘镜像的实用工具，能在创建逻辑卷时建立镜像。

mklv 命令允许你为每个逻辑卷选择一个或二个附加的拷贝。使用 mklvcopy 命令能为一个已存在的逻辑卷增加镜像。

下列建立镜像所要考虑的因素能进一步改进数据的可用性：

- 数据拷贝份数：拥有 3 份拷贝的数据比仅仅持有 2 份拷贝更可靠。
- 拷贝的位置：一个逻辑分区的不同拷贝分配在不同的物理卷上比分配在同一物理卷上更可靠。这是因为磁盘系统最常见的错误之一就是单个物理磁盘发生故障。拷贝分布于不同的磁盘适配卡能进一步隔离故障。

2.3.1 mirrorvg 命令

mirrorvg 命令为一个给定的卷组镜像所有的逻辑卷。你也可以使用 mklvcopy 命令为一个卷组中每个逻辑卷单独建立镜像，手工地完成同样的功能。和 mklvcopy 一样，数据镜像的目标物理磁盘必须已经是卷组的成员。该命令仅适用于 4.2.1 或更高的 AIX 版本。

下列是为 mirrovg 命令的句法：

mirrorvg [- S |- s][- Q][- c Copies][- m] VolumeGroup [PhysicalVolume ..]

缺省情况下，mirrorvg 尝试把逻辑卷镜像到卷组中的任何磁盘上。mirrorvg 命令按所镜像逻辑卷的缺省设置对其进行镜像。如果你想改变镜像的严密性或创建镜像的策略，你必须手工地执行 mklvcopy 命令为所有的逻辑卷建立镜像。

注意：*由于错误检查比较复杂、卷组中做镜像的逻辑卷数目、和同步新镜像逻辑卷所花的时间等方面的原因，完成 mirrorvg 命令将花相当长的时间。*

另外， 你还能使用 SMIT 快捷命令 smitty mirrorvg，做卷组镜像。

下列例子演示 mirrorvg 命令的用法：

- 使用下列命令为一个卷组建立三份镜像：

```
mirrorvg -c 3 workvg
```

workvg 上所有逻辑卷的逻辑分区都将拥有 3 份拷贝

- 使用下列命令得到 rootvg 的缺省镜像。

mirrorvg rootvg

rootvg 卷组中的数据将有 2 份拷贝。

- 使用下列命令在做了镜像的卷组中更换一个发生故障的磁盘驱动器：

unmirrorvg workvg hdisk7
reducevg workvg hdisk7
rmdev - l hdisk7 - d

- 执行下列命令，把发生故障的磁盘更换为一块新的称为 hdisk7 磁盘：
  extendvg　workvg　hdisk7
  mirrorvg　workvg

- 下列命令将同步新创建的镜像：
  mirrorvg - S　-c 3　workvg
  -c 标志指定在 mirrorvg 命令执行完后，每逻辑卷必须有的最少拷贝数。
  -S 标志在 mirrorvg 命令后迅速返回，并在后台执行卷组的 syncvg 。镜像什么时候同步好并不确定的，但是一旦准备好系统立刻就能使用。

- 运行下列命令，创建一个卷组的精确映射：
  mirrorvg - m datavg hdisk2　hdisk3
  -m 标志允许在原有拷贝的顺序上，按精确的物理分区顺序建立逻辑卷镜像。

2.3.2 镜像 Rootvg

在 rootvg 镜像完成后，还需完成下列 3 项任务。

- 运行 bosboot 命令。

bosboot 命令依据内存 (随机存取存储器)磁盘文件系统和系统核心创建一个引导文件(引导映像)。必须执行 bosboot 命令设定最新被镜像驱动器的引导记录。•

- 运行 bootlist 命令。

bosboot 命令会保存磁盘的设备配置数据。但它不会更新 NVRAM(非易失随机存取存储器)中的引导设备表。NVRAM 表能通过 bootlist 命令进行修改。

- 重新启动系统。

最后, mirrorvg 命令缺省关闭卷组的临界值设定。关掉 rootvg 卷组的临界值设定，必须重新启动系统。

2.3.3 镜像非 rootvg 卷组

当镜像卷组时，将撤销临界值设定。为使撤消临界值生效，所有活动的逻辑卷必须关闭，然后先将卷组变为不激活状态，再重新激活卷组，才能使新的设定生效。

如果在卷组上没有施行这些步骤，尽管镜像仍将正确工作，但有关临界值新的设定将不会生效。

2.3.4 镜像 Rootvg 卷和非 rootvg 卷

系统转储设备，包括主设备（dev/hd6）和辅设备（dev/sysdumpnull），不应该镜像。在某些系统上，调页设备和转储设备是相同的。然而大部分用户想镜像调页设备。当 mirrorvg 发现转储设备和调页设备相同时，该逻辑卷将自动被镜像。

如果 mirrorvg 发现转储和调页设备是不同的逻辑卷，调页设备自动被镜像，而对转储逻辑卷不做镜像。可以用 sysdumpdev 指令查看和修改转储设备。