Sun StorEdge™ Availability Suite 3.2 Remote Mirror 软件配置指南

Sun Microsystems, Inc.
www.sun.com

部件号：817-4790-10
2003 年 12 月，修订版 A

请将有关本文档的意见或建议提交至：http://www.sun.com/hwdocs/feedback

请回收

Adobe PostScript

目录

前言

*《Sun StorEdge™ Availability Suite 3.2 Remote Mirror 软件配置指南》*提供了高效设置和使用此软件的信息。

使用 UNIX 命令

本文档可能不包括有关基本的 UNIX® 命令和过程 （如关闭系统、引导系统和配置设备）的信息。有关这类信息，请参阅以下资料：

- 系统附带的软件文档资料
- Solaris™ 操作环境文档资料，位于以下网址：

  `http://docs.sun.com`

Shell 提示符

Shell	提示符
C shell	*计算机名* %
C shell 超级用户	*计算机名* #
Bourne shell 和 Korn shell	$
Bourne shell 和 Korn shell 超级用户	#

排印约定

字体*	含义	示例
AaBbCc123	命令、文件和目录的名称；计算机屏幕上的输出	编辑您的 .login 文件。 使用 ls -a 列出所有文件。 % You have mail.
AaBbCc123	您键入的内容，与计算机屏幕输出相区别	% **su** Password:
AaBbCc123	书名、新词或术语以及要强调的词。用实际名称或值来替代命令行变量。	请阅读 *《用户指南》*的第六章。 这些称为*类*选项。 要执行该操作，您*必须*是超级用户。 要删除文件，键入 rm *文件名*。

* 您浏览器上的设置可能与这些设置不同。

相关文档资料

应用	书名	部件号
手册页	sndradm iiadm dsstat kstat svadm	无
最新发行信息	*《Sun StorEdge Availability Suite 3.2 软件发行说明》*	817-4775
	《Sun Cluster 3.0/3.1 和 Sun StorEdge Availability Suite 3.2 软件发行说明补充资料》	817-4785

应用	书名	部件号
安装和用户	*《Sun StorEdge Availability Suite 3.2 软件安装指南》*	817-4765
系统管理	*《Sun StorEdge Availability Suite 3.2 Point-in-Time Copy 软件管理和操作指南》*	817-4760
	《Sun StorEdge Availability Suite 3.2 Remote Mirror 软件管理和操作指南》	817-4770

访问 Sun 文档资料

您可以查看、打印或购买内容广泛的精选 Sun 文档资料，包括本地化版本，其网址如下：

`http://www.sun.com/documentation`

联系 Sun 技术支持

如果您遇到本文档资料无法解决的技术问题，请访问以下网址：

`http://www.sun.com/service/contacting`

Sun 欢迎您提出宝贵意见

Sun 致力于提高文档资料的质量，并十分乐意收到您的意见和建议。可以将您的意见和建议提交至以下网址：

`http://www.sun.com/hwdocs/feedback`

请在您的反馈信息中包含文档的书名和部件号：

《Sun StorEdge Availability Suite 3.2 Remote Mirror 软件配置指南》，部件号 817-4790-10

配置 Remote Mirror 软件

Sun StorEdge™ Availability Suite 3.2 Remote Mirror 软件是用于 Solaris™ 8 和 9（Update 3 及更高版本）操作系统的卷级*复制*工具。 Remote Mirror 软件在物理上独立的*主要*和*次级*站点间实时地复制磁盘卷的写操作。 Remote Mirror 软件可以与任何支持 TCP/IP 的 Sun™ 网络适配器和网络链接一起使用。

由于此软件是基于卷的，因此它的存储是独立的，并且支持 Sun 及第三方产品的原始卷或任何卷管理器。另外，此产品还支持只有一台运行 Solaris 系统的主机进行写入数据操作的任何应用程序或数据库。它不支持配置为允许多台运行 Solaris 系统的主机向共享的卷写入数据的数据库、应用程序或文件系统。（例如：Oracle 9iRAC、 Oracle Parallel Server。）

作为灾难恢复和商业持续计划的一部分， Remote Mirror 软件可以在远程站点保存重要数据的最新副本。 Remote Mirror 软件允许预演和试验商业持续计划。对于高度可用的解决方案， Sun StorEdge Availability Suite 软件可配置为在 Sun Cluster 3.x 环境中进行故障转移。

当应用程序访问数据卷、连续向远程站点复制数据或更改记录以允许以后快速重新同步时， Remote Mirror 软件是活动的。

Remote Mirror 软件既允许从主要站点到次级站点 （通常称为*正向同步*），也允许从次级站点到主要站点 （通常称为*逆向同步*）手动初始化重新同步。

Remote Mirror 软件中的复制和配置是基于集来完成的。远程镜像集包括主要站点和次级站点上的主卷、次级卷和位图卷 （用于跟踪和记录更改以进行快速重新同步），以及用于*异步复制*模式的可选的*异步队列*卷。建议将主卷和次级卷设为相同大小。您可以使用 `dsbitmap` 工具确定位图卷所需的大小。有关配置远程镜像集或 `dsbitmap` 工具的更多信息，请参阅 *《Sun StorEdge Availability Suite 3.2 Remote Mirror 软件管理和操作指南》*。

操作原理

复制既可以同步进行，也可以异步进行。在同步模式下，只有主要主机和次级主机都确认了应用程序的写操作，此写操作才得到确认。在异步模式下，只要应用程序的写操作得到本地存储的确认并写入异步队列，写操作即得到确认。此队列将写操作异步推进至次级站点。

同步复制

同步操作的数据流如下：

1. 在位图卷中设置记录位。

2. 并行初始化本地写操作和网络写操作。

3. 两项写操作完成后，清除记录位 (*lazy clear*)。

4. 写操作得到应用程序的确认。

同步复制 的优点在于主要站点和次级站点总是同步的。此复制类型只有当链接的等待时间很少，并且链接能够满足应用程序所需带宽时才实用。这些限制通常会将同步解决方案局限于校园内或大城市中。

这种情况下，一个写操作的平均服务时间为：

位图写操作 + MAX （本地数据写操作，网络传输往返时间 + 远程数据写操作）

在校园内和大城市中，网络传输往返时间可以忽略，因此服务时间大约是未安装 Remote Mirror 软件时所观测到的时间的两倍。

假设一个写操作需要 5 毫秒，则：

5 毫秒 + MAX （5 毫秒， 1 毫秒 + 5 毫秒） = 11 毫秒

注意 – 在轻负荷的系统上， 5 毫秒是一个合理的假设值。在更符合实际情况的负荷系统上，排队等待累积会使该值增大。

不过，如果网络传输往返时间达到大约 50 毫秒 （这对于远距离复制来说很平常），那么网络等待时间会使同步复制解决方案变得不切实际，如下例所示：

5 毫秒 + MAX （5 毫秒， 50 毫秒 + 5 毫秒） = 60 毫秒

异步复制

异步复制将远程写操作与应用程序写操作分开。此模式下，网络写操作在添加到异步队列时进行确认。这意味着次级站点与主要站点可能不同步，直到所有写操作均发送至次级站点。在此模式下，数据流如下：

1. 设置记录位。
2. 并行执行本地写操作和异步队列写操作。
3. 写操作得到应用程序的确认。
4. 清理线程读取异步队列项并执行网络写操作。
5. 清除记录位 (lazy clear)。

服务时间为以下操作所需的时间：

位图写操作 + MAX （本地写操作，异步队列项数据）

用 5 毫秒作为一个写操作所需的服务时间值，则异步写操作所需的服务时间大约为：

5 毫秒 + MAX （5 毫秒， 5 毫秒） = 10 毫秒

如果在较长的一段时间内，卷或*一致性组* 的写入速率超过网络排出速率，则异步队列将被填满。因此，设置适当的大小非常重要。本文档后面的章节将会讨论估计适当卷大小的方法。

以下两种模式可控制 Remote Mirror 软件在异步磁盘队列填满时的操作。

- *阻止模式*

  在阻止模式 （缺省设置）下， Remote Mirror 软件会阻止并等待异步磁盘队列排出到一定程度，然后再向异步队列添加写操作。这将影响应用程序的写操作，但是能够维护链接上写操作的顺序。

- *非阻止模式*

  在非阻止模式 （不适用于基于内存的队列）下， Remote Mirror 软件在异步磁盘队列填满时并不阻止，但会进入*记录* 模式并记录写操作。随后的*更新式同步* 将从 0 位向前读取，并且不保存写操作的顺序。如果使用这种模式，当异步磁盘队列填满而写操作顺序丢失时，则相关联的卷或一致性组不再一致。强烈建议在启动更新式同步 （例如，使用自动同步守护程序）前，在次级站点上执行即时复制操作。

一致性组

在同步模式下，涉及多卷的应用程序的写操作排序是确定的。因为在需要排序时，应用程序会等一个操作完成后才发出另一个 I/O 操作。而且只有写操作到达主要和次要站点后， Remote Mirror 软件才会发出完成信号。

在缺省的异步模式下，每个卷的队列都由一个或多个独立线程进行排出操作。由于此操作独立于应用程序，因此不会保留写入多个卷时的写操作顺序。

若应用程序需要对写操作排序，则 Remote Mirror 软件提供了一致性组功能。每个一致性组都有单一的网络队列，并且尽管允许并行执行多个写操作，写操作顺序仍可通过序列号保留下来。

计划远程复制

计划远程复制时，需要考虑商业需求、应用程序写负荷及网络特性。

商业需求

决定复制商业数据时，您需要考虑到最长延迟时间：对于次级站点上的数据，您能允许的最长过期时间是多久？这决定了复制模式和快照安排。另外，务必要了解正在复制的应用程序是否要求以正确的顺序将写操作复制到次级卷。

应用程序写负荷

了解写负荷的平均值和峰值对于决定主要站点和次级站点之间的网络连接类型十分重要。要确定配置，请收集以下信息：

- 数据写操作的平均速率和大小

  平均速率为应用程序在一般负荷情况下的数据写操作量。应用程序读操作对于准备和计划远程复制并不重要。

- 数据写操作的峰值速率和大小

  峰值速率是应用程序在一段测量持续时间内写入的最大数据量。

- 峰值写操作速率的持续时间和频率

持续时间为峰值写操作速率持续的时间长短，频率为这种情况发生的频繁程度。

如果这些应用程序特性未知，可在应用程序运行时使用工具 （例如 `iostat` 或 `sar`）测量写入流量。

网络特性

了解应用程序写负荷后就可以确定网络链接的需求。需要考虑的最重要的网络特性是网络带宽及主要站点和次级站点间的网络等待时间。如果在安装 Sun StorEdge Availability Suite 软件之前网络链接已存在，则可使用工具 （例如 `ping`）来帮助您确定站点间的链接特性。

要使用同步复制，网络等待时间必须足够低，这样应用程序响应时间便不会因每次写操作的网络传输往返时间而受到较大的影响。而且，网络带宽必须足以处理应用程序峰值写操作期间产生的写操作流量。若网络无法随时处理写操作流量，则应用程序响应时间将受到影响。

要使用异步复制，网络链接带宽必须足以处理应用程序平均值写操作期间产生的写操作流量。在应用程序峰值写操作阶段，过量的写操作将写入本地异步队列，然后在以后网络流量允许时写入次级站点。只要设置了适当的异步队列大小，在突发的写操作量超过网络限制时，仍然可以使应用程序响应时间减到最低。

请参阅此文档中第 5 页的 “配置异步队列”一节。选择的远程镜像异步选项模式（阻止模式或非阻止模式）决定了软件处理队列已满时的方式。

配置异步队列

若您使用异步复制，则本节中说明了有关配置设置的计划。这些设置基于远程镜像集或一致性组。

磁盘或内存队列

在其 3.2 版中， Remote Mirror 软件添加了对基于磁盘的异步队列的支持。为了便于从以前的版本升级，将仍支持基于内存的队列，但新的基于磁盘的队列提供了创建更大更高效队列的能力。更大的队列允许更大的突发写操作，而不会影响应用程序的响应时间。而且，基于磁盘的队列比基于内存的队列对系统资源的影响小。

异步队列的大小必须足以处理应用程序峰值写操作期间有关的突发写操作流量。大的队列能够处理长时间的突发写操作，但同时会进一步扩大次级站点和主要站点不同步的可能性。请使用峰值写操作速率、峰值写操作持续时间、写操作大小和网络链接特性来确定队列的大小。请参阅第 9 页的 “设置基于磁盘的异步队列的正确大小”。

选择的队列选项 （阻止模式或非阻止模式）决定了软件处理队列已满时的方式。请使用 dsstat 工具确定异步队列的统计信息，包括显示所使用的异步队列的最大数量的高水印 (hwm)。要将异步队列添加到远程镜像集或一致性组，请使用带 -q 选项的 sndradm 命令：sndradm -q a

队列大小

可使用 dsstat(1SCM) 命令监视异步队列以检查*高水印* (hwm)。若 hwm （用于应用程序写入了超过队列处理能力的数据而导致）经常达到队列总大小的 80% 到 85%，请增加队列大小。此原则同时适用于基于磁盘的队列和基于内存的队列。但是，重新调整不同类型队列大小的程序是不同的。

基于内存的队列

- 队列 （可调整）中写操作量的缺省最大值是 4096。可使用 sndradm -W 命令更改此值。
- 512 字节数据块 （缺省队列大小）（可调整）的缺省最大值是 16384，即 8 MB 数据。可使用 sndradm -F 命令更改此值。

基于磁盘的队列

磁盘队列的有效大小为磁盘队列卷的大小。只能通过将磁盘队列卷替换为不同大小的卷来重新调整其大小。例如，队列大小为 16384 块，请确认 hwm 未超过 13000 到 14000 块。如果超过此数量，则使用以下程序重新调整队列大小。

注意 – 磁盘队列大小的最大值为 1 TB 减去一个块的大小，即 2147483647 块。请勿使用大于最大值的卷。

▼ 重新调整队列大小

1. **将卷设置为记录模式 （使用 sndradm -l 命令）。**
2. **重新调整队列大小。**
   - 基于内存的：使用 sndradm -F 命令。

- 基于磁盘的：使用 sndradm -q 命令将现有的磁盘队列卷替换为更大的卷。

3. 使用 sndradm -u 命令执行更新式同步。

▼ 显示当前队列大小、长度和 hwm

1. 键入以下命令显示队列大小：

   - 基于内存的：

```
# sndradm -P
/dev/vx/rdsk/data_t3_dg/vol0 -> priv-2-
230:/dev/vx/rdsk/data_t3_dg/vol0
autosync: off, max q writes:4096, max q fbas: 16384, async
threads: 8, mode: async, state: replicating
```

   队列中的块数由 max q fbas 指定 （此示例中为 16384 块）。队列中项目的最大值由 max q writes 指定 （此示例中为 4096）。此示例中，这表示队列中一个项目的平均大小为 2K。

   - 基于磁盘的：

```
# sndradm -P
/dev/vx/rdsk/data_t3_dg/vol0 -> priv-
230:/dev/vx/rdsk/data_t3_dg/vol0
autosync: off, max q writes: 4096, max q fbas: 16384, async
threads: 1, mode: async, blocking diskqueue:
/dev/vx/rdsk/data_t3_dg/dq_single, state: replicating
```

   将磁盘队列卷显示 (/dev/vx/rdsk/data_t3_dg/dq_single)。可通过检查卷的大小来确定队列大小。

2. 键入以下命令显示当前队列长度及其 hwm。

```
# dsstat -m sndr -d q
name               q role     qi      qk  qhwi    qhwk
data_a5k_dg/vol0   D  net      4      13     5     118
```

   其中：

   - qi 为当前队列中的项目数
   - qk 为当前队列中的数据总大小 （以 KB 为单位）
   - qhwi 为队列中曾经出现过的最大项目数。
   - qhwi 为队列中曾经出现过的数据最大值 （以 KB 为单位）。

3. 要显示流摘要和磁盘队列信息，请键入：

```
# dsstat -m sndr -r bn -d sq 2
```

4. 要显示更多信息，请运行带其它显示选项的 dsstat(1SCM) 。

大小设置正确的队列的 dsstat *输出示例*

注意 – 此示例仅显示了与本节所需的命令输出部分；实际上 dsstat 命令可显示更多信息。

以下 dsstat(1SCM) 内核统计信息输出显示了有关异步队列的信息。在这些示例中，队列设置为正确的大小，并且队列当前未满。此示例显示以下设置和统计信息：

基于磁盘的示例

```
# dsstat -m sndr -r n -d sq -s \ priv-2-230:/dev/vx/rdsk/data_t3_dg/vol67
name              q role    qi     qk  qhwi   qhwk    kps   tps  svt
data_t3_dg/vol67  D  net    48    384   240   1944     10     1   54
```

其中：

- qi 项共有 48 个写操作已放入队列中
- qk 项表示已有 384 KB 放入队列中
- qhwi 项显示已排队项目的 hwm 为 240 个项目；当前尚未达到
- qhwk 项显示已排队数据 （以 KB 为单位）的 hwm 为 1944 ；当前尚未达到

假设磁盘队列卷大小为 1 GB （或 2097152 个磁盘块），则 1944 个块的 hwm 为远远低于全部 80% 的良好状态。磁盘队列针对写负荷的大小是正确的。

大小设置不正确的磁盘队列的 dsstat *输出示例*

以下 dsstat(1M) 内核统计信息输出显示了有关异步队列的信息，此队列的大小设置不正确。

基于内存的示例

```
# sndradm -P
/dev/vx/rdsk/data_a5k_dg/vol0 -> priv-230:/dev/vx/rdsk/data_a5k_dg/vol0
autosync: off, max q writes: 4096, max q fbas: 16384, async threads: 2, mode:
async, state: replicating

# dsstat -m sndr -d sq
name               q role     qi      qk  qhwi    qhwk     kps    tps  svt
data_a5k_dg/vol0   M  net   3609    8060  3613    8184      87     34   57
k/bitmap_dg/vol0      bmp      -       -     -       -       0      0    0
```

此示例显示了缺省的队列设置，但应用程序写入的数据超出了队列的处理能力。8184 KB 的 qhwk 值与 16384 个块 （8192 KB）的 max q fbas 之间的差异表明应用程序正逐渐接近允许的 512 字节块的最大限制。接下来的几个 I/O 操作很有可能无法进入队列。

这种情况下，增大队列是一种可行的解决方案。不过，请考虑改善网络链接 （例如使用更大带宽的接口）以满足长期效益。还可以考虑使用即时卷副本并复制影像卷。请参阅 *《Sun StorEdge Availability Suite 3.2 Point-in-Time Copy 软件管理和操作指南》*。

摘要

- 若填充速率小于或等于排出速率，则缺省的队列大小即足够。
- 若排出速率小于填充速率，则增加队列大小可提供临时的解决方案。但是，如果写操作的持续了较长的一段时间，则队列最终仍会填满。

设置基于磁盘的异步队列的正确大小

请考虑以下示例。此示例中， iostat 每小时运行一次以记录将要复制的 I/O 负荷。此示例中，假设链接为 DS3 (45MB/S)。同时假设此应用程序使用单一的一致性组，因此含有单一的队列。

收集了 24 小时数据后，假设这是所述应用程序平常状况的一天，则可以确定平均写操作速率、异步队列的适当大小、远程站点在一天过后可能会变得多过时以及选择的网络带宽是否合适此应用程序。

时间	kwr/s	wr/s	网络吞吐量	队列增长	队列大小
	A	B	C	A/1000 - C)*3600	
6am	0	0	4MB/S		
7am	1000	400	4MB/S		
8am	2000	1000	4MB/S		
9am	2000	1000	4MB/S		
10am	4000	1800	4MB/S		
11am	5000	2400	4MB/S	3.6GB	3.6GB
12pm	1000	400	4MB/S	-10GB	
1pm	1200	600	4MB/S		
2pm	1000	500	4MB/S		
3pm	1200	400	4MB/S		
4pm	2000	600	4MB/S		
5pm	1000		4MB/S		
6pm	800		4MB/S		
7pm	800		4MB/S		
8pm	3200	1000	4MB/S		
9pm	8000	2500	4MB/S	14GB	14GB
10pm	8000	2500	4MB/S	14GB	28GB
11pm	1000	400	4MB/S	-10	18
12pm	0		4MB/S	-14	4
1am	0		4MB/S	-14	
2am	0		4MB/S		
3am	0		4MB/S		
4am	0		4MB/S		
5am	0		4MB/S		
平均带宽	1.8MB/S				

填写好上表并计算队列增长和大小后，很明显 30GB 的队列已足够。尽管队列会增大，并且次级站点会因此逐渐脱离同步，但在夜间运行的批工作能够保证队列在翌日的正常工作时间之前已为空，而且两个站点同步。

此试验还证明网络带宽适合应用程序产生的写负荷。

配置异步队列清理线程

Sun StorEdge Availability Suite 3.2 软件提供了设置清理异步队列线程数的功能。更改此数值可允许网络上的每个卷或一致性组同时存在多重 I/O。次级节点上的 Remote Mirror 软件可使用序列号处理 I/O 的写操作顺序。

确定对于复制配置最有效的队列清理线程数时必须考虑许多变量。这些变量包括集或一致性组的数量、可用的系统资源、网络特性，以及是否存在文件系统。如果集或一致性组的数量较少，则较多的清理线程数可能更高效。建议您进行一些基本的测试或以稍有不同的值与此变量原型加以比较，以确定对配置最有效的设置。

配置知识、网络特性及 Remote Mirror 软件的操作可以指导您选择合适的网络线程数。 Remote Mirror 软件使用 Solaris RPC 作为传输机制：这些 RPC 是同步的。对于每个网络线程，独立的线程可达到的最大吞吐量为 I/O 大小 / 传输往返时间。考虑工作负荷超过 2k I/O，传输往返时间为 60 毫秒的情况。每个网络线程的吞吐量为：

2K/0.060S = 33K/S

在只有单个卷或包含多个卷的单个一致性组时，缺省的两个网络线程会将网络复制限制在 66 K/S。建议增加网络线程数。如果复制网络设定为 4MB/S，则理论上 2K 工作负荷的最佳网络线程数为：

(4096K/S) / (2K/0.060 IO/S) = 123

这里假设的是线性的可调节性。而实际观察表明，添加的网络线程超过 64 个后将不再受益。考虑在没有一致性组的情况下， 30 个卷在 4MB/S 链接上以 8K I/O 进行复制 。缺省的每卷 2 个网络线程会产生 60 个网络线程，如果工作负荷平均分散在这些卷上，则理论上带宽为：

60 * (8K / 0.060 IO/S) = 8MB/S

这超过了网络带宽。不需要进行调整。

异步队列清理线程数的缺省设置为 2。可使用带 `-A` 选项的 `sndradm` CLI 更改此设置。 `-A` 选项的说明为：`sndradm -A` 指定在异步模式下复制集时，可创建的用于处理异步队列的最大线程数 （缺省值为 2）。

要确定当前配置的服务于异步队列的清理线程数，可使用 sndradm -P 命令。例如，您可以看到下面的集配置了 2 个异步清理线程。

```
# sndradm -P
/dev/md/rdsk/d52 -> lh1:/dev/md/sdsdg/rdsk/d102
autosync: off, max q writes: 4096, max q fbas: 16384, async threads: 2, mode:
async, group: butch, blocking diskqueue: /dev/md/rdsk/d100, state: replicating
```

以下为有关如何使用 sndradm -A 选项将异步队列清理线程数更改为 3 的示例：

```
# sndradm -A 3 lh1:/dev/md/sdsdg/rdsk/d102
```

调整网络

Remote Mirror 软件将自身直接插入到系统的 I/O 路径中，监视所有流量，以确定其目标是否为远程镜像卷。将会跟踪目标为远程镜像卷的 I/O 命令，并管理这些写操作的副本。由于 Remote Mirror 软件直接插入到系统的 I/O 路径中，因此会对系统产生某些性能方面的影响。网络复制所需的额外 TCP/IP 处理也会消耗主机 CPU 资源。

在主要和次级远程镜像主机上执行本节所述的程序。

TCP 缓冲区大小

TCP 缓冲区 大小为传输控制协议在等待确认前允许传输的字节数。要获得最大吞吐量，请务必使用所用链接的最佳 TCP 发送和接收套接字缓冲区大小。若缓冲区太小，则 TCP 拥塞窗口将永远无法完全打开。若接收端缓冲区太大，则 TCP 流控制会中断，且发送端超过接收端，从而导致 TCP 窗口关闭。若发送主机比接收主机快，则可能发生这种情况。只要仍有多余的内存，发送端的窗口过大不会造成问题。

注意 – 在共享的网络上将缓冲区大小增加到过高的值可能会影响网络性能。请参阅 Solaris System Administrator Collection 以获得有关调整大小的信息。

表 1 显示了 100BASE-T 网络可能的最大吞吐量。

表 1 网络吞吐量和缓冲区大小

等待时间	缓冲区大小 = 24KB	缓冲区大小 = 256KB
10 毫秒	18.75 MB/ 秒	100 MB/ 秒
20 毫秒	9.38 MB/ 秒	100 MB/ 秒
50 毫秒	3.75 MB/ 秒	40 MB/ 秒
100 毫秒	1.88 MB/ 秒	20 MB/ 秒
200 毫秒	0.94 MB/ 秒	10 MB/ 秒

查看和调整 TCP 缓冲区大小

您可以使用 `/usr/bin/netstat`(1M) 和 `/usr/sbin/ndd`(1M) 命令来查看和调整 TCP 缓冲区大小。调整时需要考虑的 TCP 参数包括：

- `tcp_max_buf`
- `tcp_cwnd_max`
- `tcp_xmit_hiwat`
- `tcp_recv_hiwat`

更改其中一个参数后，请使用 `shutdown` 命令重新启动 Remote Mirror 软件，以允许软件使用新的缓冲区大小。但是关闭并重新启动服务器后， TCP 缓冲区又恢复到缺省大小。为了保存更改，需要在启动脚本中设置这些值，如本节后面的部分所述。

调整网络以查看 TCP 缓冲区和值

▼ 查看所有 TCP 缓冲区

- **键入以下命令：**

```
# /usr/sbin/ndd /dev/tcp ?| more
```

▼ 按缓冲区名称查看设置

- 此命令显示值 1073741824。

```
# /usr/sbin/ndd /dev/tcp tcp_max_buf
1073741824
```

▼ 查看套接字的缓冲区大小

- 可使用 /usr/bin/netstat(1M) 命令来查看特定网络套接字的缓冲区大小。
  例如，查看端口 121 的大小 （缺省的远程镜像端口）：

```
# netstat -na |grep "121 "
*.121 *.* 0 0 262144 0 LISTEN
192.168.112.2.1009 192.168.111.2.121 263536 0 263536 0 ESTABLISHED
192.168.112.2.121 192.168.111.2.1008 263536 0 263536 0 ESTABLISHED

# netstat -na |grep rdc
*.rdc *.* 0 0 262144 0 LISTEN
ip229.1009 ip230.rdc 263536 0 263536 0 ESTABLISHED
ip229.rdc ip230.ufsd 263536 0 263536 0 ESTABLISHED
```

此示例显示的值 263536 为 256 KB 的缓冲区大小。在主要主机和次级主机上的设置必须是相同的。

▼ 在启动脚本中设置和验证缓冲区大小

注意 – 在主要主机和次级主机上创建此脚本。

---

1. 使用以下值在文本编辑器中创建脚本文件：

```
#!/bin/sh
ndd -set /dev/tcp tcp_max_buf 16777216
ndd -set /dev/tcp tcp_cwnd_max 16777216

# increase DEFAULT tcp window size
ndd -set /dev/tcp tcp_xmit_hiwat 262144
ndd -set /dev/tcp tcp_recv_hiwat 262144
```

2. 将文件另存为 /etc/rc2.d/S68ndd，然后退出该文件。

3. 设置 /etc/rc2.d/S68ndd 文件的权限和所有权。

```
# /usr/bin/chmod 744 /etc/rc2.d/S68ndd
# /usr/bin/chown root /etc/rc2.d/S68ndd
```

4. 关闭并重新启动服务器。

```
# /usr/sbin/shutdown -y g0 -i6
```

5. 验证以前显示的大小。

Remote Mirror 使用的 TCP/IP 端口

主要节点和次级节点上的 Remote Mirror 软件会监听 /etc/services 中指定的一个公认的端口 （端口 121）。 Remote Mirror 写入通过套接字 （在主要站点上为任意指定的地址；在次级站点上为公认的地址）从主要站点到次级站点的流量。而运作状况监视 heartbeat 则在不同的连接上进行 （在主要站点上为任意指定的地址；在次级站点上为公认的地址）。 Remote Mirror 协议在这些连接上使用 SUN RPC。

图 1　　Remote Mirror 使用的 TCP 端口地址

缺省的 TCP 监听端口

端口 121 是 Remote Mirror sndrd 守护程序使用的缺省 TCP 端口。要更改端口号，请使用文本编辑器编辑 /etc/services 文件。有关更多信息，请参阅 *《Sun StorEdge Availability Suite 3.2 Remote Mirror 软件安装指南》*。

如果您更改了该端口号，则必须更改配置集内所有主机的端口号 （即，主要主机和次级主机；一对多、多对一和多重中继配置中的所有主机）。另外，您还必须关闭和重新启动所有受影响的主机，以使端口号更改生效。

使用 Remote Mirror 和防火墙

由于 RPC 需要确认，因此必须打开*防火墙*，以允许数据包的源或目的字段中有公认的端口地址。如果该选项可用，则请确保配置防火墙以允许 RPC 流量。

在写入复制流量时，目标为次级站点的数据包的目标字段包含公认的端口号，这些 RPC 的确认将在源字段包含公认的端口号。

对于运作状况监视，来自次级站点的 heartbeat 目标字段中带有公认的端口号，其确认将在源字段中包含此地址。

Remote Mirror 软件与 Point-in-Time Copy 软件

为了保证在常规操作中两个站点上最高级别的数据完整性和系统性能，建议将 Sun StorEdge Availability Suite Point-in-Time Copy 软件与 Remote Mirror 软件结合使用。

作为整体灾难恢复计划的一部分，即时副本可以复制到物理上的远程站点，提供卷的一致性副本。通常这种方式又称为批量复制，其实践过程和优点如最佳的实践指南 （*《Sun StorEdge Availability Suite Software - Improving Data Replication over a Highly Latent Link》*）中所述。

远程镜像次级卷的即时副本可在从主要站点 （主卷所在的站点）启动次级卷的同步之前建立。开始重新同步之前，可在次级站点上启用 Point-in-Time Copy 软件创建复制数据的即时副本，以防止双重故障。若在重新同步的过程中产生了并发的故障，则即时副本可用作返回位置，且在并发故障问题解决后继续进行重新同步。一旦次级站点与主要站点完全同步后，便可以禁用 Point-in-Time Copy 软件卷集，或将其用于次级站点的其它目的 （远程备份、远程数据分析或其他功能）。

在启用、复制或更新操作中内部执行的 Point-in-Time Copy 软件 I/O 操作可以更改影像卷的内容，而不使任何新的 I/O 进入 I/O 堆栈。当这种情况发生时， I/O 不会在 SV 层被打断。如果该影像卷也是远程镜像卷，则 Remote Mirror 软件将不会察觉到这些 I/O 操作。在这种情况下， I/O 操作更改的数据将不会复制到目标远程镜像卷。

为支持这种复制， Point-in-Time Copy 软件可配置为向 Remote Mirror 软件提供已更改的位图。如果 Remote Mirror 软件处于记录模式，则它会接受位图，然后将 Point-in-Time Copy 软件位图与自身中该卷的位图进行 OR 比较，并将 Point-in-Time Copy 软件位图的变化添加到自身中要复制到远程节点的变化列表中。如果

Remote Mirror 软件处于卷的复制模式，则拒绝来自 Point-in-Time Copy 软件的位图。于是，启动、复制或更新操作将失败。一旦重新启用远程镜像记录模式，便可重新进行 Point-in-Time Copy 软件操作。

注意 – 必须把远程镜像卷集置于记录模式下，以便 Point-in-Time Copy 软件在远程镜像卷上成功执行启用、复制、更新或复位操作。否则，即时复制操作将失败，Remote Mirror 软件报告操作被拒绝。

远程复制配置

Remote Mirror 软件可以创建一对多、多对一和多重中继卷集。

- 一对多复制可用于将数据从主卷复制到驻留在一台或多台主机上的多个次级卷。一个主要站点卷和每个次级站点卷分别组成一个单独的卷集。例如，对于一个主要主机卷和三个次级主机卷，您需要配置三个卷集：主要主机卷 A 和次级主机卷 B1、主要主机卷 A 和次级主机卷 B2，以及主要主机卷 A 和次级主机卷 B3。
- 多对一复制可用于通过多个网络连接、在两台以上的主机间复制卷。本软件支持将多台不同主机上的卷复制到单台主机上的卷中。此术语不同于卷到卷的一对多配置。
- 多重中继卷集是指，一个卷集的次级主机卷可以作为另一个卷集的主要主机卷。在一个主要主机卷 A 和一个次级主机卷 B 的情况下，由次级主机卷 B 充当次级主机卷 B1 的主要主机卷 A1。

Remote Mirror 软件还支持将上述几种配置结合使用。

词汇表

dsstat 一种 Sun StorEdge Availability Suite 工具，可用于显示来自远程镜像和即时快照产品的内核统计信息。

hwm 请参阅高水印

lazy clear 记录 由位图记录对磁盘的写入，而不记录每个 I/O 事件的运行日志的模式。当远程服务中断或受损时，此方法可以记录尚未复制到远程站点的已更新磁盘数据。每个源卷中不再与其远程副本匹配的块都被标记出来。本软件使用此日志通过优化的更新式同步而不是卷对卷的复制来重新构建远程镜像。

TCP 缓冲区 TCP 缓冲区大小为传输控制协议在等待确认前允许传输的字节数。

次级或远程：主机或卷 主组件的远程副本，在此对数据副本进行读写。远程副本在对等服务器上传送，主机不介入。一台服务器可同时充作某些卷的主存储器和其它卷的次级（远程) 存储器。

防火墙 一台用作两网络间接口并控制这些网络间通讯的计算机，目的是保护内部网络免受来自外部网络的电子攻击。

非阻止模式 （异步队列）在非阻止模式下，如果异步队列已满，则 Remote Mirror 软件进入记录模式并放弃队列的内容。非阻止模式不能保证发往次级站点的数据包的写操作顺序，但是能够保证异步队列已满时不会影响应用程序响应时间。

复制 完成卷集的初始同步之后，该软件将确保主卷和次级卷始终含有相同的数据。复制是由用户层应用程序的写操作所触发的；复制是一个持续的进程。

高水印 高水印是指所使用的异步队列最大容量。

更新式同步 更新式同步只复制那些由记录模式标记的磁盘块，减少了恢复远程镜像集的时间。

卷集文件 一个含有特定卷集相关信息的文本文件。此文本文件与配置位置不同，配置位置包含所有 Remote Mirror 和 Point-in-Time Copy 软件用到的已配置卷集的有关信息。

逆向同步 恢复预演时所用的操作。日志记录预演过程中次级系统上的测试更新。当主系统恢复时，测试更新将被主系统映像的数据块覆盖，从而恢复匹配的远程数据集。

配置位置 存储 Sun StorEdge Availability Suite 软件用到的关于所有启用的卷的配置信息的位置。

同步 在目标磁盘上建立源磁盘的副本的过程，是软件镜像的前提条件。

同步复制 由于 I/O 响应时间的传输延迟的不利影响，同步复制只限于较短的距离 （几十公里）。

一致性组 一致性组是指共享同一个异步队列以维护写操作顺序的一组远程卷。

异步队列 用于存储要复制到远程站点的写操作的本地磁盘或内存。写操作进入队列后由应用程序确认，然后在网络性能允许的情况下稍后转发到远程站点。

异步复制 异步复制是在远程映像更新前即向源主机确认主 I/O 事务已完成。即，当本地写操作已结束并且远程写操作已进入队列时，才向主机确认 I/O 事务的完成。推迟次级副本除去了 I/O 响应时间中长距离传播的延时。

整卷式同步 整卷式同步执行卷对卷的完全复制，是最耗时的同步操作。大部分情况下，是使用主卷对次级卷进行同步。然而，在恢复出现故障的主磁盘时，可能需要使用幸存的远程镜像来执行逆向同步。

正向重新同步 请参阅更新式同步。

主要或本地：主机或卷 主机应用程序主要依赖的系统或卷。例如，产品数据库由此获取访问数据。本软件将把此数据复制到次级。

自动同步 在主要主机上启用自动同步选项后，如果系统重新引导或者发生链接失败，则同步守护程序 (autosyncd) 会试图重新同步卷集。

阻止模式 （异步队列）在阻止模式下，如果异步队列已满，则之后的写操作都将延迟，直到队列腾出足够的空间允许进行写操作。中断模式是缺省的异步运行选项，能够保证发送到次级站点的数据包的写操作顺序。设置中断选项后，如果异步队列已满，则应用程序响应时间可能会受影响。