

Tru64 UNIX Best Practice

Aligning LSM Disks and Volumes to Hardware RAID Devices

December 2001

Product Version: **Tru64 UNIX Version 5.0 or higher**

This Best Practice describes how to align LSM disks and volumes to hardware RAID devices for the Tru64 UNIX operating system.

Contents

Aligning LSM Disks and Volumes to Hardware RAID Devices

| | |
|--------------------------------------------------------|----|
| Is This Best Practice Right for You? | 1 |
| Before You Begin | 2 |
| Applying the Best Practice | 5 |
| Creating a Hardware Stripeset | 5 |
| Making the System Aware of the New Device | 7 |
| Initializing a Stripeset as an LSM Disk | 7 |
| Creating a Sliced Disk | 8 |
| Creating a Simple Disk | 8 |
| Creating an LSM Volume with an Aligned Stripe Width .. | 9 |
| Verifying Success | 10 |
| Displaying Disk Properties | 11 |
| Displaying Volume Properties | 11 |
| Troubleshooting | 12 |
| Comments and Questions | 13 |
| Legal Notice | 13 |

Aligning LSM Disks and Volumes to Hardware RAID Devices

This Best Practice is intended for administrators who have configured hardware stripesets or RAIDsets with a specified I/O chunk size and who want to configure those hardware devices as Logical Storage Manager (LSM) disks and volumes, ensuring alignment to the underlying hardware chunk size.

Use the commands described in your hardware documentation to create stripesets or RAIDsets with a specified chunk size. This Best Practice shows the commands for an HSG60 or HSG80 Array Controller and describes how to initialize the device as an LSM disk and then use it to create an LSM volume to maximize the I/O throughput from the volume to the LSM disk to the hardware device.

See the Tru64 UNIX Best Practices Web page for more information about Best Practices documentation:

http://www.tru64unix.compaq.com/docs/best_practices/

Is This Best Practice Right for You?

Not all Best Practices apply to all configurations, so you must be sure that this Best Practice is appropriate for your system and circumstances. To use this Best Practice, you must meet the requirements described in the following table:

| Requirement | Description |
|----------------------|----------------------------------------------------------------------|
| Operating System | Tru64 UNIX Version 5.0 or higher, with LSM installed and initialized |
| System Configuration | Disks connected to a RAID Array Controller such as an HSG60 or HSG80 |

| Requirement | Description |
|------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Impact on Availability | None. No reboot required to configure LSM disks or volumes |
| Knowledge or Skills | <p>Knowledge of your application's I/O size and your site's performance requirements</p> <p>Knowledge of how to configure hardware devices for your storage solution</p> <p>Knowledge of how your hardware device names map to disk names recognized by the operating system (LUNs)</p> |

It is generally not recommended to use software striping on top of hardware striping (for example, creating an LSM striped volume on disks that are themselves hardware stripesets). This tends to diminish performance as each I/O request is broken into inefficiently small units and can involve more drives than intended or necessary.

It is also not recommended to use software RAID 5 (striping with parity) on top of hardware RAIDsets, for similar reasons. In particular, it is not recommended to create multiple back-end units and then use LSM to concatenate, mirror, or stripe them, if they all belong to the same RAID array. However, you can use LSM to manage the devices, to mirror across multiple RAID arrays (for high availability), and to stripe efficiently across multiple controllers if an application has a large I/O request size.

If you create hardware stripesets or RAIDsets, you can use each set to create one LSM disk.

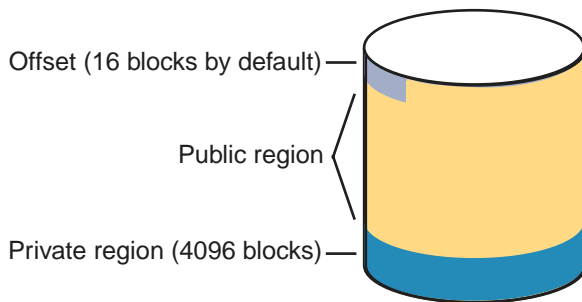
Before You Begin

Before you apply this Best Practice for aligning LSM disks and volume stripe widths to the chunk size of the underlying hardware devices, you must understand some background information.

- An LSM sliced disk with the default settings (Figure 1) consists of a public region, for storage allocation, at the beginning of the disk (typically the `g` partition) and a private region, for internal metadata, at the end (typically the `h` partition).

The private region is 4096 blocks by default, and the public region is the remainder of the disk. However, LSM skips over the first 16 blocks of the `g` partition to preserve the disk label and, if applicable, the bootstrap information.

Figure 1: LSM Sliced Disk with Default Offset (16 Blocks)



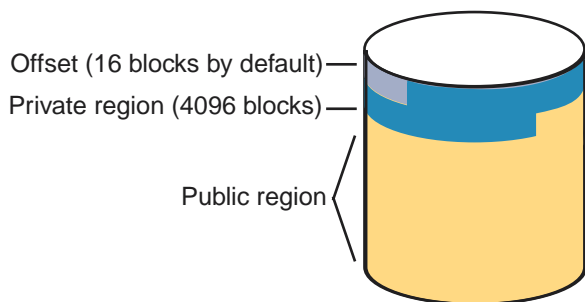
Note: This drawing is not to scale.

ZK-1863U-AI

- An LSM simple disk, created by specifying a disk partition, also consists of a public and private region but their order is reversed (Figure 2).

The private region uses the first 4096 blocks of the given partition and the public region uses the remainder. (However, if the partition specified is either `a` or `c` — a partition that starts at the beginning of the disk, LSM skips over the first 16 blocks to preserve the disk label and, if applicable, the bootstrap information.)

Figure 2: LSM Simple Disk with Default Offset (16 Blocks)



Note: This drawing is not to scale.

ZK-1864U-AI

For sliced disks, you specify an offset for the start of the public region, and for simple disks, you specify an offset for the private region, which has the effect of also offsetting the public region. By specifying an offset, you

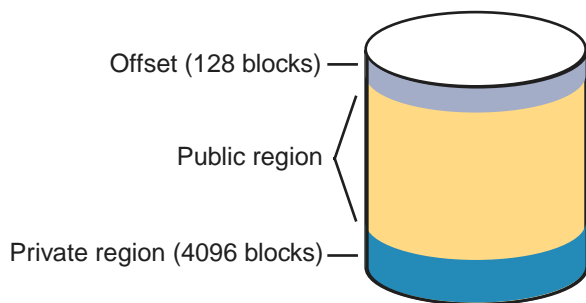
can force LSM to begin allocating space at a point that aligns with the beginning of a chunk at the hardware level.

Consider the following example: You have a hardware stripeset with a chunk size of 128 blocks. You initialize that stripeset as an LSM sliced disk with the default settings (no offset). Because LSM skips the first 16 blocks of the disk, an I/O request of 128 blocks actually begins at block 16 and extends to block 144 of the disk. If each stripe in the hardware stripeset resides on a different physical disk, or on noncontiguous regions of the same disk, this one I/O request results in two reads or writes.

However if you initialize the LSM disk with an offset that is a multiple of the chunk size (128 blocks in this example), an I/O request to the LSM disk will always align to the chunk size on the underlying hardware device.

Figure 3 shows an LSM sliced disk with a public region offset of 128 blocks. Figure 4 shows an LSM simple disk with a private region offset of 128 blocks.

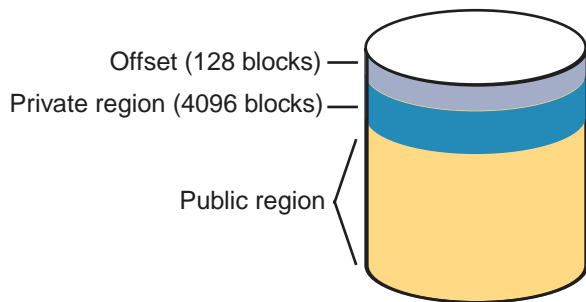
Figure 3: LSM Sliced Disk with 128 Block Offset



Note: This drawing is not to scale.

ZK-1861U-AI

Figure 4: LSM Simple Disk with 128 Block Offset



Note: This drawing is not to scale.

ZK-1862U-AI

Applying the Best Practice

There are four main steps to align your LSM disks and volume stripe widths to the hardware chunk size:

1. Create the hardware stripesets with the desired chunk size.
2. Make the system aware of the devices.
3. Initialize the hardware stripesets as LSM disks with an offset that aligns to the chunk size.
4. Create an LSM volume from the disks, specifying a stripe width that is equal to, or a multiple of, the chunk size.

Creating a Hardware Stripese

Be sure to follow the storage planning guidelines in the documentation for your controller.

To create a stripeset with a chunk size of 128 blocks on an HSG80 controller:

1. Create the stripeset:

```
HSG80> ADD STRIPESET SET_NAME DISKnnnnnnn DISKnnnnnnn ...
```

For example:

```
HSG80> ADD STRIPESET STRIPE1 DISK10100 DISK20100
```

2. Initialize the stripeset and specify the chunk size:

```
HSG80> INITIALIZE SET_NAME CHUNKSIZE=128
```

For example:

```
HSG80> INITIALIZE STRIPE1 CHUNKSIZE=128
```

3. Optionally, confirm the chunksize. For example:

```
HSG80> SHOW STRIPE1
```

| Name | Storageset | Uses | Used by |
|-----------------------------------------------------------|------------|------------------------|---------|
| STRIPE1 | stripeset | DISK10100 DISK20100 | |
| Switches: | | | |
| CHUNKSIZE = 128 blocks | | | |
| State: | | | |
| UNKNOWN -- State only available when configured as a unit | | | |
| Size: 35538354 blocks | | | |

4. Assign the stripeset a unit number to make it accessible to the host:

```
HSG80> ADD UNIT UNIT_NUMBER SET_NAME
```

For example:

```
HSG80> ADD UNIT D10 STRIPE1
```

5. Assign an identifier to the unit, to help you associate the unit with its future disk name. The identifier must be a number (1 through 9999) and cannot already be assigned to another device:

```
HSG80> SET D10 IDENTIFIER=0110
```

6. Optionally, confirm the settings. For example:

```
HSG80> SHOW D10
```

| LUN | Uses | Used by |
|-------------------------------------------------|-----------------|------------|
| D10 | STRIPE1 | |
| LUN ID: 6000-1FE1-0006-3F10-0009-0270-0662-0031 | | |
| IDENTIFIER = 0110 | | |
| Switches: | | |
| RUN | NOWRITE_PROTECT | READ_CACHE |
| READAHEAD_CACHE | WRITEBACK_CACHE | |
| MAXIMUM_CACHED_TRANSFER_SIZE = 32 | | |
| Access: | | |
| ALL | | |
| State: | | |
| ONLINE to this controller | | |
| Not reserved | | |
| NOPREFERRED_PATH | | |
| Size: 35538354 blocks | | |
| Geometry (C/H/S): (10515 / 20 / 169) | | |

7. Configure any additional storage devices.

As you configure your hardware devices, make note of the chunk size you specify for each set and the set's name and identifier.

Making the System Aware of the New Device

To make the operating system aware of the new device:

1. At the Tru64 UNIX prompt, create device special files for the new device.

- On a standalone system, enter:

```
# hwmgr -scan scsi
```

- On a cluster, enter:

```
# hwmgr scan component -category scsi_bus -cluster
```

2. Find the new disk number, using the identifier you assigned the device to restrict the search (in this example, 0110):

```
# hwmgr -view devices | grep IDENTIFIER=0110
126: /dev/disk/dsk12c DEC HSG80 IDENTIFIER=0110
```

The system now recognizes the new unit as *dskn* (in this example, *dsk12c*).

3. To ensure that the disk label reflects the new size and layout of the composite device, delete the old disk label and apply a default disk label:

```
# disklabel -z dskn
# disklabel -rwn dskn
```

Initializing a Stripeset as an LSM Disk

Depending on your configuration needs, you can initialize the hardware devices as either sliced or simple LSM disks.

- For a sliced disk, specify an offset for the public region.
- For a simple disk, specify an offset for the private region.

Creating a Sliced Disk

To initialize a hardware device as a sliced disk with an offset for the public region:

1. Set up the disk with a default LSM disk label for the public and private regions (do not include a partition letter):

```
# voldisksetup dskn
```

2. Initialize the disk and apply an offset to the public region that is equal to or a multiple of the hardware chunk size:

```
# voldisk init dskn puboffset=value
```

For example:

```
# voldisk init dsk12 puboffset=128
```

3. Add the disk to the desired disk group (the default is rootdg) or create a new disk group, depending on your configuration requirements.

- To add the disk to an existing disk group, enter:

```
# voldg [-g diskgroup] adddisk dskn
```

For example:

```
# voldg adddisk dsk12
```

- To create a new disk group, enter:

```
# voldg init new_diskgroup dskn ...
```

For example:

```
# voldg init dg01 dsk12
```

Creating a Simple Disk

For a simple disk, you specify an offset for the private region. This consequently moves the beginning of the public region to the block number equal to the size of the private region (4096 by default) plus the offset you specified.

Make sure that the total of the offset plus the private region size is a multiple of the hardware chunk size.

The following table shows where the public region will begin based on different offsets and a default private region size of 4096 blocks.

| Private Region Offset in Blocks | Start of Public Region (Block Number) |
|---------------------------------------|---------------------------------------|
| 0 (for a partition other than a or c) | 4096 |
| 16 (default for the a or c partition) | 4112 |
| 128 (user-specified) | 4224 |

To initialize a hardware device as a simple disk with an offset for the private region:

1. Initialize the disk partition and apply an offset to the private region that is equal to or a multiple of the hardware chunk size:

```
# voldisk init dsknp privoffset=value
```

For example:

```
# voldisk init dsk10d privoffset=128
```

2. Add the disk to the desired disk group (the default is rootdg) or create a new disk group, depending on your configuration requirements.

- To add the disk to an existing disk group, enter:

```
# voldg [-g diskgroup] adddisk dskn
```

For example:

```
# voldg adddisk dsk10d
```

- To create a new disk group, enter:

```
# voldg init new_diskgroup dskn ...
```

For example:

```
# voldg init dg01 dsk10d
```

Creating an LSM Volume with an Aligned Stripe Width

After you create the underlying hardware devices with the specified chunk size and turn those devices into LSM disks with offsets that align to the chunk size, you can create a striped volume whose stripe width is also aligned to the LSM disks.

For the example that follows, assume you created six new LSM sliced disks with a `puboffset` of 128 blocks named `dsk12`, `dsk13`, `dsk14`, `dsk15`, `dsk16`, and `dsk17`.

To create an LSM striped volume and align the stripe width to the chunk size, enter the following command line, where `stwid` is equal to or a multiple of the hardware chunk size, and `ncolumn` is the number of disks to stripe across:

```
# volassist make volume_name length layout=stripe stwid=width \  
ncolumn=num_disks dskn ...
```

For example, to create a 10 MB volume named `volA` with a stripe width of 128 blocks on the six disks mentioned previously, enter:

```
# volassist make volA 10m layout=stripe stwid=128 ncolumn=6 \  
dsk12 dsk13 dsk14 dsk15 dsk16 dsk17
```

Note

To create a mirrored and striped volume on the same number of disks, halve the number of columns and specify no DRL log, to prevent LSM from creating the log on one of the disks that will contain data.

Add the log in a separate step, specifying a different LSM disk. The log disk does not have to be one for which you configured a hardware chunk size.

The commands would look like the following:

```
# volassist make volA 10m layout=stripe,nolog \  
stwid=128 ncolumn=3 nmirror=2 dsk12 dsk13 dsk14 dsk15 \  
dsk16 dsk17  
  
# volassist addlog volA dsk20
```

Each mirror will be striped across three of the six disks with a chunk size, and the log will be on a different LSM disk.

Verifying Success

After you apply this Best Practice, you can verify whether it was successful by displaying the properties of the LSM disks you created from the hardware devices and the layout of the volumes using those disks.

Additionally, if you have performance statistics from an application using an LSM volume that was not aligned to the hardware chunk size, or if you know what the application's theoretical throughput is, you can compare the performance of the volume or application after applying this Best Practice and confirm that the performance is improved.

If the Best Practice was not successful, see *Troubleshooting* for information about identifying and solving problems.

Displaying Disk Properties

To display the disk properties, enter:

```
# voldisk list dskn
```

For example:

```
# voldisk list dsk12
```

Information similar to the following is displayed. Confirm that the public region offset (for a sliced disk) or private region offset (for a simple disk) is what you specified (indicated in bold):

```
Device:      dsk12
devicetag:   dsk12
type:        sliced
hostid:      hostname.com
disk:        name=dsk12 id=1002921154.2836.hostname.com
group:       name=rootdg id=992637777.1026.hostname.com
flags:       online ready autoimport imported
pubpaths:    block=/dev/disk/dsk12g char=/dev/rdisk/dsk12g
privpaths:   block=/dev/disk/dsk12h char=/dev/rdisk/dsk12h
version:     2.1
iosize:      min=512 (bytes) max=2048 (blocks)
public:     slice=6 offset=128 len=4106256
private:    slice=7 offset=0 len=4096
update:      time=1002917000 seqno=0.3
headers:     0 248
configs:     count=1 len=2993
logs:        count=1 len=453
Defined regions:
config  priv  17-   247[  231]: copy=01 offset=000000 disabled
config  priv  249- 3010[ 2762]: copy=01 offset=000231 disabled
log     priv  3011- 3463[  453]: copy=01 offset=000000 disabled
```

Displaying Volume Properties

To display the volume properties, enter:

```
# volprint -vht volume
```

For example:

```
# volprint -vht volA
```

For a striped volume, information similar to the following is displayed, showing that the volume is striped over six columns of disks and has a stripe width of 128 blocks (indicated in bold):

Disk group: rootdg

| V NAME | USETYPE | KSTATE | STATE | LENGTH | READPOL | PREFPLEX | | |
|-------------|---------|---------|----------|--------|------------|----------|------|--|
| PL NAME | VOLUME | KSTATE | STATE | LENGTH | LAYOUT | NCOL/WID | MODE | |
| SD NAME | PLEX | DISK | DISKOFFS | LENGTH | [COL/]OFF | DEVICE | MODE | |
| v volA | fsgen | ENABLED | ACTIVE | 20480 | SELECT | volA-01 | | |
| pl volA-01 | volA | ENABLED | ACTIVE | 20736 | STRIPE | 6/128 | RW | |
| sd dsk12-01 | volA-01 | dsk12 | 0 | 3456 | 0/0 | dsk12 | ENA | |
| sd dsk13-01 | volA-01 | dsk13 | 0 | 3456 | 1/0 | dsk13 | ENA | |
| sd dsk14-01 | volA-01 | dsk14 | 0 | 3456 | 2/0 | dsk14 | ENA | |
| sd dsk15-01 | volA-01 | dsk15 | 0 | 3456 | 3/0 | dsk15 | ENA | |
| sd dsk16-01 | volA-01 | dsk16 | 0 | 3456 | 4/0 | dsk16 | ENA | |
| sd dsk17-01 | volA-01 | dsk17 | 0 | 3456 | 5/0 | dsk17 | ENA | |

For a striped and mirrored volume, information similar to the following is displayed, showing that each plex of the volume is striped over three columns of disks and has a stripe width of 128 blocks:

Disk group: rootdg

| V NAME | USETYPE | KSTATE | STATE | LENGTH | READPOL | PREFPLEX | | |
|-------------|---------|---------|----------|---------|------------|----------|------|--|
| PL NAME | VOLUME | KSTATE | STATE | LENGTH | LAYOUT | NCOL/WID | MODE | |
| SD NAME | PLEX | DISK | DISKOFFS | LENGTH | [COL/]OFF | DEVICE | MODE | |
| v volA | fsgen | ENABLED | ACTIVE | 20480 | SELECT | - | | |
| pl volA-01 | volA | ENABLED | ACTIVE | 20736 | STRIPE | 3/128 | RW | |
| sd dsk12-01 | volA-01 | dsk12 | 65 | 6912 | 0/0 | dsk12 | ENA | |
| sd dsk13-01 | volA-01 | dsk13 | 0 | 6912 | 1/0 | dsk13 | ENA | |
| sd dsk14-01 | volA-01 | dsk14 | 0 | 6912 | 2/0 | dsk14 | ENA | |
| pl volA-02 | volA | ENABLED | ACTIVE | 20736 | STRIPE | 3/128 | RW | |
| sd dsk15-01 | volA-02 | dsk15 | 0 | 6912 | 0/0 | dsk15 | ENA | |
| sd dsk16-01 | volA-02 | dsk16 | 0 | 6912 | 1/0 | dsk16 | ENA | |
| sd dsk17-01 | volA-02 | dsk17 | 0 | 6912 | 2/0 | dsk17 | ENA | |
| pl volA-03 | volA | ENABLED | ACTIVE | LOGONLY | CONCAT | - | RW | |
| sd dsk20-01 | volA-03 | dsk20 | 0 | 65 | LOG | dsk20 | ENA | |

Troubleshooting

If you determine that this Best Practice was not successful, as described in *Verifying Success*, use the following table to identify and solve problems:

| Problem | Possible Solutions |
|----------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| puboffset or privoffset is not what you specified. | Ensure you specified the appropriate offset for the type of disk you initialized (puboffset for sliced disks, privoffset for simple disks). |
| Performance is not what you expected. | Confirm the chunk size of the hardware device, and confirm that the LSM offsets are equal to or multiples of that value. |

Comments and Questions

We value your comments and questions on the information in this document. Please mail your comments to us at this E-mail address:

`best_practices@zk3.dec.com`

Legal Notice

Compaq, the Compaq logo, and Tru64 are trademarks of Compaq Information Technologies Group, L.P. in the U.S. and/or other countries. UNIX is a trademark of The Open Group in the U.S. and/or other countries. All other product names mentioned herein may be trademarks of their respective companies.

Confidential computer software. Valid license from Compaq required for possession, use, or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Compaq shall not be liable for technical or editorial errors or omissions contained herein. The information is provided "as is" without warranty of any kind and is subject to change without notice. The warranties for Compaq products are set forth in the express limited warranty statements accompanying such products. Nothing herein should be construed as constituting an additional warranty.