Tunables for the Sun StorEdge™ SAN Foundation Suite: Optimizing for Performance and Failover

Chris Wood, Sajid Zia, and Dennis Kleppen

June 2007

Sun Microsystems, Inc.

Abstract: *This article describes various tunables for Sun StorEdge™ SAN Foundation Suite (also known as Leadville) software, specifically for failover optimization and performance.*

Table of Contents

Introduction

There are several areas of performance – failover times, SCSI queue depth management, and so on – that can be optimized by judicious and careful use of certain user-tunable kernel-level variables described below. ***Because many of these variables are interdependent with other variables and are global in nature, please approach any changes to them with extreme caution.***

Sun's SAN engineering philosophy is to provide a tested, optimized stack that operates with **all** Sun-sold storage devices in a seamless, interoperable manner, without requiring any user configuration of the driver stack. This is very different from other vendors, who often **require** that you edit their driver stacks to make them work. That stated, there may be circumstances where changing these values to optimize failover times or array performance for **specific** configurations and/or applications may be desirable.

This paper presents three (3) general matrices describing the various types of tunables and/or default values related to performance, failover times and error recovery and logging. Type one (1) variables may be adjusted to fit specific customer configuration requirements. Type two (2) variables are adjustable, but in general no adjustment is recommended: these variables need to be tuned only in very rare cases. Type three (3) variables are hard-coded and are not adjustable; they are presented herein for completeness only.

Additionally, two additional, device-specific matrices are supplied that cover the Sun StorEdge T3/6x20, the Sun StorEdge 99x0 and the Sun StorEdge 3510 systems. There are, in some cases, different, device-specific Type 1 and Type 2 values used in the driver stack based on the driver issuing a SCSI Inquiry command and identifying a specific type of device. These are identified in the device-specific matrix section.

Matrix Type 1 (User-Adjustable)

Tunable	Description	Default (Value)	Practical Ranges	Dependencies (see "Dependencies and Interrelationships" section)
`set ssd:ssd_io_time=nn` Use: Improve failover times	Sets the command timeout value for SCSI commands.	60 (seconds)	30-120	1, 3
`set ssd:ssd_ua_retry_count=nn` Use: Trigger potential failover faster versus retrying the operation on the same port.	Sets the retry count for SCSI commands, typically after a timeout or when requested by the target device.	3 (retries)	2-5	2, 3 See Note 1
`set ssd:ssd_max_throttle=nn` Use: Limit command queue depths to keep one server from dominating a device and/or limit the time required to flush commands during a failover operation.	Sets the maximum SCSI command queue depth that is issued to any given target device.	(See following table.)	8-256	3

Tunable	Description	Default (Value)	Practical Ranges	Dependencies (see "Dependencies and Interrelationships" section)
`set fp:fp_offline_ticker=nn` Use: Detect a dead port faster and trigger upper-level error recovery.	Controls the number of seconds that the HBA waits before it off-lines a port due to loss of "light."	90 (seconds)	5-120	4, 5
`set fp:fp_retry_count=nn` Use: Limit the number of frame retries on the same port, thus driving upper-level error recovery faster.	Controls the retry count at the FCP (Frame) level.	5 (retries)	2-5	4, 5

Note 1: On a global basis, there are certain SCSI retry-related values for certain specific device responses versus a general timeout:

- Device or bus reset directed to a different device on the fabric: *`ssd_retry_count * 2 (6)`* retries to the target device.
- Delay between retries: 100 msec. in most cases. A totally unresponsive device incurs the `ssd_io_time` (60 seconds) timeout per retry.
- Retries issued before the initiator (Leadville) issues a bus or device reset: `ssd_retry_count/2` (1 or 2).

The default values shown above have been developed over the past several years to address generic customer requirements. They represent a tested set of parameter values which has been shown to work in many large installations. Before implementing any changes to the default values in a production system, you should verify that the new values work correctly in your customer's configuration first.

These variables can be adjusted by setting them in the `/etc/system` file.

Dependencies and Interrelationships

1. Non-Disruptive Firmware Upgrades: Certain arrays require up to 120 seconds to perform firmware (micro code) upgrades. Sun StorEdge 99x0 systems require 60 seconds. Sun StorEdge 6xxx and T3 systems require up to 90 seconds. Setting `ssd_io_time *(ssd_retry_count +1)` to a value lower than these could result in failed I/O (after all retries are exhausted) in *less* time than it takes for the firmware upgrade to complete. This either results in failed IO, or in the case where an alternate path is available, path failover.

2. SCSI Retry Count: Retries are driven either by hitting the `ssd_io_time` or when a target device hits its queue depth and returns queue full or busy. By carefully limiting the `ssd_max_throttle` value, the device queuing limit is hit more rarely (or never). Setting the retry value to zero should not be done since that cripples the ability of the driver to recover from occasional glitches in the system.

3. SCSI Queue depth: Arrays have hard limits on the SCSI queue depth they support; these limits should not be exceeded. The Sun StorEdge 9910 and 9960 systems, for instance, set a queue depth of 32 per LUN and a maximum of 256 per port, but in newer Sun StorEdge ST9990 and Sun StorEdge ST9990V systems, the maximum is 1024 per port. Exceeding the queue depth supported drives retry recovery scenarios that hurt performance. Older arrays may not handle this situation

properly and may have to be rebooted. Measurements have shown that values larger than 32 have no measurable performance impact. Values lower than 8 have a significant impact on performance. Reducing `ssd_io_time` should usually be accompanied by a reduction in `ssd_max_throttle`. An `ssd_max_throttle` value of 16 seems to be a nice sweet spot that balances performance against failover times.

4. FP Level Tunables: `fp_offline_ticker` and `fp_retry_count` work at the FCP frame level and are designed to either retry a frame transmission, or trigger upper level error recovery if the HBA detects a loss of "light" on the link. `fp_retry_count` must be at least two, and preferably higher in more complex SANs that have a higher chance of having to retransmit a frame due to having many links. `fp_offline_ticker` should be set to at least five seconds to avoid reporting an error when sensing a temporary loss of light on a link. Certain switches, when under going reconfiguration (for example, re-zoning, and so on), lose light on a port during the operation. Most 2-Gbit switches are very fast and low values are not an issue. Older 1-Gbit switches take a long time to reconfigure ports. The default values should not be changed.

5. Arbitrated Loop (FC-AL) Special Case: `fp_offline_ticker` and `fp_retry_count` should **not** be changed (reduced) in SANs that are configured to run in arbitrated loop mode – both public and private. Due to long retraining times, "LIP Storms" and other anomalies specific to loop configurations, a long time is typically needed to allow the loop to stabilize. Lowering these settings arbitrarily causes unpredictable and inappropriate off-lining and/or failovers. Devices where this tends to be an issue are Sun StorEdge 6x20 and Sun StorEdge 3510 FC systems configured to operate in loop mode. The **default** operating mode of the Sun StorEdge 3510 Array is FC-AL, so special care should be taken with Sun StorEdge 3510-based SANs to verify the operating mode before making any changes.

Approximate Time to I/O Failure

Approximate expected behavior at the target driver level in terms of an I/O on a given path failing with the above tunable settings is shown in the table below. Each example assumes that the error condition persists through all the I/O retries. Note that the I/O may ultimately succeed if MPxIO can re-route the I/O down a different path successfully, or if a volume manager above the target driver has an alternate source to use to get the data (for example, mirror).

The following values are currently calculated values. Specific testing is planned to validate these calculations and/or modify them as appropriate.

Error Condition	~Time to I/O Failure
Device not responding to I/O	240 seconds
Device returning busy status	20 seconds
Device returning busy status (Sun StorEdge T3, SESS01, Sun StorEdge SE6920, PSX1000)	305 seconds See Note 1
Frame level transmit error	18 seconds
Loss of light on FC link	90 seconds

Note 1: These products have special handling in the driver stack. Their values are calculated based upon determination of the specific device type by the driver stack.

Generic Example

A tested, generic set of values designed to minimize failover with **all** current Sun arrays is:

```
set ssd:ssd_io_time=30
set ssd:ssd_retry_count=2
set ssd:ssd_max_throttle=16
set fp:fp_offline_ticker=5
set fp:fp_retry_count=2
```

This set of tunables would be appropriate for cases where there is more than one type of array in a SAN. Note that the `ssd_io_time` parameter disables the ability of the Sun StorEdge 99xx and 6xxx systems to perform transparent firmware upgrades. `ssd_max_throttle` has been limited to 16 to make sure that there are not a large number of outstanding commands when the failover takes place. As a last word, **test these configurations settings first** in your customer's environment before going into production.

Approximate Time to I/O Failure

Approximate expected behavior at the target driver level in terms of an I/O on a given path failing with the above tunable settings is in the table below. Each example assumes that the error condition persists through all the I/O retries. Note that the I/O may ultimately succeed if MPxIO can re-route the I/O down a different path successfully, or if a volume manager above the target driver has an alternate source to use to get the data (for example, mirror).

The following values are currently calculated values. Specific testing is planned to validate these calculations and/or modify them as appropriate.

Error Condition	~Time to I/O Failure
Device not responding to I/O	90 seconds
Device returning busy status	15 seconds
Device returning busy status (Sun StorEdge T3, SESS01, Sun StorEdge SE6920, PSX1000)	305 seconds See Note 1
Frame level transmit error	9 seconds
Loss of light on FC link	5 seconds

Note 1: These products have special handling in the driver stack. Their values are calculated based upon determination of the specific device type by the driver stack.

The following pages show tested sets of tunables that can be used to optimize for reduced failover time for specific Sun products. Note that if you are optimizing for performance and stability across all supported storage products, then just stay with the tunable defaults.

Sun StorEdge T3/6x20/99xx Type Devices

The following table indicates how to set the tunables in order to optimize for reduced failover time in fabric environments. Do not use these values if the environment is FC-AL or DAS. If optimizing for performance, just stay with the tunable defaults.

Tunable	Default	Set to:	Notes
`set ssd:ssd_io_time`	60	20	Lowering this value below 60 affects the ability of the array to perform non-disruptive firmware upgrades.
`set ssd:ssd_retry_count`	3	1	Three retries is the default for FC devices. Lowering to 1 keeps the device from looping through the timeout cycle, and significantly reduces failover times.
`set ssd:ssd_max_throttle`	64	8	Limiting this value ensures that there are fewer commands to time out internally, thus reducing busy time during a failover. Timing out potentially hundreds of commands can lead to greatly extended (multiple minutes) failover times. Values below 8 significantly reduce performance. Values over 16 do little to enhance performance.
`set fp:fp_offline_ticker`	90	5	Five seconds allows any Sun-sold switch time to retrain.
`set fp:fp_retry_count`	5	1	

Approximate Time to I/O Failure

Approximate expected behavior at the target driver level in terms of an I/O on a given path failing with the above tunable settings is shown in the table below. Each example assumes that the error condition persists through all the I/O retries. Note that the I/O may ultimately succeed if MPxIO can re-route the I/O down a different path successfully or if a volume manager above the target driver has an alternate source to use to get the data (for example, mirror).

The following values are currently calculated values. Specific testing is planned to validate these calculations and/or modify them as appropriate.

Error Condition	~Time to I/O Failure
Device not responding to I/O	40 seconds
Device returning busy status	10 seconds
Device returning busy status (Sun StorEdge T3, SESS01, Sun StorEdge 6920, PSX1000)	305 seconds See Note 1
Frame level transmit error	6 seconds
Loss of light on FC link	5 seconds

Note 1: These products have special handling in the driver stack. Their values are calculated based upon determination of the specific device type by the driver stack.

Sun StorEdge 351x Type Devices

The following table indicates how to set the tunables to optimize for reduced failover time in fabric, FC-AL, or DAS environments. If optimizing for performance, just stay with the tunable defaults.

Tunable	Default	Set to:	Notes
`set ssd:ssd_io_time`	60	20 (Fabric) 30 (FC-AL or DAS)	• FC_AL Mode: 30 – Allows the loop to stabilize in addition to performing all takeover activities described below. • Fabric Mode: 20 – Allows the takeover controller to assume the "identity" of the failed controller and take over all LUNs. • These values cover both link and controller failures.
`set ssd:ssd_retry_count`	3	2	Lowering to 1 keeps the device from looping through the timeout cycle and significantly reduces failover times. Set this value to 2 to ensure that the command can be re-driven enough times to cover for a long cache flush cycle.
`set ssd:ssd_max_throttle`	64	8	Limiting this value ensures that there are fewer commands to time out internally, thus reducing busy time during a failover. Timing out potentially hundreds of commands can lead to greatly extended (multiple minutes) failover times. Values below 8 significantly reduce performance. Values over 16 do little to enhance performance.
`set fp:fp_offline_ticker`	90	5 (Fabric) 90 (FC-AL or DAS)	For FC-AL topologies or direct connect (DAS), this value must **not** be changed. For example, temporary light loss in a DAS environment may be caused by factors other than a port retraining. For fabric topologies, 5 seconds allows any Sun-sold switch time to retrain.
`set fp:fp_retry_count`	5	1 (Fabric) 5 (FC-AL or DAS)	Only change this in fabric (switch) topologies. Never change in DAS and FC-AL configurations.

The Sun StorEdge 3510 array fails over in a unique manner: When a controller fails, in either loop or fabric mode, the surviving controller takes over the "identity" (Target WWN) of the failed controller, and the integrated bypass hubs in the device automatically re-route the host-facing ports to the surviving controller. MPxIO, per se, does not actually failover to an alternate path. It is required, however, to have multiple paths to each controller for high availability. These paths are symmetric on a per-controller basis, and MPxIO will load-balance across all of them.

Approximate Time to I/O Failure

Approximate expected behavior at the target driver level in terms of an I/O on a given path failing with the above tunable settings is shown in the table below. Each example assumes that the error condition persists through all the I/O retries. Note that the I/O may ultimately succeed if MPxIO can re-route the I/O down a different path successfully or if a volume manager above the target driver has an alternate source to use to get the data (for example, mirror).

The following values are currently calculated values. Specific testing is planned to validate these calculations and/or modify them as appropriate.

Error Condition	**~Time to I/O Failure**
Device not responding to I/O	60 seconds (Fabric) 90 seconds (FC-AL/DAS)
Device returning busy status	15 seconds
Frame level transmit error	6 seconds (Fabric) 18 seconds (FC-AL/DAS)
Loss of light on FC link	5 seconds (Fabric) 90 seconds (FC-AL/DAS)

Matrix Type 2 (Adjustment Not Recommended)

Tunable	Description	Default (Value)	Practical Ranges	Notes
`Set fp:fp_retry_delay=nn` Use: Reduce or expand the wait time between retries.	Controls the wait time between retries. Allows temporary anomalies to subside and increases the chance that the retry succeeds (for example, link speed renegotiation).	3 (seconds)	1-5	See Note 1
`Set fp:fp_log_size=nnnn * nnnn` Use: Increases or decreases the area reserved for error logging at the FP (link) level.	Controls the log-out area for FP errors. Set as a well-formed buffer of `nnnn` size.	1024*1024 (1 MB)	N/A	This value provides sufficient space for even very large configurations.
`Set fp:fp_cmd_wait_cnt=n` Use: Increase or decrease the wait time on completion of all internal commands during a DR operation.	Controls the wait time for *all* internal commands (for example, PLOGI, PRLI, Report LUNs, SCSI Inquires) to complete during a DR (Dynamic Reconfiguration) operation.	240 (seconds)	N/A	This is a tested and verified value that works across all servers that support DR.
`Set ssfcp:ssfcp_log_size=nnnn * nnnn` Use: Same as `fp_log_size`	Same as `fp_log_size`, except logs out errors at the FCP level.	1024*1024 (1 MB)	N/A	This value provides sufficient space for even very large configurations.
`Set ssfcp:ssfcp enable_auto_configuration=n` Use: Allows for the automatic enumeration and configuring of any and all discovered targets and LUNs.	Controls whether all discovered targets and LUNs are automatically configured and enumerated or not. The default in operating systems prior to the Solaris™ 10 Operating System is that discovered targets and LUNs are discovered but **not** configured.	0 - (pre-Solaris 10 OS) 1 - (Solaris 10+ OS)	0, 1	See Note 2

Tunable	Description	Default (Value)	Practical Ranges	Notes
Set `ssfcp:ssfcp_lun_ready_retry=nnn` Use: Shorten or lengthen the time `cfadm` waits before it fails a configure operation.	Controls the total wait time at the FCP layer during a `cfgadm configure` operation for all LUNs associated with the specified SCSI target to become ready (that is, configured and enumerated).	300	N/A	No need to change. Does not affect performance in any way. It just makes sure that the wait time is not "forever."
Set `ssfcp:ssfcp_max_target_retries=nnn` Use: Similar to `ssfcp_lun_ready` except that this is operating on SCSI targets.	Controls the number of configuration retries for a SCSI **target** (not a LUN) to be configured. Issued at 1-second intervals (for example, Target Busy conditions).	120	N/A	No need to change. Does not affect performance in any way. It just makes sure that the retry time is not "forever."

Notes:

1. Reducing the retry delay below three seconds may not allow for a temporary condition (for example, link retraining, buffer credit overflow, and so on) to subside before retrying the operation.
2. A value of one can have the following consequences:

    a. Boot time becomes longer.

    b. The ability to host mask-specific SCSI targets to specific hosts is lost. A setting of one allows the host to see all storage from all HBAs. To prevent unintended security issues and/or data corruption, non-host mechanisms (for example, zoning) should be used to prevent hosts from seeing storage not intended for them.

Changing this value to (1) one (pre-Solaris 10 OS) should **only** be done after careful consideration of the risks to an operational SAN, and when explicit procedures are in place to protect production data from misuse or corruption.

Matrix Type 3 (Not Adjustable – Information Only)

Variable	Description	Value	Notes
MAX_TARGETS_PORT	Sets the maximum number of SCSI targets allowed per port.	256 (Targets)	Should not require any more.
FCP_OFFLINE_DELAY	Controls the wait time to off-line a SCSI target when the target may temporarily "disappear" due to a switch RSCN or other asynchronous events.	20 (Seconds)	This setting is designed to safeguard against the accidental or inappropriate off-lining of a SCSI target due to a short-duration "disappearance" from the fabric while SAN reconfiguration information is being exchanged.
ELS_TIMEOUT	Controls the wait time for a target to perform an ELS type operation (for example, PLOGI, and so on).	20 (Seconds)	Normally completes in under a second. A subset of DISCOVERY_DEADLINE.
FCP_MAX_RETRIES	Controls the number of retries for internal discovery commands only (PLOGI, PRLI, Report LUNs and SCSI inquiry).	4	Should not require more.
DISCOVERY_DEADLINE	Controls the **total** time for all the internal discovery commands to complete (PLOGI, PRLI, Report LUNs and SCSI inquiry).	120 (Seconds)	Normally completes in a few seconds.
FP_NS_TIMEOUT	Controls the timeout value for a switch's name server to respond.	120 (Seconds)	Normally completes in under a second.

Summary and Conclusion

Sun is providing this information to allow you to better understand the discovery, path failover, and general performance tunables and values incorporated in the Leadville driver stack. Sun has chosen and tested default values that are guaranteed to work across Sun's complete range of servers, switches and I/O devices. Sun makes no representation as to the effects, intended or unintended, that may result from changing these values other than as indicated in this paper. These values are provided solely as potentially useful information to a SAN architect who is designing to a specific set of business and functional requirements.

For More Information

For further resources, please see the Storage Administration hub on the BigAdmin System Administration Portal: http://www.sun.com/bigadmin/hubs/storage/

Storage and Information Life Cycle Management Courses
http://www.sun.com/training/catalog/storage/index.xml

Training Courses for the Solaris Operating System
http://www.sun.com/training/catalog/operating_systems/index.xml

Licensing Information