

SAS[®] Foundation 9.3 on Oracle SPARC T4 Server

An I/O Performance and Tuning Summary

Abstract

This paper documents a study on how to improve I/O for intensive SAS workloads running on Oracle SPARC T4 and Solaris 11. Oracle has released a new version of the T-series SPARC processor (T4) with an enhanced floating point capability, and testing shows it can be used effectively for SAS[®] Foundation workloads and the middle tier. Previous versions of this processor (T1 and T2) were aimed primarily at middle-tier or Web-based workloads.

During the study, there were several key observations to highlight. We have listed them here for your review:

- Oracle SPARC T4, Solaris 11, ZFS file systems provides excellent, robust I/O throughput for demanding workloads.
- Varied storage subsystems and I/O tests were used.
- Sustained sequential write throughput demonstrated in the 1.5-2GBps range.
- Sustained sequential read throughput (through cache) demonstrated in the range of 10GBps
- Results consistent with a previous study where 16 simultaneous SAS jobs yielded an effective, combined read/write throughput of 2.1GBps.

Test details

The test suites used in the study included the following:

- vappbench (dd(1) based)
- iotest
- vdbench
- Scalability test composed of SAS programs.

Vappbench and iotest are I/O microbenchmarks, so the focus will be on vdbench and the scalability test composed of SAS programs.

[Vdbench](#) is an enterprise grade I/O workload generator for verifying data integrity and measuring performance of direct attached and network storage on Windows, Solaris, Linux, AIX, OS/X, and HP-UX. The generated workload simultaneously created 64 files of approximately 5GB each and used for both sequential read/ write throughput testing.

Scalability test composed of SAS programs – a rigorous SAS test suite, which performs testing up to a user designated number of concurrent SAS processes. Each SAS process runs through multistep, resource intensive steps that include DATA step, PROC SQL for index creation and numerous PROC SUMMARY steps. Up to 32 concurrent processes were used, to match the number of CPU cores)

System details

Software

- SAS Foundation 9.3
- Solaris 11, SRU 3
- ZFS file systems
- vdbench 503 rc11

Hardware

- SPARC T4-4 (four 2.9GHz (8core))
- 512GB RAM
- Three dual ported, 8gb fibre Host Bus Adapted (HBAs)
- Storage (All ZFS file systems)
 - Six disk pool from T4 internal drives
 - 3PAR - high-end fibre attached storage - 128 spindles/LUN
- /sasdata –one LUN
- /sasdata2 – four LUNs

Results

Storage workloads are captured using Storage Workload Analysis Tool (SWAT) from Oracle/Sun.

Vdbench workload

Vdbench write throughput to the ZFS pool with six internal drives.

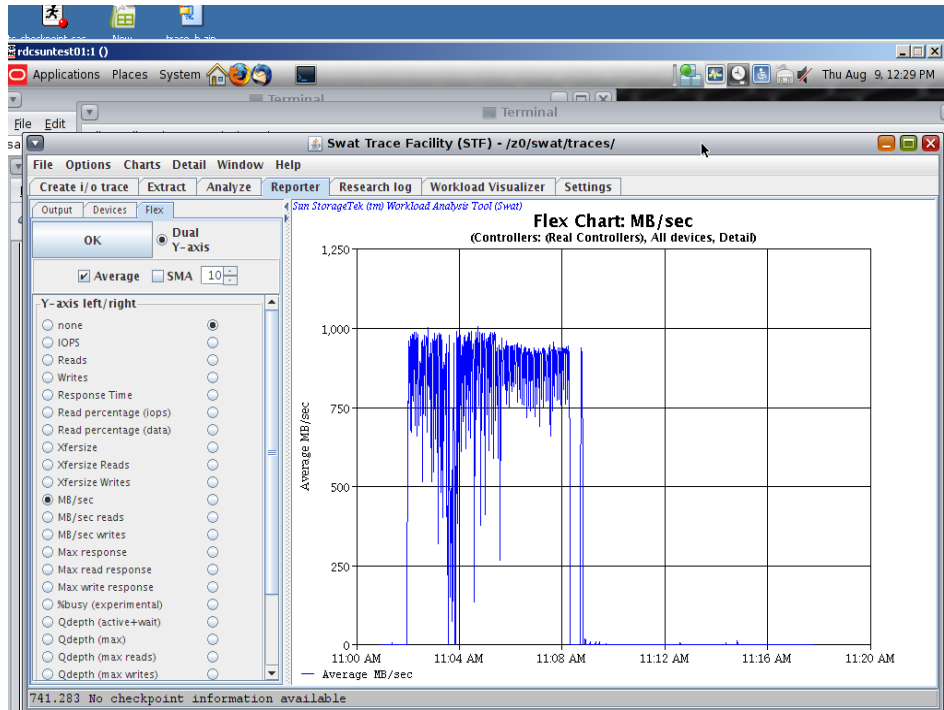


Figure 1: t4pool: ~1GBps Sustained Write Throughput

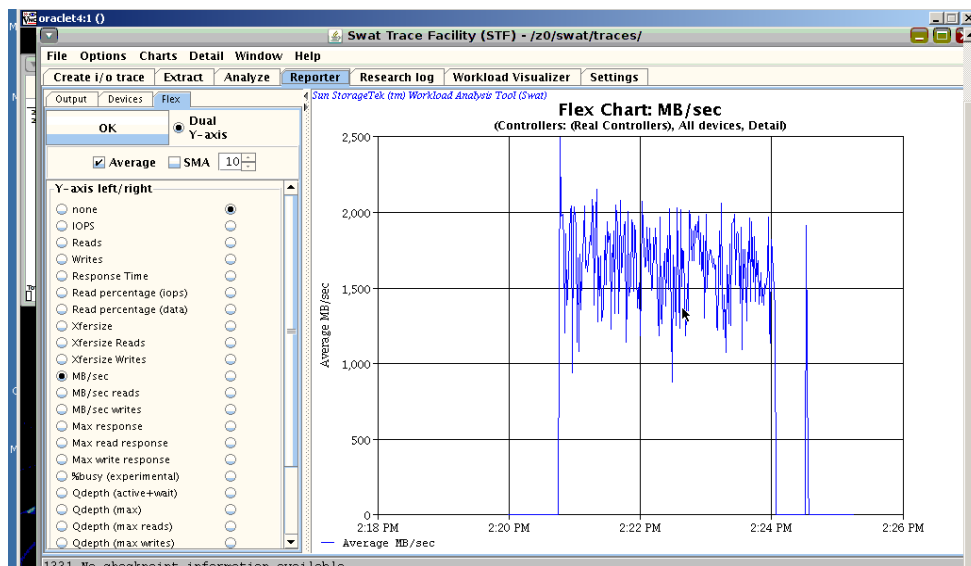


Figure 2: sasdata2 pool: ~2GBps Sustained Write Throughput

Vdbench Write Throughput to Four LUN Fibre Attached Pool (sasdata2)

A read throughput test with vdbench shows extremely high throughput as data is coming out of the file system cache. Here we see the output from vdbench showing samples of 10GBps read throughput

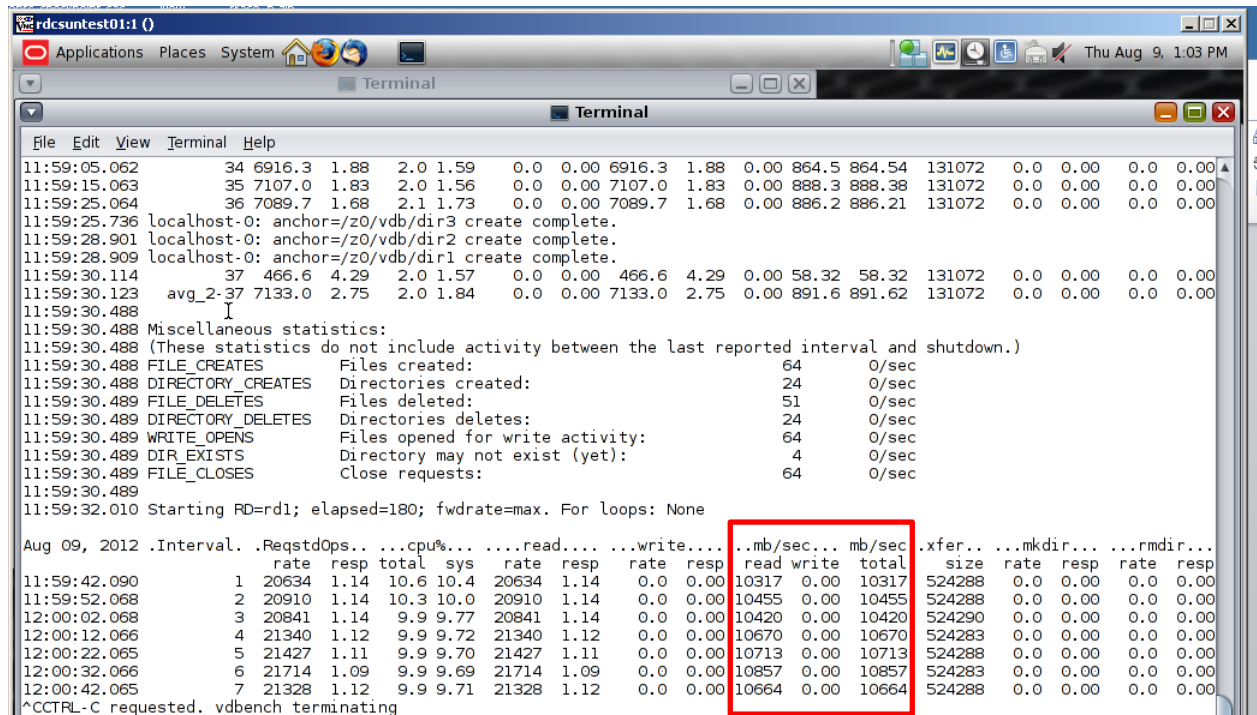


Figure 3: vdbench read throughput of 10GBps

Scalability Test Composed of SAS Programs

For the scalability test composed of SAS programs that alternates heavy I/O and compute resource utilization, we can see sustained peak write throughput ~2GBps.

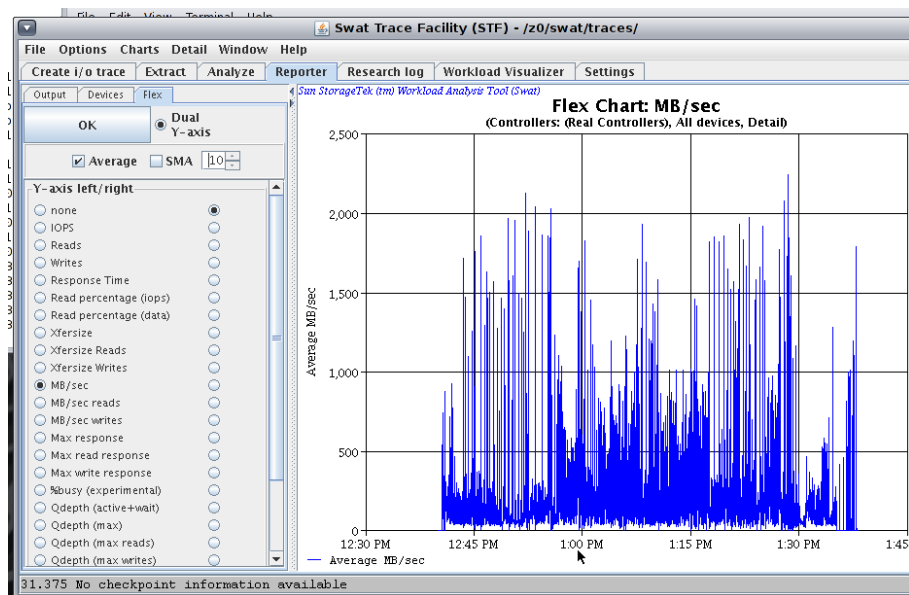


Figure 4: Scalability Test composed of SAS programs I/O Profile at 32 Concurrent Flows

Performance Considerations

- Solaris 11 ZFS Performance Patch for read performance improvement
 - IDR 216.1(x64) / 217.1(SPARC) - compatible with SRU3
- Use More than 1 LUN
 - regardless of the number of spindles backing the LUN
- ZFS tuning parameters
 - ZFS recordsize – default 128K (done on a per ZFS pool basis)
 - Zfs_vdev_max_pending – kernel parameter, Solaris 11 default == 10

Solaris 11 ZFS Performance Patch

A ZFS-related performance optimization patch for CR 7048222 could provide 10-20 percent read performance improvement and is available as IDR 216.1(x64) or IDR 217.1(SPARC). An Interim Diagnostic Relief (IDR) is not generally available and built or released on a case by case basis. This IDR is compatible with SRU3 and planned integration in Solaris 11, Update 1 (also named Solaris 11.1).

The command to install an IDR: `pkg install -g idr217.1.p5p idr217`

Use More than One LUN

Although the fibre attached LUNs each had 128 spindles backing them, the SCSI protocol is limited to 256 outstanding requests per LUN.

Some experimentation was done using one, two, four, six, eight, 10 and 12 LUNs. Using `dd(1)` to raw devices, peak throughput for the larger number of LUNs was higher. It was between 3-4GBps for eight, 10, and 12 LUNs, but with more variability (ie: less consistency). The four LUN configuration was chosen (results for the three LUN configuration were similar) based on the graph in Figure 5 below. The numbers in red represent # concurrent `dd(1)` streams / #LUNs used. So 16/4 means 16 concurrent `dd` streams multiplexed over four LUNs.

Throughput for 4,8,16 jobs on 3&4 LUNs

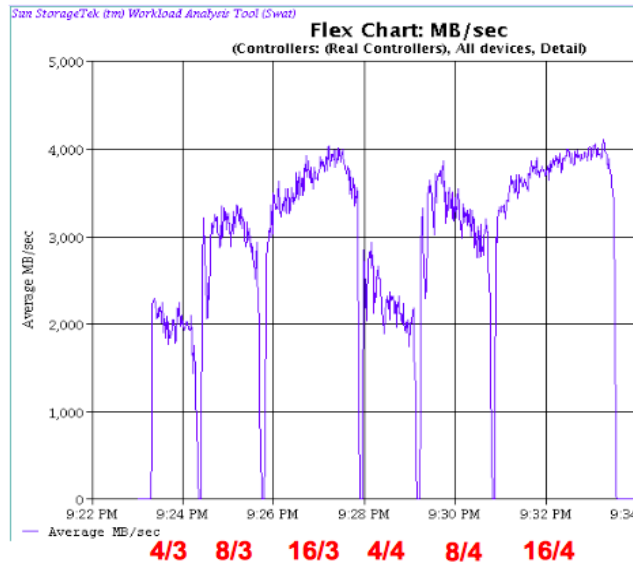


Figure 5: Four LUN Configuration, 16 Concurrent dd Streams: Approximately 4GBps Write Throughput

Running raw device performance tests, the system demonstrated the capability to sustain approximately 3.5GBps write throughput.

Note: Since 8gb fibre channel uses [8b/10b](#) encoding, HBA port can drive a maximum of 750-800 MBps. In order to drive the throughput above, three or four of the six ports (two dual ported HBAs in this case) are required.

Once a ZFS pool was created with four fibre attached LUNs, we saw immediate I/O performance improvement over a ZFS pool with a single fibre attached LUN. The table below highlights the approximate performance improvements

	Single LUN pool	4 LUN pool	Improvement Percentage
iozone	1GBps write throughput	1.5 GBps write throughput	~50%
32 concurrent SAS PROC SUMMARY	Approximately three minutes per job	Approximately one and half and two minutes per job	~30-50%
32 concurrent flows iteration average from the scalability test suite composed of SAS programs.	874 seconds	658 seconds	~25%

Table 1: Performance Improvement, Comparing One to Four LUN ZFS Pool

There was between ~25-50 percent performance improvement using 4 LUNs versus a single LUN. Your mileage can vary depending on the backing storage configuration

ZFS Tuning Parameters

ZFS recordsize

It is prudent to set the ZFS recordsize to match the I/O sizes. The default record size is 128K. If the ZFS file system is singularly purposed (such as to hold Oracle data files), which have mostly known I/O sizes (ie: 16K), It is a good idea to change it. However, in general, if the record sizes are variable and user dependent, we recommend leaving it alone. We saw only a minor improvement in increasing the recordsize from 128K to 256K. To query the record size for the sasdata2 ZFS pool:

```
root@rdcsuntest01:~# zfs get recordsize sasdata2
NAME          PROPERTY      VALUE      SOURCE
sasdata2     recordsize    128K       local
```

To set the record size:

```
root@rdcsuntest01:~# zfs set recordsize=256k t4pool
```

Note: 1) Changing the recordsize only affects files that are newly created. 2) Prior to Solaris 11, the largest recordsize supported was 128K. Solaris 11 now supports a recordsize up to 1 MB. However, if a recordsize larger than 128K is used, the ZFS pool would not longer be compatible (ie: importable) into an older release of Solaris.

Increase zfs_vdev_max_pending

ZFS controls the I/O queue depth for a given LUN with the zfs_vdev_max_pending parameter ([ZFS Evil Tuning Guide](#)). The default is 10 in Solaris 11 and in general, it's best to leave it alone if there are less than 10 disks in a LUN. Since our LUNs have 128 disks per, we increased this to 50.

When multipathing is configured, run "iostat -dxnzY 10" to gauge I/O throughput on a per path basis. If the "actv" field (or queue depth) is consistently at 10, then increasing this parameter could help. We initially set this to 35 and found that we were hitting that threshold and eventually increased to 50.

The field can be modified in 1 of 2 ways – on a temporary basis

```
root@rdcsuntest01:~# echo zfs_vdev_max_pending/W0t50 | mdb -kw
```

To verify

```
root@rdcsuntest01:~# echo zfs_vdev_max_pending/D | mdb -k
zfs_vdev_max_pending:
zfs_vdev_max_pending:
```

To set it on a permanent basis, add to /etc/system and reboot

```
set zfs:zfs_vdev_max_pending=50
```

After setting to 50, we can verify the "**actv**" column from `iostat -cmmxz 5` that there is margin left in the setting as nothing (in this sample) exceeds 30.

```

              extended device statistics
r/s      w/s      kr/s      kw/s      wait actv  wsvc_t  asvc_t   %w %b device
0.0 1218.5    0.0 152695.1  0.0 29.7    0.0    24.4    0 99 c0t5000C5004311D10Fd0
0.0 1215.5    0.0 152666.1  0.0 28.9    0.0    23.8    0 97 c0t5000C50043119247d0
0.0 1228.1    0.0 155169.5  0.0 29.2    0.0    23.8    0 98 c0t5000C5004311CC7Bd0
0.0 1186.9    0.0 150014.9  0.0 27.6    0.0    23.3    0 94 c0t5000C5004311C7B7d0
0.0 1237.1    0.0 155274.0  0.0 29.5    0.0    23.8    0 98 c0t5000C5004311CF33d0
0.0 1247.3    0.0 156431.3  0.0 28.9    0.0    23.2    0 97 c0t5000C500431176CFd0

```

Summary

Four different I/O intensive tests were run on both an internal 6 drive ZFS pool and high end fibre attached ZFS pools. Excellent and robust I/O performance was demonstrated. The results were in line with performance testing, done by SAS staff, where calculated throughput for 16 simultaneous SAS jobs yielded 2.1GBps throughput (combined read/write)

Significant I/O performance improvements were seen after adding the Solaris 11 ZFS performance patch and increasing the number of fibre attached LUNs.

Some ZFS parameter tunings helped performance; in general, no kernel or ZFS tunings are recommended as a blanket recommendation. However, some discussion on increasing the ZFS queue depth was presented.

Appendix

Some useful commands

System Basics (show platform / CPU config)

```
sas@rdcsuntest01:~$ prtdiag -v | more
```

```
System Configuration:  Oracle Corporation  sun4v SPARC T4-4
```

```
Memory size: 523776 Megabytes
```

```
sas@rdcsuntest01:~$ psrinfo -pv
```

```
The physical processor has 8 cores and 64 virtual processors (0-63)
```

```
  The core has 8 virtual processors (0-7)
```

```
  The core has 8 virtual processors (8-15)
```

```
  The core has 8 virtual processors (16-23)
```

```
  The core has 8 virtual processors (24-31)
```

```
  The core has 8 virtual processors (32-39)
```

```
  The core has 8 virtual processors (40-47)
```

```
  The core has 8 virtual processors (48-55)
```

```
  The core has 8 virtual processors (56-63)
```

```
    SPARC-T4 (chipid 0, clock 2998 MHz)
```

```
The physical processor has 8 cores and 64 virtual processors (64-127)
```

```
  The core has 8 virtual processors (64-71)
```

```
  The core has 8 virtual processors (72-79)
```

```
  The core has 8 virtual processors (80-87)
```

```
  The core has 8 virtual processors (88-95)
```

```
  The core has 8 virtual processors (96-103)
```

```
  The core has 8 virtual processors (104-111)
```

```
  The core has 8 virtual processors (112-119)
```

```
  The core has 8 virtual processors (120-127)
```

```
    SPARC-T4 (chipid 1, clock 2998 MHz)
```

```
The physical processor has 8 cores and 64 virtual processors (128-191)
```

```
  The core has 8 virtual processors (128-135)
```

```
  The core has 8 virtual processors (136-143)
```

```
  The core has 8 virtual processors (144-151)
```

```
  The core has 8 virtual processors (152-159)
```

```
  The core has 8 virtual processors (160-167)
```

```
  The core has 8 virtual processors (168-175)
```

```
  The core has 8 virtual processors (176-183)
```

```
  The core has 8 virtual processors (184-191)
```

```
    SPARC-T4 (chipid 2, clock 2998 MHz)
```

```
The physical processor has 8 cores and 64 virtual processors (192-255)
```

```
  The core has 8 virtual processors (192-199)
```

```
  The core has 8 virtual processors (200-207)
```

```
  The core has 8 virtual processors (208-215)
```

```
  The core has 8 virtual processors (216-223)
```

```
  The core has 8 virtual processors (224-231)
```

```
  The core has 8 virtual processors (232-239)
```

```
  The core has 8 virtual processors (240-247)
```

```
  The core has 8 virtual processors (248-255)
```

```
    SPARC-T4 (chipid 3, clock 2998 MHz)
```

Solaris 11 Version Information

Show what is installed (SRU 3 in this case)

```
root@rdcsuntest01:~# pkg info entire
Name: entire
Summary: entire incorporation including Support Repository Update (Oracle Solaris 11
11/11 SRU 03). For more information see:
https://support.oracle.com/CSP/main/article?cmd=show&type=NOT&doctype=REFERENCE&id=1372094
.1
Description: This package constrains system package versions to the same
build. WARNING: Proper system update and correct package
selection depends on the presence of this incorporation.
Removing this package results in an unsupported system.
Category: Meta Packages/Incorporations
State: Installed
Publisher: Solaris
Version: 0.5.11
Build Release: 5.11
Branch: 0.175.0.3.0.4.0
Packaging Date: December 29, 2011 07:15:05 PM
Size: 5.45 kB
FMRI: pkg://solaris/entire@0.5.11,5.11-0.175.0.3.0.4.0:20111229T191505Z
```

Show the version of the current, latest repository (requires support entitlement)

```
root@rdcsuntest01:~# pkg info -r entire
Name: entire
Summary: entire incorporation including Support Repository Update (Oracle Solaris
11 11/11 SRU 9.5).
Description: This package constrains system package versions to the same
build. WARNING: Proper system update and correct package
selection depends on the presence of this incorporation.
Removing this package results in an unsupported system. For
more information go to: https://support.oracle.com/CSP/main/article
?cmd=show&type=NOT&doctype=REFERENCE&id=1372094.1.
Category: Meta Packages/Incorporations
State: Not installed
Publisher: Solaris
Version: 0.5.11 (Oracle Solaris 11 SRU 9.5)
Build Release: 5.11
Branch: 0.175.0.9.0.5.0
Packaging Date: July 5, 2012 06:23:03 PM
Size: 5.45 kB
FMRI: pkg://solaris/entire@0.5.11,5.11 0.175.0.9.0.5.0:20120705T182303Z
```

Multiple boot environments available - *beadm activate sru7* to enable

```
root@rdcsuntest01:~# beadm list
```

BE	Active	Mountpoint	Space	Policy	Created
GA	-	-	12.64M	static	2012-01-25 13:29
solaris-backup-1	-	-	176.0K	static	2012-01-30 14:35
solaris-backup-2	-	-	214.0K	static	2012-03-08 07:42
sru3	-	-	19.62M	static	2012-05-25 10:07
sru3-1	NR	/	118.44G	static	2012-05-25 10:25
sru3_clone	-	-	65.0K	static	2012-05-25 10:22
sru7	-	-	1.14G	static	2012-05-24 12:21

Fibre Channel and Multipathing info

Verify all fibre ports are online and available

```
root@rdcsuntest01:~# fcinfo hba-port
```

```
HBA Port WWN: 10000000c9df8800
```

```
Port Mode: Initiator
```

```
Port ID: 1
```

```
OS Device Name: /dev/cfg/c16
```

```
Manufacturer: Emulex
```

```
Model: LPem12002E-S
```

```
Firmware Version: 2.00a3 (U3D2.00A3)
```

```
FCode/BIOS Version: Boot:5.03a4 Fcode:3.10a3
```

```
Serial Number: 0999VM0-1211001ZZY
```

```
Driver Name: emlxs
```

```
Driver Version: 2.61i (2011.08.10.11.40)
```

```
Type: L-port
```

```
State: online
```

```
Supported Speeds: 2Gb 4Gb 8Gb
```

```
Current Speed: 8Gb
```

```
Node WWN: 20000000c9df8800
```

```
NPIV Not Supported
```

```
HBA Port WWN: 10000000c9df8801
```

```
Port Mode: Initiator
```

```
Port ID: 1
```

```
OS Device Name: /dev/cfg/c17
```

```
Manufacturer: Emulex
```

```
Model: LPem12002E-S
```

```
Firmware Version: 2.00a3 (U3D2.00A3)
```

```
FCode/BIOS Version: Boot:5.03a4 Fcode:3.10a3
```

```
Serial Number: 0999VM0-1211001ZZY
```

```
Driver Name: emlxs
```

```
Driver Version: 2.61i (2011.08.10.11.40)
```

```
Type: L-port
```

```
State: online
```

```
Supported Speeds: 2Gb 4Gb 8Gb
```

```
Current Speed: 8Gb
```

```
Node WWN: 20000000c9df8801
```

```
NPIV Not Supported
```

```
HBA Port WWN: 10000000c9df8764
```

```
Port Mode: Initiator
```

```
Port ID: 1
```

```
OS Device Name: /dev/cfg/c18
```

```
Manufacturer: Emulex
```

```
Model: LPem12002E-S
```

```
Firmware Version: 2.00a3 (U3D2.00A3)
```

```
FCode/BIOS Version: Boot:5.03a4 Fcode:3.10a3
```

```
Serial Number: 0999VM0-1211001ZZA
```

```
Driver Name: emlxs
```

```
Driver Version: 2.61i (2011.08.10.11.40)
```

```
Type: L-port
```

```
State: online
```

```
Supported Speeds: 2Gb 4Gb 8Gb
```

```
Current Speed: 8Gb
```

```
Node WWN: 20000000c9df8764
```

```
NPIV Not Supported
```

```
HBA Port WWN: 10000000c9df8765
```

```

Port Mode: Initiator
Port ID: 1
OS Device Name: /dev/cfg/c19
Manufacturer: Emulex
Model: LPem12002E-S
Firmware Version: 2.00a3 (U3D2.00A3)
FCode/BIOS Version: Boot:5.03a4 Fcode:3.10a3
Serial Number: 0999VM0-1211001ZZA
Driver Name: emlxs
Driver Version: 2.61i (2011.08.10.11.40)
Type: L-port
State: online
Supported Speeds: 2Gb 4Gb 8Gb
Current Speed: 8Gb
Node WWN: 20000000c9df8765
NPIV Not Supported
HBA Port WWN: 10000000c994b424
Port Mode: Initiator
Port ID: 1
OS Device Name: /dev/cfg/c13
Manufacturer: Emulex
Model: LPem12002E-S
Firmware Version: 2.00a3 (U3D2.00A3)
FCode/BIOS Version: Boot:5.03a4 Fcode:3.10a3
Serial Number: 0999VM0-1028001D6M
Driver Name: emlxs
Driver Version: 2.61i (2011.08.10.11.40)
Type: L-port
State: online
Supported Speeds: 2Gb 4Gb 8Gb
Current Speed: 8Gb
Node WWN: 20000000c994b424
NPIV Not Supported
HBA Port WWN: 10000000c994b425
Port Mode: Initiator
Port ID: 1
OS Device Name: /dev/cfg/c14
Manufacturer: Emulex
Model: LPem12002E-S
Firmware Version: 2.00a3 (U3D2.00A3)
FCode/BIOS Version: Boot:5.03a4 Fcode:3.10a3
Serial Number: 0999VM0-1028001D6M
Driver Name: emlxs
Driver Version: 2.61i (2011.08.10.11.40)
Type: L-port
State: online
Supported Speeds: 2Gb 4Gb 8Gb
Current Speed: 8Gb
Node WWN: 20000000c994b425
NPIV Not Supported

```

Show multipathing configuration

```

root@rdcsuntest01:~# mpathadm list lu
/dev/rdsd/c0t5000C50043118DC7d0s2
    Total Path Count: 1
    Operational Path Count: 1
/scsi_vhci/ses@g50002acffffff1593

```

```

        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC0002E1593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC0002D1593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC0002C1593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC0002B1593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC0002A1593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000291593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000281593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000271593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000261593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000251593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000241593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000231593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t50002AC000201593d0s2
        Total Path Count: 6
        Operational Path Count: 6
/dev/rdisk/c0t5000C50043107DD3d0s2
        Total Path Count: 1
        Operational Path Count: 1
/dev/rdisk/c0t5000C5004311D10Fd0s2
        Total Path Count: 1
        Operational Path Count: 1
/dev/rdisk/c0t5000C50043119247d0s2
        Total Path Count: 1
        Operational Path Count: 1
/dev/rdisk/c0t5000C5004311CC7Bd0s2
        Total Path Count: 1
        Operational Path Count: 1
/dev/rdisk/c0t5000C5004311C7B7d0s2
        Total Path Count: 1
        Operational Path Count: 1
/dev/rdisk/c0t5000C5004311CF33d0s2

```

```
Total Path Count: 1
Operational Path Count: 1
/dev/rdisk/c0t5000C500431176CFd0s2
Total Path Count: 1
Operational Path Count: 1
```

CONTACT INFORMATION

Your comments and questions are valued and encouraged. For any questions or sample codes described in this paper, please contact the authors at:

Maureen Chew

Email: maureen.chew@oracle.com

Oracle (on-site), S0058

(919) 531-5852



SAS Institute Inc. World Headquarters +1 919 677 8000

To contact your local SAS office, please visit: **sas.com/offices**

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2012, SAS Institute Inc. All rights reserved. T4_0812