ORACLE
SOLARIS

An Oracle White Paper
May 2010

Quick-Start Tuning Guide for Oracle Solaris Running Databases and Business Intelligence Applications With Sun Storage 7000 Unified Storage Systems

ORACLE

1 Objective

The purpose of this guide is to get the maximum benefit from Oracle's Sun Storage 7000 Unified Storage Systems when running databases and business applications on the Oracle Solaris OS. This guide does not go into detail on the background of features and how particular tunable parameters work, but it describes common tuning parameters that have been found to improve performance.

2 Oracle Solaris Tuning

This section describes Oracle Solaris tuning that helps improve performance with Sun Storage 7000 Unified Storage Systems. Oracle Solaris tuning primarily includes tuning variables in `/etc/system`. The network stack can be additionally tuned using the `ndd` utility. The various device drivers typically have their own configuration file, which needs to be tuned (for Jumbo Frames).

2.1 Tuning `/etc/system` Parameters

This subsection describes the basic tunings for each such component by function. The lists show typical tunings used for the individual components being stressed. Depending on what is used on the server running Oracle Solaris, the tunables will help to improve performance over the default out-of-the-box performance.The Network Cache and Acceleration (NCA) section is typically tuned when a server running Oracle Solaris is used as a Web Server. For more information on NCA, refer to *The Solaris Network Cache and Accelerator* white paper:

http://www.sun.com/software/whitepapers/Oracle Solaris9/networkcache.pdf

The NFS-related tunings help improve performance of NFS to get greater throughput than the default settings. Oracle Solaris Zettabyte File System (ZFS) tuning applies to pools created on shares from Unified Storage using 10GE ports. The following table shows typical tunings used for the individual components being stressed. Now all of them have to be applied. Depending on what is used on the server running Oracle Solaris, the tunables will help to improve performance over the default out-of-the-box performance.

Set 1: Oracle Solaris `/etc/system` Tunables for Different Workloads

```
* NCA tuning per sun doc 817-0404 for a
* 4GB RAM system with 64bit kernel.
set sq_max_size=0 set ge:ge_intr_mode=1
set nca:nca_conn_hash_size=82500 set nca:nca_conn_req_max_q=100000
set nca:nca_conn_req_max_q0=100000
set nca:nca_ppmax=393216
set nca:nca_vpmax=39321
* For NFS
set nfs:nfs3_max_threads=256
set nfs:nfs4_max_threads=256
set nfs:nfs3_nra=32
set nfs:nfs4_nra=32
set nfs:nfs3_bsize=1048576
set nfs:nfs3_max_transfer_size=1048576
```

```
*For NFS throughput
set rpcmod:clnt_max_conns = 8
set hires_tick=1
*For Oracle Solaris ZFS
set zfs:zfetch_max_streams=64
set zfs:zfetch_block_cap=2048
set zfs:zfs_txg_synctime=1
set zfs:zfs_vdev_max_pending = 8
*For IP Stack
set ip:ip_squeue_fanout=1
set ip:ip_squeue_bind=0
set ip:ip_squeue_worker_wait=1

*Next one only for Solaris 10 05/09 (update 7) or earlier
set ip:ip_soft_rings_cnt=16
* http://www.Oracle Solarisinternals.com/wiki/index.php/Networks
* For ixgbe or nxge
set ddi_msix_alloc_limit=8
* For nxge
set nxge:nxge_bcopy_thresh=1024
set pcplusmp:apic_multi_msi_max=8
set pcplusmp:apic_msix_max=8
set pcplusmp:apic_intr_policy=1
set nxge:nxge_msi_enable=2
```

2.2 Tuning Network Stack With `ndd`

Apart from the NFS and IP tuning done using `/etc/system` in the previous section, there are other tunables that can only be modified using the `ndd` utility. It is convenient to put all `ndd` tunables in a single shell script and execute them as part of a startup routine by placing a link to the file in `/etc/init.d/` and then using a symbolic link from `/etc/rc3.d/S99name.sh` to the original file in `/etc/init.d`.

Set 2: Oracle Solaris `ndd` Tunables Using `/etc/rc3.d/S99network.sh`

```
#!/bin/sh
# increase max tcp window
# Rule-of-thumb: max_buf = 2 x cwnd_max (congestion window)
ndd -set /dev/tcp tcp_max_buf 4194304
ndd -set /dev/tcp tcp_cwnd_max 2097152
# increase DEFAULT tcp window size
ndd -set /dev/tcp tcp_xmit_hiwat 131072
ndd -set /dev/tcp tcp_recv_hiwat 131072
ndd -set /dev/tcp tcp_conn_req_max_q 16384
ndd -set /dev/tcp tcp_conn_req_max_q0 16384
ndd -set /dev/tcp tcp_naglim_def 1
```

2.3 Tuning Network Drivers for Jumbo Frames

Jumbo frames help improve NFS performance. However, they need to be enabled and are not enabled by default by the network drivers. If you are using an ixgb, nxge, e1000g, or ce network driver to plumb your network interface, it can be tuned by editing the respective network driver configuration file, as seen in the examples below. The changes take place only when the driver is reloaded, which typically requires a reboot.

Set 3: Oracle Solaris Network Driver Tuning for igxb, nxge, e1000g, and ce

```
#vi /kernel/drv/ixgb.conf default_mtu=8150;
tx_copy_threshold=1024;

# vi /platform/i86pc/kernel/drv/nxge.conf
accept_jumbo = 1;
soft-lso-enable = 1;
rxdma-intr-time=1;
rxdma-intr-pkts=8;

# vi /kernel/drv/e1000g.conf
MaxFrameSize=3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3;
#MaxFrameSize=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;

#vi /platform/sun4u/kernel/drv/ce.conf
accept_jumbo = 1;
```

When jumbo frames are enabled, the `ifconfig` output will show the higher MTU. For example, an e1000g0 interface configured to use jumbo frames shows MTU of 10244 in the following example.

```
# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu
8232 index 1
        inet 127.0.0.1 netmask ff000000
e1000g0: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 10244
index 3
        inet 192.168.200.2 netmask ffffff00 broadcast 192.168.200.255
        ether 0:14:4f:ad:df:2d
```

3 Accessing Shares From Sun Storage 7000 Unified Storage System

There are three ways to access the shares from Sun Storage 7000 Unified Storage Systems via Oracle Solaris. The three ways are NFS, iSCSI, or FC. This section describes how to access the shares from Sun Storage 7000 Unified Storage Systems depending on the protocol.

3.1 Accessing NFS Shares From Oracle Solaris

NFS shares can be simply mounted as follows:

```
# mount -F nfs 10.9.168.93:/export/mynfsshare /mountpoint
```

Note: In the next chapter, we discuss more about tuning NFS shares.

3.2 Accessing iSCSI Shares From Oracle Solaris

Quick reference steps to mount iSCSI shares from an Oracle Solaris client are as follows:Add and display the iSCSI target discovery address by giving the IP address of the Sun Storage 7000 Unified Storage System.

```
# iscsiadm add discovery-address 10.9.168.93
# iscsiadm list discovery-address
Discovery Address: 10.9.168.93:3260
```

Display iSCSI targets discovered from the Sun Storage 7000 Unified Storage System.

```
# iscsiadm list discovery-address -v 10.9.168.93
Discovery Address: 10.6.140.151:3260
Target name: iqn.1986-03.com.sun:02:a4602145-85f8-64fa-c0ef-
a059394d9a12
Target address: 10.9.168.93:3260, 1
Target name: iqn.1986-03.com.sun:02:0449398a-486f-4296-9716-
bcba3c1be41c  Target address: 10.9.168.93:3260, 1
```

Enable and display static discovery.

```
# iscsiadm modify discovery --static enable
# iscsiadm list discovery
Discovery:
Static: enabled
```

```
Send Targets: disabled
iSNS: disabled
```

Add a target to the list of statically configured targets. A connection to the target will not be attempted unless the static configuration method of discovery has been enabled.

```
# iscsiadm add static-config iqn.1986-03.com.sun:02:9e0b0e03-8823-
eb7e-d449-f9c21930ba15,10.9.168.93

# iscsiadm add static-config iqn.1986-03.com.sun:02:2cc4fe10-c7ba-
697f-d95f-fa75efe50239,10.9.168.93
```

Use Oracle Solaris `devfsadm`(1M) to create iSCSI device nodes.

```
# devfsadm -i iscsi
```

Use the `format`(1M) command to access iSCSI disks. The disk(s) to be selected contain `/scsi_vhci` in their path name. Local disks are listed before iSCSI disks in the `format` command list. The following shows disk numbers 4 and 5 are iSCSI disks.

```
# format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
0. c0t0d0 <DEFAULT cyl 17830 alt 2 hd 255 sec 63>
/pci@0,0/pci8086,25f8@4/pci108e,286@0/sd@0,0
1. c0t1d0 <DEFAULT cyl 17830 alt 2 hd 255 sec 63>
/pci@0,0/pci8086,25f8@4/pci108e,286@0/sd@1,0
2. c0t2d0 <DEFAULT cyl 17830 alt 2 hd 255 sec 63>
/pci@0,0/pci8086,25f8@4/pci108e,286@0/sd@2,0
3. c0t3d0 <DEFAULT cyl 17830 alt 2 hd 255 sec 63>
/pci@0,0/pci8086,25f8@4/pci108e,286@0/sd@3,0
4. c2t600144F04890703F0000144FA6CCAC00d0 <DEFAULT cyl 13051 alt 2 hd
255 sec 63>  /scsi_vhci/disk@g600144f04890703f0000144fa6ccac00
5. c2t600144F0489070250000144FA6CCAC00d0 <DEFAULT cyl 13051 alt 2 hd
255 sec 63>  /scsi_vhci/disk@g600144f0489070250000144fa6ccac00
```

3.3 Accessing FC Shares From Oracle Solaris

Typically shares exported via Fibre Channel will be available for access from Oracle Solaris by rebooting with `reconfigure` or `devfsadm` (provided that Fibre Channel HBA drivers and fiber connections are working as expected).

```
# devfsadm
```

```
# format
Searching for disks...done

AVAILABLE DISK SELECTIONS:
       0. c0t600144F0DA56519F00004B63451A0005d0 <SUN-Sun Storage
7410-1.0-128.00GB>
/scsi_vhci/disk@g600144f0da56519f00004b63451a0005
       1. c0t600144F0DA56519F00004B63451B0006d0 <SUN-Sun Storage
7410-1.0-128.00GB>
/scsi_vhci/disk@g600144f0da56519f00004b63451b0006
       2. c0t600144F0DA56519F00004B63451C0007d0 <SUN-Sun Storage
7410-1.0-128.00GB>
/scsi_vhci/disk@g600144f0da56519f00004b63451c0007
       3. c0t600144F0DA56519F00004B63451D0008d0 <SUN-Sun Storage
7410-1.0-128.00GB>
/scsi_vhci/disk@g600144f0da56519f00004b63451d0008
       4. c0t600144F0DA56519F00004B63451E0009d0 <SUN-Sun Storage
7410-1.0-128.00GB>
```

Important Note: For shares accessed over Fiber Channel or iSCSI, depending on the label selected (EFI or Oracle Solaris VTOC), there could be alignment issues with the blocks presented by the share. There is a separate paper titled *Detecting and Resolving Oracle Solaris LUN Alignment Problem* that will help format the labels with the right offset to resolve alignment problems.

4 Tuning Mounted Shares on Oracle Solaris

This section describes how to tune the respective file systems. There are three ways to access the shares of Sun Storage 7000 Unified Storage Systems through Oracle Solaris. The three ways are NFS, iSCSI, or FC. Typically, NFS shares will be used directly as NFS mounted file systems, while iSCSI or FC shares will be used raw by applications running on Oracle Solaris or via a file system such as UFS or Oracle Solaris ZFS.

4.1 Tuning NFS Mount Options

If you have jumbo frames enabled and need more throughput through NFS for databases or business intelligence applications, the following mount options will provide improved throughput over the default NFS mount. For example, the following allows a 1-MB size chunk of IO being transmitted through NFS (depending on the OS version).

```
# mount -F nfs -o
hard,nointr,rsize=1048576,wsize=1048576,proto=tcp,vers=3,noac,forcedi
rectio,rw 10.9.168.93:/export/mynfsshare /mountpoint
```

The mount options above typically enable the best single-access throughput speeds. However, when there are multiple users accessing the file systems, it has been found that 128kB sizes are more scalable than 1MB size. Also, for older versions of Oracle Solaris wherein 1MB sizes may not be supported, NFS can be mounted with 128kB.

```
# mount -F nfs -o
hard,nointr,rsize=131072,wsize=131072,proto=tcp,vers=3,noac,forcedire
ctio,rw 10.9.168.93:/export/mynfsshare /mountpoint
```

Alternatively, in cases where the number of users accessing data on the shares is very large, the 32kB size may be the best option.

```
# mount -F nfs -o
hard,nointr,rsize=32768,wsize=32768,proto=tcp,vers=3,noac,forcedirect
io,rw 10.9.168.93:/export/mynfsshare /mountpoint
```

The `forcedirectio` flag above indicates to the NFS client that it will not use client memory as cache for the NFS shares (useful for databases that cache data in their own bufferpool).

4.2 Tuning UFS Over iSCSI or FC Shares

If you are using UFS for a LUN presented over an iSCSI or FC target, the following steps will help you set up and fine tune the file system for the Sun Storage 7000 Unified Storage System shares.

Creating UFS file systems on iSCSI disks:

```
# newfs /dev/rdsk/c2t600144F04890703F0000144FA6CCAC00d0s0
```

```
# mount  /dev/dsk/c2t600144F0489070250000144FA6CCAC00d0s0 /mountpoint
```

The `maxcontig` tunable can be set using `tunefs`(1M) to match the `recordsize` used for the share. If the share `recordsize` is 128K, then the following will set the `maxcontig` to 128K for the UFS file system used.

```
# tunefs -a 14  /dev/dsk/c2t600144F0489070250000144FA6CCAC00d0s0
```

The value is calculated as follows: 14 = 128K/8K (recordsize/ pagesize of UFS).

The tunable `maxphys` should be at least equal to the `recordsize` or bigger. It is set in `/etc/system`.

```
set maxphys=131072  # or greater than this value
```

Also, if UFS is mounted with default mount options (which uses `noforcedirectio`) or without `forcedirectio`, then you need to verify other parameters to verify that a value `ufs:klustsize` matches the `recordsize`.

```
set ufs:klustsize=131072 # For Oracle Solaris x64
set klustsize=131072 # For Oracle Solaris SPARC
```

4.3 Tuning Oracle Solaris ZFS Over iSCSI or FC Shares

To use Oracle Solaris ZFS as the file system for an exported share (over iSCSI or FC), a zpool will need to be created over the raw share.

```
zpool create myzpool c2t600144F04890703F0000144FA6CCAC00d0s0
```

The default ZFS dataset properties should be changed to match the share properties. For example, the `recordsize` should be changed from the default to match the share `recordsize`.

```
zfs set recordsize=32k myzpo
```

5 For More Information

Here are additional resources.

Oracle Solaris resources:

- BigAdmin web applications and device lists:
  - Hardware Compatibility Lists for Solaris OS: http://www.sun.com/bigadmin/hcl/data/sol/
  - Hardware Certification Test Suite: http://www.sun.com/bigadmin/hcl/hcts/index.jsp
  - Solaris 10 Applications Library: http://www.sun.com/bigadmin/apps/
  - Sun Device Detection Tool: http://www.sun.com/bigadmin/hcl/hcts/device_detect.jsp
  - Installation Check Tool: http://www.sun.com/bigadmin/hcl/hcts/install_check.jsp
  - Device and Third-party Solaris Device Driver Reference Lists: http://www.sun.com/bigadmin/drivers
- BigAdmin Oracle Solaris technology resource centers:
  - Solaris Information Center: http://www.sun.com/bigadmin/hubs/documentation
  - Solaris Patching Center: http://www.sun.com/bigadmin/patches/solaris/index.jsp
  - Solaris 10 Upgrade Resources for System Administrators: http://www.sun.com/bigadmin/topics/upgrade/
  - Solaris Containers (Zones): http://www.sun.com/bigadmin/content/zones/
  - Logical Domains (LDoms): http://www.sun.com/bigadmin/hubs/ldoms/
  - DTrace: http://www.sun.com/bigadmin/content/dtrace/index.jsp
  - ZFS: http://www.sun.com/bigadmin/topics/zfs/
  - Predictive Self-Healing: http://www.sun.com/bigadmin/content/selfheal/
  - Solaris 8 Vintage Support: http://www.sun.com/bigadmin/topics/vintagepatch/
- BigAdmin Oracle Solaris resource collections (which include community submissions):
  - Solaris resource collection: http://www.sun.com/bigadmin/collections/solaris.jsp
  - Solaris 10 resource collection: http://www.sun.com/bigadmin/collections/solaris10.jsp
  - Solaris 9 resource collection: http://www.sun.com/bigadmin/collections/solaris9.jsp
  - Solaris 8 resource collection: http://www.sun.com/bigadmin/collections/solaris8.jsp
  - Solaris on x86 resource collection: http://www.sun.com/bigadmin/collections/solarisx86.jsp
- Discussions, such as the Solaris OS forums: http://forums.sun.com/category.jspa?categoryID=65
- BigAdmin Operating System wiki page: http://wikis.sun.com/display/BigAdmin/Operating+System

Storage resources:

- Sun Storage 7000 Unified Storage Systems web site: http://www.sun.com/storage/disk_systems/unified_storage/
- Sun Disk Storage web page: http://www.sun.com/storagetek/open.jsp
- Sun Storage Solutions web page: http://www.sun.com/storagetek/solutions.jsp
- Discussions, such as the Storage forums (http://forum.java.sun.com/category.jspa?categoryID=66) and the Sun Hardware - Servers forums (http://forums.sun.com/forum.jspa?forumID=830)
- Wikis, such as the Storage Administration wiki (http://wikis.sun.com/display/StorageAdmin/Home) and the BigAdmin Storage Tech Tips wiki (http://wikis.sun.com/display/BigAdmin/Storage+Tech+Tips)
- Storage Stop Blog: http://blogs.sun.com/storage
- Resources on BigAdmin, such as the Storage resource collection (includes community submissions): http://www.sun.com/bigadmin/collections/storage.jsp

General links:

- Sun download site: http://www.sun.com/download/
- Oracle University web site: http://www.sun.com/training/
- Discussions, such as Sun forums (http://forums.sun.com/index.jspa) and the BigAdmin Discussions collection (http://www.sun.com/bigadmin/discussions/)
- Sun product documentation at http://docs.sun.com and the Sun Documentation Center (http://www.sun.com/documentation/)
- Sun wikis, such as the Sun BluePrints wiki (http://wikis.sun.com/display/BluePrints/Main)and the BigAdmin wiki (http://wikis.sun.com/display/BigAdmin/Home)
- Support:
  - Sun resources:

o Register your gear: https://inventory.sun.com/inventory/

o Sun Services: http://www.sun.com/service/index.jsp

o SunSolve Online: http://sunsolve.sun.com

  - Community system administration experts: http://www.sun.com/bigadmin/content/communityexperts/

Oracle is committed to developing practices and products that help protect the environment

Quick-Start Tuning Guide for Oracle Solaris Running Databases and Business Intelligence Applications With Sun Storage 7000 Unified Storage Systems
May 2010
Author: Jignesh Shah, AIE
Contributing Authors: Sridhar Ranganathan, AIE

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

SOFTWARE. HARDWARE. COMPLETE.