

Oracle® Solaris 11.1 Tunable Parameters Reference Manual

Copyright © 2000, 2013, Oracle and/or its affiliates. All rights reserved.

This software and related documentation are provided under a license agreement containing restrictions on use and disclosure and are protected by intellectual property laws. Except as expressly permitted in your license agreement or allowed by law, you may not use, copy, reproduce, translate, broadcast, modify, license, transmit, distribute, exhibit, perform, publish, or display any part, in any form, or by any means. Reverse engineering, disassembly, or decompilation of this software, unless required by law for interoperability, is prohibited.

The information contained herein is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

If this is software or related documentation that is delivered to the U.S. Government or anyone licensing it on behalf of the U.S. Government, the following notice is applicable:

U.S. GOVERNMENT END USERS. Oracle programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, delivered to U.S. Government end users are "commercial computer software" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, use, duplication, disclosure, modification, and adaptation of the programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, shall be subject to license terms and license restrictions applicable to the programs. No other rights are granted to the U.S. Government.

This software or hardware is developed for general use in a variety of information management applications. It is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury. If you use this software or hardware in dangerous applications, then you shall be responsible to take all appropriate fail-safe, backup, redundancy, and other measures to ensure its safe use. Oracle Corporation and its affiliates disclaim any liability for any damages caused by use of this software or hardware in dangerous applications.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group.

This software or hardware and documentation may provide access to or information on content, products, and services from third parties. Oracle Corporation and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services. Oracle Corporation and its affiliates will not be responsible for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services.

Ce logiciel et la documentation qui l'accompagne sont protégés par les lois sur la propriété intellectuelle. Ils sont concédés sous licence et soumis à des restrictions d'utilisation et de divulgation. Sauf disposition de votre contrat de licence ou de la loi, vous ne pouvez pas copier, reproduire, traduire, diffuser, modifier, breveter, transmettre, distribuer, exposer, exécuter, publier ou afficher le logiciel, même partiellement, sous quelque forme et par quelque procédé que ce soit. Par ailleurs, il est interdit de procéder à toute ingénierie inverse du logiciel, de le désassembler ou de le décompiler, excepté à des fins d'interopérabilité avec des logiciels tiers ou tel que prescrit par la loi.

Les informations fournies dans ce document sont susceptibles de modification sans préavis. Par ailleurs, Oracle Corporation ne garantit pas qu'elles soient exemptes d'erreurs et vous invite, le cas échéant, à lui en faire part par écrit.

Si ce logiciel, ou la documentation qui l'accompagne, est concédé sous licence au Gouvernement des Etats-Unis, ou à toute entité qui délivre la licence de ce logiciel ou l'utilise pour le compte du Gouvernement des Etats-Unis, la notice suivante s'applique:

U.S. GOVERNMENT END USERS. Oracle programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, delivered to U.S. Government end users are "commercial computer software" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, use, duplication, disclosure, modification, and adaptation of the programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, shall be subject to license terms and license restrictions applicable to the programs. No other rights are granted to the U.S. Government.

Ce logiciel ou matériel a été développé pour un usage général dans le cadre d'applications de gestion des informations. Ce logiciel ou matériel n'est pas conçu ni n'est destiné à être utilisé dans des applications à risque, notamment dans des applications pouvant causer des dommages corporels. Si vous utilisez ce logiciel ou matériel dans le cadre d'applications dangereuses, il est de votre responsabilité de prendre toutes les mesures de secours, de sauvegarde, de redondance et autres mesures nécessaires à son utilisation dans des conditions optimales de sécurité. Oracle Corporation et ses affiliés déclinent toute responsabilité quant aux dommages causés par l'utilisation de ce logiciel ou matériel pour ce type d'applications.

Oracle et Java sont des marques déposées d'Oracle Corporation et/ou de ses affiliés. Tout autre nom mentionné peut correspondre à des marques appartenant à d'autres propriétaires qu'Oracle.

Intel et Intel Xeon sont des marques ou des marques déposées d'Intel Corporation. Toutes les marques SPARC sont utilisées sous licence et sont des marques ou des marques déposées de SPARC International, Inc. AMD, Opteron, le logo AMD et le logo AMD Opteron sont des marques ou des marques déposées d'Advanced Micro Devices. UNIX est une marque déposée d'The Open Group.

Ce logiciel ou matériel et la documentation qui l'accompagne peuvent fournir des informations ou des liens donnant accès à des contenus, des produits et des services émanant de tiers. Oracle Corporation et ses affiliés déclinent toute responsabilité ou garantie expresse quant aux contenus, produits ou services émanant de tiers. En aucun cas, Oracle Corporation et ses affiliés ne sauraient être tenus pour responsables des pertes subies, des coûts occasionnés ou des dommages causés par l'accès à des contenus, produits ou services tiers, ou à leur utilisation.

Contents

Preface	13
1 Overview of Oracle Solaris System Tuning	17
What's New in Oracle Solaris System Tuning?	17
Tuning an Oracle Solaris System	19
Tuning Format of Tunable Parameters Descriptions	19
Tuning the Oracle Solaris Kernel	21
/etc/system File	21
kldb Command	22
mdb Command	22
Special Oracle Solaris tune and var Structures	23
Viewing Oracle Solaris System Configuration Information	24
sysdef Command	24
kstat Utility	24
2 Oracle Solaris Kernel Tunable Parameters	25
Where to Find Tunable Parameter Information	25
General Kernel and Memory Parameters	26
physmem	26
default_stksize	26
lwp_default_stksize	27
logevent_max_q_sz	28
segkpsize	29
noexec_user_stack	30
fsflush and Related Parameters	30
fsflush	30
tune_t_fsflushr	31

autoup	32
dopageflush	33
doiflush	33
Process-Sizing Parameters	34
maxusers	34
reserved_procs	35
pidmax	36
max_nprocs	37
maxuprc	37
ngroups_max	38
Paging-Related Parameters	39
lotsfree	40
desfree	41
minfree	42
throttlefree	43
pageout_reserve	44
pages_pp_maximum	45
tune_t_minarmem	45
fastscan	46
slowscan	47
min_percent_cpu	47
handspreadpages	48
pages_before_pager	48
maxpgio	49
Swapping-Related Parameters	50
swapfs_reserve	50
swapfs_minfree	51
Kernel Memory Allocator	52
kmem_flags	52
kmem_stackinfo	53
General Driver Parameters	54
moddebug	54
ddi_msix_alloc_limit	55
Network Driver Parameters	56
igb Parameters	56
ixgbe Parameters	57

General I/O Parameters	61
maxphys	61
rlim_fd_max	61
rlim_fd_cur	62
General File System Parameters	63
ncsize	63
dnlc_dir_enable	64
dnlc_dir_min_size	64
dnlc_dir_max_size	65
dnlc_dircache_percent	65
TMPFS Parameters	66
tmpfs:tmpfs_maxkmem	66
tmpfs:tmpfs_minfree	67
Pseudo Terminals	67
pt_cnt	68
pt_pctofmem	69
pt_max_pty	69
STREAMS Parameters	70
nstrpush	70
strmsgsz	70
strctlsz	71
System V Message Queues	71
System V Semaphores	72
System V Shared Memory	72
segspt_minfree	72
Scheduling	73
disp_rechoose_interval	73
Timers	74
hires_tick	74
timer_max	74
SPARC System Specific Parameters	75
tsb_alloc_hiwater_factor	75
default_tsb_size	76
enable_tsb_rss_sizing	77
tsb_rss_factor	77
Locality Group Parameters	78

lpg_alloc_prefer	78
lgrp_mem_pset_aware	79
3 Oracle Solaris ZFS Tunable Parameters	81
Where to Find Tunable Parameter Information	81
Tuning ZFS Considerations	82
ZFS ARC Parameters	82
zfs_arc_min	82
zfs_arc_max	83
ZFS File-Level Prefetch	83
zfs_prefetch_disable	83
ZFS Device I/O Queue Depth	84
zfs_vdev_max_pending	84
Tuning ZFS When Using Flash Storage	85
Adding Flash Devices as ZFS Log or Cache Devices	86
Ensuring Proper Cache Flush Behavior for Flash and NVRAM Storage Devices	87
Tuning ZFS for Database Products	89
Tuning ZFS for an Oracle Database	90
Using ZFS with MySQL Considerations	93
4 NFS Tunable Parameters	95
Where to Find Tunable Parameter Information	95
Tuning the NFS Environment	95
NFS Module Parameters	96
nfs:nfs3_pathconf_disable_cache	96
nfs:nfs4_pathconf_disable_cache	96
nfs:nfs_allow_preepoch_time	97
nfs:nfs_cots_timeo	98
nfs:nfs3_cots_timeo	98
nfs:nfs4_cots_timeo	99
nfs:nfs_do_symlink_cache	100
nfs:nfs3_do_symlink_cache	100
nfs:nfs4_do_symlink_cache	101
nfs:nfs_dynamic	101
nfs:nfs3_dynamic	102

nfs:nfs_lookup_neg_cache	102
nfs:nfs3_lookup_neg_cache	103
nfs:nfs4_lookup_neg_cache	104
nfs:nfs_max_threads	105
nfs:nfs3_max_threads	106
nfs:nfs4_max_threads	107
nfs:nfs_nra	107
nfs:nfs3_nra	108
nfs:nfs4_nra	109
nfs:nrnode	109
nfs:nfs_shrinkreaddir	110
nfs:nfs3_shrinkreaddir	111
nfs:nfs_write_error_interval	112
nfs:nfs_write_error_to_cons_only	112
nfs:nfs_disable_rddir_cache	113
nfs:nfs3_bsize	114
nfs:nfs4_bsize	114
nfs:nfs_async_clusters	115
nfs:nfs3_async_clusters	116
nfs:nfs4_async_clusters	117
nfs:nfs_async_timeout	118
nfs:nacache	118
nfs:nfs3_jukebox_delay	119
nfs:nfs3_max_transfer_size	120
nfs:nfs4_max_transfer_size	121
nfs:nfs3_max_transfer_size_clts	122
nfs:nfs3_max_transfer_size_cots	122
rpcmod Module Parameters	123
rpcmod:clnt_max_conns	123
rpcmod:clnt_idle_timeout	124
rpcmod:svc_idle_timeout	124
rpcmod:svc_default_stksize	124
rpcmod:maxdupreqs	125
rpcmod:cotsmaxdupreqs	126

5 Internet Protocol Suite Tunable Parameters	129
Where to Find Tunable Parameter Information	129
Overview of Tuning IP Suite Parameters	129
IP Suite Parameter Validation	130
Internet Request for Comments (RFCs)	130
IP Tunable Parameters	130
_icmp_err_interval and _icmp_err_burst	130
_respond_to_echo_broadcast and _respond_to_echo_multicast (ipv4 or ipv6) ..	131
_send_redirects (ipv4 or ipv6)	131
forwarding (ipv4 or ipv6)	132
ttl	132
hoplimit (ipv6)	132
_addr_per_if	133
hostmodel (ipv4 or ipv6)	133
ip_queue_fanout	134
IP Tunable Parameters With Additional Cautions	135
TCP Tunable Parameters	136
_deferred_ack_interval	136
_local_dack_interval	137
_deferred_acks_max	137
_local_dacks_max	138
_wscale_always	138
_tstamp_always	139
send_buf	139
recv_buf	140
max_buf	140
_cwnd_max	141
_slow_start_initial	141
_local_slow_start_initial	142
_slow_start_after_idle	142
sack	142
_rev_src_routes	143
_time_wait_interval	143
ecn	144
_conn_req_max_q	145
_conn_req_max_q0	146

_conn_req_min	146
_rst_sent_rate_enabled	147
_rst_sent_rate	147
TCP/IP Parameters Set in the /etc/system File	149
TCP Parameters With Additional Cautions	150
UDP Tunable Parameters	154
send_buf	154
recv_buf	154
max_buf	155
smallest_anon_port	155
largest_anon_port	156
IPQoS Tunable Parameter	156
_policy_mask	156
SCTP Tunable Parameters	157
_max_init_retr	157
_pa_max_retr	157
_pp_max_retr	158
_cwnd_max	158
_ipv4_ttl	159
_heartbeat_interval	159
_new_secret_interval	160
_initial_mtu	160
_deferred_ack_interval	161
_ignore_path_mtu	161
_initial_ssthresh	161
send_buf	162
_xmit_lowat	162
recv_buf	162
max_buf	163
_rto_min	163
_rto_max	164
_rto_initial	164
_cookie_life	164
_max_in_streams	165
_initial_out_streams	165
_shutack_wait_bound	165

_maxburst	166
_addip_enabled	166
_prsctp_enabled	166
smallest_anon_port	167
largest_anon_port	167
Per-Route Metrics	168
6 System Facility Parameters	171
System Default Parameters	172
autofs	172
cron	172
devfsadm	172
dhcpgent	172
fs	173
ftp	173
inetinit	173
init	173
ipsec	174
kbd	174
keyserv	174
login	175
mpathd	175
nfs	175
nfslogd	175
nss	175
passwd	175
su	176
syslog	176
tar	176
telnetd	176
utmpd	176
A Tunable Parameters Change History	177
Kernel Parameters	177
General Kernel and Memory Parameters (Oracle Solaris 11 and 11.1)	177

Paging-Related Parameters (Oracle Solaris 11)	178
Process-Sizing Tunables (Oracle Solaris 11.1)	178
Swapping-Related Parameter (Oracle Solaris 11.1)	179
General File System Parameter (Oracle Solaris 11.1)	179
General Driver Parameter (Oracle Solaris 11)	179
Network Driver Parameters (Oracle Solaris 11)	179
TCP/IP Tunable Parameters (Oracle Solaris 11 and 11.1)	180
[tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port	180
_local_slow_start_initial	180
IP Parameter Name Changes (Oracle Solaris 11)	180
TCP Parameter Name Changes (Oracle Solaris 11)	181
UDP Parameter Name Changes (Oracle Solaris 11)	183
SCTP Parameter Name Changes (Oracle Solaris 11)	183
Parameters That Are Obsolete or Have Been Removed (Oracle Solaris 11 and 11.1)	184
NCA Parameters	185
consistent_coloring	185
rstchown	185
Obsolete TCP/IP Module Parameters (Oracle Solaris 11)	186
B Revision History for This Manual	187
Current Version: Oracle Solaris 11.1 Release	187
New or Changed Parameters in the Oracle Solaris Release	187
Index	189

Preface

The Oracle Solaris 11.1 Tunable Parameters Reference Manual provides reference information about Oracle Solaris OS kernel and network tunable parameters. This manual does not provide tunable parameter information about desktop systems or Java environments.

This manual contains information for both SPARC based and x86 based systems.

Note – This Oracle Solaris release supports systems that use the SPARC and x86 families of processor architectures. The supported systems appear in the *Oracle Solaris Hardware Compatibility List* at <http://www.oracle.com/webfolder/technetwork/hcl/index.html>. This document cites any implementation differences between the platform types.

Who Should Use This Book

This book is intended for experienced Oracle Solaris system administrators who might need to change kernel tunable parameters in certain situations. For guidelines on changing Oracle Solaris tunable parameters, refer to “[Tuning an Oracle Solaris System](#)” on page 19.

How This Book Is Organized

The following table describes the chapters and appendixes in this book.

Chapter	Description
Chapter 1, “Overview of Oracle Solaris System Tuning”	An overview of tuning an Oracle Solaris system. Also provides a description of the format used in the book to describe the kernel tunables.
Chapter 2, “Oracle Solaris Kernel Tunable Parameters”	A description of Oracle Solaris kernel tunables such as kernel memory, file system, process size, and paging parameters.
Chapter 3, “Oracle Solaris ZFS Tunable Parameters”	A description of Oracle Solaris ZFS tunable parameters. This chapter also includes ZFS tuning information for database products.
Chapter 4, “NFS Tunable Parameters”	A description of NFS tunables such as caching symbolic links, dynamic retransmission, and RPC security parameters.

Chapter	Description
Chapter 5, “Internet Protocol Suite Tunable Parameters”	A description of TCP/IP tunables such as IP forwarding, source routing, and buffer-sizing parameters.
Chapter 6, “System Facility Parameters”	A description of parameters used to set default values of certain system facilities. Changes are made by modifying files in the <code>/etc/default</code> directory.
Appendix A, “Tunable Parameters Change History”	A history of parameters that have changed or are now obsolete.
Appendix B, “Revision History for This Manual”	A history of this manual's revisions including the current Oracle Solaris release.

Other Resources for Oracle Solaris Tuning Information

This table describes other resources for Oracle Solaris tuning information.

Tuning Resource	For More Information
In-depth technical white papers	http://www.oracle.com/technetwork/server-storage/solaris/overview/index.html

Access to Oracle Support

Oracle customers have access to electronic support through My Oracle Support. For information, visit <http://www.oracle.com/pls/topic/lookup?ctx=acc&id=info> or visit <http://www.oracle.com/pls/topic/lookup?ctx=acc&id=trs> if you are hearing impaired.

Typographic Conventions

The following table describes the typographic conventions that are used in this book.

TABLE P-1 Typographic Conventions

Typeface	Description	Example
AaBbCc123	The names of commands, files, and directories, and onscreen computer output	Edit your <code>.login</code> file. Use <code>ls -a</code> to list all files. <code>machine_name% you have mail.</code>

TABLE P-1 Typographic Conventions (Continued)

Typeface	Description	Example
AaBbCc123	What you type, contrasted with onscreen computer output	machine_name% su Password:
<i>aabbcc123</i>	Placeholder: replace with a real name or value	The command to remove a file is <i>rm filename</i> .
<i>AaBbCc123</i>	Book titles, new terms, and terms to be emphasized	Read Chapter 6 in the <i>User's Guide</i> . <i>A cache</i> is a copy that is stored locally. Do <i>not</i> save the file. Note: Some emphasized items appear bold online.

Shell Prompts in Command Examples

The following table shows UNIX system prompts and superuser prompts for shells that are included in the Oracle Solaris OS. In command examples, the shell prompt indicates whether the command should be executed by a regular user or a user with privileges.

TABLE P-2 Shell Prompts

Shell	Prompt
Bash shell, Korn shell, and Bourne shell	\$
Bash shell, Korn shell, and Bourne shell for superuser	#
C shell	machine_name%
C shell for superuser	machine_name#

Overview of Oracle Solaris System Tuning

This section provides overview information about the format of the tuning information in this manual. This section also describes the different ways to tune an Oracle Solaris system.

- “What's New in Oracle Solaris System Tuning?” on page 17
- “Tuning an Oracle Solaris System” on page 19
- “Tuning Format of Tunable Parameters Descriptions” on page 19
- “Tuning the Oracle Solaris Kernel” on page 21
- “Special Oracle Solaris tune and var Structures” on page 23
- “Viewing Oracle Solaris System Configuration Information” on page 24
- “kstat Utility” on page 24

What's New in Oracle Solaris System Tuning?

This section describes new or changed parameters in the Oracle Solaris 11.1 release.

- Oracle Solaris 11.1: Oracle Solaris ZFS tunable information is provided in [Chapter 3](#), “Oracle Solaris ZFS Tunable Parameters.”
- Oracle Solaris 11.1: The `maxusers`, `max_nprocs`, `ngroups_max`, `pidmax`, and `segkpsize` parameters have been revised. For more information, see [Chapter 2](#), “Oracle Solaris Kernel Tunable Parameters.”
- Oracle Solaris 11.1: The `dnlc_dircache_percent` parameter is new. For more information, see “[dnlc_dircache_percent](#)” on page 65.
- Oracle Solaris 11: The `rstchown` parameter that was previously set in the `/etc/system` file is obsolete. If you set this parameter in the `/etc/system` file, the following error message is displayed:

```
sorry, variable 'rstchown' is not defined in the 'kernel'
```

This parameter has been replaced by the ZFS `rstchown` file system property and a general file system mount option. For more information, see [Oracle Solaris 11.1 Administration: ZFS File Systems](#) and `mount(1M)`.

- Oracle Solaris 11: The following system configuration parameters that were previously configured by editing files in the `/etc/default` directory have changed to SMF services:
 - `autofs`
 - `init`
 - `kbd`
 - `nfs`

For information about changing SMF service properties, see [Chapter 6, “System Facility Parameters.”](#)

- Oracle Solaris 11: The `ipadm` command replaces the `nnd` command for setting network properties. TCP, IP, UDP, and SCTP properties are set as follows:

```
ipadm set-prop -p parameter ip|ipv4|ipv6|tcp|udp|sctp
```

In addition, most of the network tunable names have changed slightly to correlate better with the `ipadm` format.

For more information, see [“Overview of Tuning IP Suite Parameters” on page 129.](#)

- Oracle Solaris 11: This release includes the `disp_rechoose_interval` parameter. For more information, see [“disp_rechoose_interval” on page 73.](#)
- Oracle Solaris 11: This release includes the `ngroups_max` parameter description. For more information, see [“ngroups_max” on page 38.](#)
- Oracle Solaris 11: This release includes the `zfs_arc_min` and `zfs_arc_max` parameter descriptions. For more information, see [“zfs_arc_min” on page 82](#) and [“zfs_arc_max” on page 83.](#)
- Oracle Solaris 11: This release includes several `igb` and `ixgbe` network driver parameters. For more information, see [“igb Parameters” on page 56](#) and [“ixgbe Parameters” on page 57.](#)
- Oracle Solaris 11: This release includes the `ddi_msix_alloc_limit` parameter that can be used to increase the number of MSI-X interrupts that a device instance can allocate. For more information, see [“ddi_msix_alloc_limit” on page 55.](#)
- Oracle Solaris 11: This release includes the `kmem_stackinfo` parameter, which can be enabled to monitor kernel thread stack usage. For more information, see [“kmem_stackinfo” on page 53.](#)
- Oracle Solaris 11: Memory locality group parameters are provided in this release. For more information about these parameters, see [“Locality Group Parameters” on page 78.](#)

Tuning an Oracle Solaris System

The Oracle Solaris OS is a multi-threaded, scalable UNIX operating system that runs on SPARC and x86 processors. It is self-adjusting to system load and demands minimal tuning. In some cases, however, tuning is necessary. This book provides details about the officially supported kernel tuning options available for the Oracle Solaris OS.

The Solaris kernel is composed of a core portion, which is always loaded, and a number of loadable modules that are loaded as references are made to them. Many variables referred to in the kernel portion of this guide are in the core portion. However, a few variables are located in loadable modules.

A key consideration in system tuning is that setting system parameters (or system variables) is often the least effective action that can be done to improve performance. Changing the behavior of the application is generally the most effective tuning aid available. Adding more physical memory and balancing disk I/O patterns are also useful. In a few rare cases, changing one of the variables described in this guide will have a substantial effect on system performance.

Remember that one system's `/etc/system` settings might not be applicable, either wholly or in part, to another system's environment. Carefully consider the values in the file with respect to the environment in which they will be applied. Make sure that you understand the behavior of a system before attempting to apply changes to the system variables that are described here.

We recommend that you start with an empty `/etc/system` file when moving to a new Oracle Solaris release. As a first step, add only those tunables that are required by in-house or third-party applications. After baseline testing has been established, evaluate system performance to determine if additional tunable settings are required.



Caution – The tunable parameters described in this book can and do change from Oracle Solaris release to Oracle Solaris release. Publication of these tunable parameters does not preclude changes to the tunable parameters and their descriptions without notice.

Tuning Format of Tunable Parameters Descriptions

The format for the description of each tunable parameter is as follows:

- Parameter Name
- Description
- Data Type
- Default
- Range
- Units
- Dynamic?
- Validation

- Implicit
- When to Change
- Zone Configuration
- Commitment Level
- Change History

Parameter Name Is the exact name that is typed in the `/etc/system` file, or found in the `/etc/default/facility` file.

Most parameters names are of the form *parameter* where the parameter name does not contain a colon (:). These names refer to variables in the core portion of the kernel. If the name does contain a colon, the characters to the left of the colon reference the name of a loadable module. The name of the parameter within the module consists of the characters to the right of the colon. For example:

module_name:variable

Description

Briefly describes what the parameter does or controls.

Data Type

Indicates the signed or unsigned short integer or long integer. A long integer is twice the width in bits as an integer. For example, an unsigned integer = 32 bits, an unsigned long integer = 64 bits.

Units

(Optional) Describes the unit type.

Default

What the system uses as the default value.

Range

Specifies the possible range allowed by system validation or the bounds of the data type.

- MAXINT – A shorthand description for the maximum value of a signed integer (2,147,483,647)
- MAXUINT – A shorthand description for the maximum value of an unsigned integer (4,294,967,295)

Dynamic?

Yes, if the parameter can be changed on a running system with the `mdb` or `kldb` debugger. No, if the parameter is a boot time initialization only.

Validation

Checks that the system applies to the value of the variable either as specified in the `/etc/system` file or the default value, as well as when the validation is applied.

Implicit

(Optional) Provides unstated constraints that might exist on the parameter, especially in relation to other parameters.

When to Change

Explains why someone might want to change this value. Includes error messages or return codes.

Zone Configuration	Identifies whether the parameter can be set in a exclusive-IP zone or must be set in the global zone. None of the parameters can be set in shared-IP zones.
Commitment Level	Identifies the stability of the interface. Many of the parameters in this manual are still evolving and are classified as unstable. For more information, see attributes(5) .
Change History	(Optional) Contains a link to the Change History appendix, if applicable.

Tuning the Oracle Solaris Kernel

The following table describes the different ways tunable parameters can be applied.

Apply Tunable Parameters in These Ways	For More Information
Modify the <code>/etc/system</code> file	“/etc/system File” on page 21
Use the kernel debugger (<code>kldb</code>)	“kldb Command” on page 22
Use the modular debugger (<code>mdb</code>)	“mdb Command” on page 22
Use the <code>ipadm</code> command to set TCP/IP parameters	Chapter 5, “Internet Protocol Suite Tunable Parameters”
Modify the <code>/etc/default</code> files	Chapter 6, “System Facility Parameters”

`/etc/system` File

The `/etc/system` file provides a static mechanism for adjusting the values of kernel parameters. Values specified in this file are read at boot time and are applied. Any changes that are made to the file are not applied to the operating system until the system is rebooted.

One pass is made to set all the values before the configuration parameters are calculated.

Example-Setting a Parameter in `/etc/system`

The following `/etc/system` entry sets the ZFS ARC maximum (`zfs_arc_max`) to 30 GB.

```
set zfs:zfs_arc_max = 0x78000000
```

Recovering From an Incorrect Value

Make a copy of the `/etc/system` file before modifying it so that you can easily recover from incorrect value. For example:

```
# cp /etc/system /etc/system.good
```

If a value specified in the `/etc/system` file causes the system to become unbootable, you can recover with the following command:

```
ok boot -a
```

This command causes the system to ask for the name of various files used in the boot process. Press the Return key to accept the default values until the name of the `/etc/system` file is requested. When the Name of system file [`/etc/system`]: prompt is displayed, type the name of the good `/etc/system` file or `/dev/null`:

```
Name of system file [/etc/system]: /etc/system.good
```

If `/dev/null` is specified, this path causes the system to attempt to read from `/dev/null` for its configuration information. Because this file is empty, the system uses the default values. After the system is booted, the `/etc/system` file can be corrected.

For more information on system recovery, see [Oracle Solaris Administration: Common Tasks](#).

kldb Command

`kldb` is a interactive kernel debugger with the same general syntax as `mdb`. An advantage of interactive kernel debugger is that you can set breakpoints. When a breakpoint is reached, you can examine data or step through the execution of kernel code.

`kldb` can be loaded and unloaded on demand. You do not have to reboot the system to perform interactive kernel debugging, as was the case with `kadb`.

For more information, see [kldb\(1\)](#).

mdb Command

The modular debugger, `mdb`, is unique among Solaris debuggers because it is easily extensible. A programming API is available that allows compilation of modules to perform desired tasks within the context of the debugger.

`mdb` also includes a number of desirable usability features, including command-line editing, command history, built-in output pager, syntax checking, and command pipelining. `mdb` is the recommended post-mortem debugger for the kernel.

For more information, see [mdb\(1\)](#).

Example—Using mdb to Display Information

Display a high-level view of a system's memory usage. For example:

```
# mdb -k
Loading modules: [ unix genunix specfs dtrace mac cpu.generic cpu_ms.AuthenticAMD.15
uppc pcplusmp scsi_vhci zfs mpt sd ip hook neti arp usba sockfs kssl qlc fctl stmf stmf_
sbd md lofs random idm fcp crypto cpc smbsrv nfs fcip sPPP ufs logindmux ptm nsmb scu
mpt_sas pmcs emlxs ]
> ::memstat
Page Summary
-----
Kernel                160876          628    16%
ZFS File Data         303401         1185    30%
Anon                   25335           98     2%
Exec and libs          1459            5     0%
Page cache             5083            19     1%
Free (cachelist)       6616            25     1%
Free (freelist)       510870         1995   50%

Total                  1013640         3959
Physical               1013639         3959
> $q
```

For more information on using the modular debugger, see the [Oracle Solaris Modular Debugger Guide](#).

When using either `kmdb` or `mdb` debugger, the module name prefix is not required. After a module is loaded, its symbols form a common name space with the core kernel symbols and any other previously loaded module symbols.

Special Oracle Solaris tune and var Structures

Oracle Solaris tunable parameters come in a variety of forms. The tune structure defined in the `/usr/include/sys/tuneable.h` file is the runtime representation of `tune_t_fsflushr`, `tune_t_minarmem`, and `tune_t_flkrec`. After the kernel is initialized, all references to these variables are found in the appropriate field of the tune structure.

The proper way to set parameters for this structure at boot time is to initialize the special parameter that corresponds to the desired field name. The system initialization process then loads these values into the tune structure.

A second structure into which various tunable parameters are placed is the `var` structure named `v`. You can find the definition of a `var` structure in the `/usr/include/sys/var.h` file. The runtime representation of variables such as `autoup` and `bufhwm` is stored here.

Do not change either the `tune` or `v` structure on a running system. Changing any field in these structures on a running system might cause the system to panic.

Viewing Oracle Solaris System Configuration Information

Several tools are available to examine system configuration information. Some tools require superuser privilege. Other tools can be run by a non-privileged user. Every structure and data item can be examined with the kernel debugger by using `mdb` on a running system or by booting under `kmdb`.

For more information, see [mdb\(1\)](#) or [kadb\(1M\)](#).

sysdef Command

The `sysdef` command provides the values of memory and process resource limits, and portions of the `tune` and `v` structures. For example, the `sysdef` “Tunable Parameters” section from a SPARC T3-4 system with 500 GB of memory is as follows:

```
2206203904    maximum memory allowed in buffer cache (bufhwm)
  65546      maximum number of processes (v.v_proc)
    99      maximum global priority in sys class (MAXCLSYSPRI)
  65541      maximum processes per user id (v.v_maxup)
    30      auto update time limit in seconds (NAUTOUP)
    25      page stealing low water mark (GPGSLO)
    1       fsflush run rate (FSFLUSHR)
    25      minimum resident memory for avoiding deadlock (MINARMEM)
    25      minimum swapable memory for avoiding deadlock (MINASMEM)
```

For more information, see [sysdef\(1M\)](#).

kstat Utility

`kstats` are data structures maintained by various kernel subsystems and drivers. They provide a mechanism for exporting data from the kernel to user programs without requiring that the program read kernel memory or have superuser privilege. For more information, see [kstat\(1M\)](#) or [kstat\(3KSTAT\)](#).

Oracle Solaris Kernel Tunable Parameters

This chapter describes most of the Oracle Solaris kernel tunable parameters.

- “General Kernel and Memory Parameters” on page 26
- “fsflush and Related Parameters” on page 30
- “Process-Sizing Parameters” on page 34
- “Paging-Related Parameters” on page 39
- “Swapping-Related Parameters” on page 50
- “Kernel Memory Allocator” on page 52
- “General Driver Parameters” on page 54
- “Network Driver Parameters” on page 56
- “General I/O Parameters” on page 61
- “General File System Parameters” on page 63
- “TMPFS Parameters” on page 66
- “Pseudo Terminals” on page 67
- “STREAMS Parameters” on page 70
- “System V Message Queues” on page 71
- “System V Semaphores” on page 72
- “System V Shared Memory” on page 72
- “Scheduling” on page 73
- “Timers” on page 74
- “SPARC System Specific Parameters” on page 75
- “Locality Group Parameters” on page 78

Where to Find Tunable Parameter Information

Tunable Parameter	For Information
Oracle Solaris ZFS tunables parameters	Chapter 3, “Oracle Solaris ZFS Tunable Parameters”
NFS tunable parameters	Chapter 4, “NFS Tunable Parameters”

Tunable Parameter	For Information
Internet Protocol Suite tunable parameters	Chapter 5, “Internet Protocol Suite Tunable Parameters”

General Kernel and Memory Parameters

This section describes general kernel parameters that are related to physical memory and stack configuration. The ZFS-related memory parameters have moved to [Chapter 3, “Oracle Solaris ZFS Tunable Parameters.”](#)

physmem

Description	Modifies the system's configuration of the number of physical pages of memory after the Oracle Solaris OS and firmware are accounted for.
Data Type	Unsigned long
Default	Number of usable pages of physical memory available on the system, not counting the memory where the core kernel and data are stored
Range	1 to amount of physical memory on system
Units	Pages
Dynamic?	No
Validation	None
When to Change	Whenever you want to test the effect of running the system with less physical memory. Because this parameter does <i>not</i> take into account the memory used by the core kernel and data, as well as various other data structures allocated early in the startup process, the value of <code>physmem</code> should be less than the actual number of pages that represent the smaller amount of memory.
Commitment Level	Unstable

default_stksize

Description	Specifies the default stack size of all threads. No thread can be created with a stack size smaller than <code>default_stksize</code> . If <code>default_stksize</code> is set, it overrides <code>lwp_default_stksize</code> . See also “<code>lwp_default_stksize</code>” on page 27.
Data Type	Integer

Default	<ul style="list-style-type: none"> ▪ 3 x PAGESIZE on SPARC systems ▪ 5 x PAGESIZE on x64 systems
Range	<p>Minimum is the default values:</p> <ul style="list-style-type: none"> ▪ 3 x PAGESIZE on SPARC systems ▪ 5 x PAGESIZE on x64 systems <p>Maximum is 32 times the default value.</p>
Units	Bytes in multiples of the value returned by the <code>getpagesize</code> parameter. For more information, see getpagesize(3C) .
Dynamic?	Yes. Affects threads created after the variable is changed.
Validation	<p>Must be greater than or equal to 8192 and less than or equal to 262,144 (256 x 1024). Also must be a multiple of the system page size. If these conditions are not met, the following message is displayed:</p> <pre>Illegal stack size, Using N</pre> <p>The value of <i>N</i> is the default value of <code>default_stksize</code>.</p>
When to Change	<p>When the system panics because it has run out of stack space. The best solution for this problem is to determine why the system is running out of space and then make a correction.</p> <p>Increasing the default stack size means that almost every kernel thread will have a larger stack, resulting in increased kernel memory consumption for no good reason. Generally, that space will be unused. The increased consumption means other resources that are competing for the same pool of memory will have the amount of space available to them reduced, possibly decreasing the system's ability to perform work. Among the side effects is a reduction in the number of threads that the kernel can create. This solution should be treated as no more than an interim workaround until the root cause is remedied.</p>
Commitment Level	Unstable

`lwp_default_stksize`

Description	Specifies the default value of the stack size to be used when a kernel thread is created, and when the calling routine does not provide an explicit size to be used.
Data Type	Integer

Default	<ul style="list-style-type: none"> ▪ 32,768 for SPARC platforms ▪ 20,480 for x64 platforms
Range	<p>Minimum is the default values:</p> <ul style="list-style-type: none"> ▪ 3 x PAGESIZE on SPARC systems ▪ 5 x PAGESIZE on x64 systems <p>Maximum is 32 times the default value.</p>
Units	Bytes in multiples of the value returned by the <code>getpagesize</code> parameter. For more information, see getpagesize(3C) .
Dynamic?	Yes. Affects threads created after the variable is changed.
Validation	<p>Must be greater than or equal to 8192 and less than or equal to 262,144 (256 x 1024). Also must be a multiple of the system page size. If these conditions are not met, the following message is displayed:</p> <pre>Illegal stack size, Using N</pre> <p>The value of <i>N</i> is the default value of <code>lwp_default_stksize</code>.</p>
When to Change	<p>When the system panics because it has run out of stack space. The best solution for this problem is to determine why the system is running out of space and then make a correction.</p> <p>Increasing the default stack size means that almost every kernel thread will have a larger stack, resulting in increased kernel memory consumption for no good reason. Generally, that space will be unused. The increased consumption means other resources that are competing for the same pool of memory will have the amount of space available to them reduced, possibly decreasing the system's ability to perform work. Among the side effects is a reduction in the number of threads that the kernel can create. This solution should be treated as no more than an interim workaround until the root cause is remedied.</p>
Commitment Level	Unstable

logevent_max_q_sz

Description	Maximum number of system events allowed to be queued and waiting for delivery to the <code>syseventd</code> daemon. Once the size of the system event queue reaches this limit, no other system events are allowed on the queue.
Data Type	Integer

Default	5000
Range	0 to MAXINT
Units	System events
Dynamic?	Yes
Validation	The system event framework checks this value every time a system event is generated by <code>ddi_log_sysevent</code> and <code>sysevent_post_event</code> . For more information, see ddi_log_sysevent(9F) and sysevent_post_event(3SYSEVENT) .
When to Change	When error log messages indicate that a system event failed to be logged, generated, or posted.
Commitment Level	Unstable

segkpsize

Description	Specifies the amount of kernel pageable memory available. This memory is used primarily for kernel thread stacks. Increasing this number allows either larger stacks for the same number of threads or more threads.
Data Type	Unsigned long
Default	2 GB x the smaller result of <code>nCPUs / 128</code> or the amount of physical memory / 256 GB
Range	512 MB to 64 GB (SPARC) 200 MB to 8 GB (x86)
Units	Pages
Dynamic?	No
Validation	Value is compared to minimum and maximum sizes. If smaller than the minimum or larger than the maximum, it is reset to 2 GB. A message to that effect is displayed. On SPARC systems, the <code>segkpsize</code> value cannot exceed twice the size of physical memory. On x86 systems, the value cannot exceed the size of physical memory.
When to Change	Required to support large numbers of processes on a system. The default size of 2 GB allows creation of 24-KB stacks for more than

65,536 kernel threads. If more than this number is needed or the stack size needs to be increased, `segkpsize` can be increased, assuming sufficient physical memory exists.

Commitment Level Unstable

Change History For information, see “[segkpsize](#)” on page 178.

noexec_user_stack

Description Enables the stack to be marked as nonexecutable, which helps make buffer-overflow attacks more difficult.

An Oracle Solaris system running a 64-bit kernel makes the stacks of all 64-bit applications nonexecutable by default. Setting this parameter is necessary to make 32-bit applications nonexecutable.

Data Type Signed integer

Default 0 (disabled)

Range 0 (disabled) or 1 (enabled)

Units Toggle (on/off)

Dynamic? Yes. Does not affect currently running processes, only processes created after the value is set.

Validation None

When to Change Should be enabled at all times unless applications are deliberately placing executable code on the stack without using `mprotect` to make the stack executable. For more information, see [mprotect\(2\)](#).

Commitment Level Unstable

fsflush and Related Parameters

This section describes `fsflush` and related tunables.

fsflush

The system daemon, `fsflush`, runs periodically to do three main tasks:

1. On every invocation, `fsflush` flushes dirty file system pages over a certain age to disk.

2. On every invocation, `fsflush` examines a portion of memory and causes modified pages to be written to their backing store. Pages are written if they are modified and if they do not meet one of the following conditions:

- Pages are kernel page
- Pages are free
- Pages are locked
- Pages are associated with a swap device
- Pages are currently involved in an I/O operation

The net effect is to flush pages from files that are mapped with `mmap` with write permission and that have actually been changed.

Pages are flushed to backing store but left attached to the process using them. This will simplify page reclamation when the system runs low on memory by avoiding delay for writing the page to backing store before claiming it, if the page has not been modified since the flush.

3. `fsflush` writes file system metadata to disk. This write is done every n th invocation, where n is computed from various configuration variables. See “`tune_t_fsflushr`” on page 31 and “`autoup`” on page 32 for details.

The following features are configurable:

- Frequency of invocation (`tune_t_fsflushr`)
- Whether memory scanning is executed (`dopageflush`)
- Whether file system data flushing occurs (`doiflush`)
- The frequency with which file system data flushing occurs (`autoup`)

For most systems, memory scanning and file system metadata synchronizing are the dominant activities for `fsflush`. Depending on system usage, memory scanning can be of little use or consume too much CPU time.

`tune_t_fsflushr`

Description	Specifies the number of seconds between <code>fsflush</code> invocations
Data Type	Signed integer
Default	1
Range	1 to MAXINT
Units	Seconds
Dynamic?	No

Validation	If the value is less than or equal to zero, the value is reset to 1 and a warning message is displayed. This check is done only at boot time.
When to Change	See the autoup parameter.
Commitment Level	Unstable

autoup

Description	<p>Along with <code>tune_t_flushr</code>, <code>autoup</code> controls the amount of memory examined for dirty pages in each invocation and frequency of file system synchronizing operations.</p> <p>The value of <code>autoup</code> is also used to control whether a buffer is written out from the free list. Buffers marked with the <code>B_DELWRI</code> flag (which identifies file content pages that have changed) are written out whenever the buffer has been on the list for longer than <code>autoup</code> seconds. Increasing the value of <code>autoup</code> keeps the buffers in memory for a longer time.</p>
Data Type	Signed integer
Default	30
Range	1 to MAXINT
Units	Seconds
Dynamic?	No
Validation	If <code>autoup</code> is less than or equal to zero, it is reset to 30 and a warning message is displayed. This check is done only at boot time.
Implicit	<p><code>autoup</code> should be an integer multiple of <code>tune_t_fsflushr</code>. At a minimum, <code>autoup</code> should be at least 6 times the value of <code>tune_t_fsflushr</code>. If not, excessive amounts of memory are scanned each time <code>fsflush</code> is invoked.</p> <p>The total system pages multiplied by <code>tune_t_fsflushr</code> should be greater than or equal to <code>autoup</code> to cause memory to be checked if <code>dopageflush</code> is non-zero.</p>
When to Change	<p>Here are several potential situations for changing <code>autoup</code>, <code>tune_t_fsflushr</code>, or both:</p> <ul style="list-style-type: none">Systems with large amounts of memory – In this case, increasing <code>autoup</code> reduces the amount of memory scanned in each invocation of <code>fsflush</code>.

- Systems with minimal memory demand – Increasing both `autoup` and `tune_t_fsflushr` reduces the number of scans made. `autoup` should be increased also to maintain the current ratio of `autoup / tune_t_fsflushr`.
- Systems with large numbers of transient files (for example, mail servers or software build machines) – If large numbers of files are created and then deleted, `fsflush` might unnecessarily write data pages for those files to disk.

Commitment Level Unstable

dopageflush

Description	Controls whether memory is examined for modified pages during <code>fsflush</code> invocations. In each invocation of <code>fsflush</code> , the number of physical memory pages in the system is determined. This number might have changed because of a dynamic reconfiguration operation. Each invocation scans by using this algorithm: total number of pages x <code>tune_t_fsflushr / autoup</code> pages
Data Type	Signed integer
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Units	Toggle (on/off)
Dynamic?	Yes
Validation	None
When to Change	If the system page scanner rarely runs, which is indicated by a value of 0 in the <code>sr</code> column of <code>vmstat</code> output.
Commitment Level	Unstable

doiflush

Description	Controls whether file system metadata syncs will be executed during <code>fsflush</code> invocations. This synchronization is done every N th invocation of <code>fsflush</code> where $N = (\text{autoup} / \text{tune_t_fsflushr})$. Because this algorithm is integer division, if <code>tune_t_fsflushr</code> is greater than <code>autoup</code> , a synchronization is done on every invocation of <code>fsflush</code> because the code checks to see if its iteration counter is greater
-------------	--

than or equal to N . Note that N is computed once on invocation of `fsflush`. Later changes to `tune_t_fsflushr` or `autoup` have no effect on the frequency of synchronization operations.

Data Type	Signed integer
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Units	Toggle (on/off)
Dynamic?	Yes
Validation	None
When to Change	When files are frequently modified over a period of time and the load caused by the flushing perturbs system behavior.
	Files whose existence, and therefore consistency of state, does not matter if the system reboots are better kept in a TMPFS file system (for example, <code>/tmp</code>). Inode traffic can be reduced on systems by using the <code>mount -noatime</code> option. This option eliminates inode updates when the file is accessed.
	For a system engaged in realtime processing, you might want to disable this option and use explicit application file synchronizing to achieve consistency.
Commitment Level	Unstable

Process-Sizing Parameters

Several parameters (or variables) are used to control the number of processes that are available on the system and the number of processes that an individual user can create. The foundation parameter is `maxusers`. This parameter drives the values assigned to `max_nprocs` and `maxuprc`.

maxusers

Description	Originally, <code>maxusers</code> defined the number of logged in users the system could support. When a kernel was generated, various tables were sized based on this setting. Current Oracle Solaris releases do much of its sizing based on the amount of memory on the system. Thus, much of the past use of <code>maxusers</code> has changed. A number of subsystems that are still derived from <code>maxusers</code> :
-------------	--

	<ul style="list-style-type: none"> ▪ The maximum number of processes on the system ▪ The number of quota structures held in the system ▪ The size of the directory name look-up cache (DNLC)
Data Type	Signed integer
Default	Lesser of the amount of memory in MB or 2048, and the greater of that value and nCPUs x 8
Range	1 to the greater of 2048 or nCPUs x 8, based on the size of physical memory, if not set in the <code>/etc/system</code> file
	1 to the greater of 4096 or the nCPUs x 8, if set in the <code>/etc/system</code> file
Units	Users
Dynamic?	No. After computation of dependent parameters is done, <code>maxusers</code> is never referenced again.
Validation	If the value is greater than the maximum allowed, it is reset to the maximum. A message to that effect is displayed.
When to Change	<p>When the default number of user processes derived by the system is too low. This situation is evident when the following message displays on the system console:</p> <pre>out of processes</pre> <p>You might also change this parameter when the default number of processes is too high, as in these situations:</p> <ul style="list-style-type: none"> ▪ Database servers that have a lot of memory and relatively few running processes can save system memory when the default value of <code>maxusers</code> is reduced. ▪ If file servers have a lot of memory and few running processes, you might reduce this value. However, you should explicitly set the size of the DNLC. See “ncsize” on page 63.
Commitment Level	Unstable
Change History	For information, see “maxusers” on page 178 .

reserved_procs

Description	Specifies the number of system process slots to be reserved in the process table for processes with a UID of root (0). For example, <code>fsflush</code> has a UID of root (0).
Data Type	Signed integer

Default	5
Range	5 to MAXINT
Units	Processes
Dynamic?	No. Not used after the initial parameter computation.
Validation	Any <code>/etc/system</code> setting is honored.
Commitment Level	Unstable
When to Change	Consider increasing to 10 + the normal number of UID 0 (root) processes on system. This setting provides some cushion should it be necessary to obtain a root shell when the system is otherwise unable to create user-level processes.

pidmax

Description	<p>Specifies the value of the largest possible process ID.</p> <p><code>pidmax</code> sets the value for the <code>maxpid</code> variable. Once <code>maxpid</code> is set, <code>pidmax</code> is ignored. <code>maxpid</code> is used elsewhere in the kernel to determine the maximum process ID and for validation checking.</p> <p>Any attempts to set <code>maxpid</code> by adding an entry to the <code>/etc/system</code> file have no effect.</p>
Data Type	Signed integer
Default	3,000
Range	5 to 999,999
Units	Processes
Dynamic?	No. Used only at boot time to set the value of <code>pidmax</code> .
Validation	Yes. Value is compared to the value of <code>reserved_procs</code> and 999,999. If less than <code>reserved_procs</code> or greater than 999,999, the value is set to 999,999.
Implicit	<code>max_nprocs</code> range checking ensures that <code>max_nprocs</code> is always less than or equal to this value.
When to Change	Required to enable support for more than 30,000 processes on a system. See also “ max_nprocs ” on page 37.
Commitment Level	Unstable
Change History	For information, see “ pidmax ” on page 178.

max_nprocs

Description	<p>Specifies the maximum number of processes that can be created on a system. Includes system processes and user processes. Any value specified in <code>/etc/system</code> is used in the computation of <code>maxuprc</code>.</p> <p>This value is also used in determining the size of several other system data structures. Other data structures where this parameter plays a role are as follows:</p> <ul style="list-style-type: none"> ▪ Determining the size of the directory name lookup cache (if <code>ncsize</code> is not specified) ▪ Verifying that the amount of memory used by configured system V semaphores does not exceed system limits ▪ Configuring Hardware Address Translation resources for x86 platforms
Data Type	Signed integer
Default	<p>$10 + (16 \times \text{maxusers})$ if <code>maxusers</code> is set in the <code>/etc/system</code> file</p> <p>The larger of 30,000 or $10 + (128 \times \text{number of CPUs})$, if <code>maxusers</code> is not set in the <code>/etc/system</code> file</p>
Range	26 to value of <code>maxpid</code>
Dynamic?	No
Validation	Yes. If the value exceeds <code>maxpid</code> , it is set to <code>maxpid</code> .
When to Change	Changing this parameter is one of the steps necessary to enable support for more than 30,000 processes on a system.
Commitment Level	Unstable
Change History	For information, see “max_nprocs” on page 178 .

maxuprc

Description	Specifies the maximum number of processes that can be created on a system by any one user.
Data Type	Signed integer
Default	<code>max_nprocs - reserved_procs</code>
Range	1 to <code>max_nprocs - reserved_procs</code>
Units	Processes

Dynamic?	No
Validation	Yes. This value is compared to <code>max_nprocs - reserved_procs</code> and set to the smaller of the two values.
When to Change	When you want to specify a hard limit for the number of processes a user can create that is less than the default value of however many processes the system can create. Attempting to exceed this limit generates the following warning messages on the console or in the messages file: <code>out of per-user processes for uid N</code>
Commitment Level	Unstable

ngroups_max

Description	Specifies the maximum number of supplemental groups per process.
Data Type	Signed integer
Default	16
Range	0 to 1024
Units	Groups
Dynamic?	No
Validation	Yes. If <code>ngroups_max</code> is set to an invalid value, it is automatically reset to the closest legal value. For example, if it is set to less than zero, it is reset to 0. If it is set to greater than 1024, it is reset to 1024.
When to Change	Review the following considerations if you are using NFS AUTH_SYS authentication and you want to increase the default <code>ngroups_max</code> value: <ol style="list-style-type: none">1. If <code>ngroups_max</code> is set to 16 or if the client's AUTH_SYS credential that is provided has 15 or fewer groups, the client's group information is used.2. If <code>ngroups_max</code> is set to greater than 16 and the client's AUTH_SYS credential from the name server contains exactly 16 groups, the maximum allowed, the NFS server consults the name server and matches the client's UID to a user name. Then, the name server computes a list of groups to which the user belongs.
Commitment Level	Unstable
Change History	For information, see “ngroups_max” on page 178 .

Paging-Related Parameters

The Solaris OS uses a demand paged virtual memory system. As the system runs, pages are brought into memory as needed. When memory becomes occupied above a certain threshold and demand for memory continues, paging begins. Paging goes through several levels that are controlled by certain parameters.

The general paging algorithm is as follows:

- A memory deficit is noticed. The page scanner thread runs and begins to walk through memory. A two-step algorithm is employed:
 1. A page is marked as unused.
 2. If still unused after a time interval, the page is viewed as a subject for reclaim.

If the page has been modified, a request is made to the pageout thread to schedule the page for I/O. Also, the page scanner continues looking at memory. Pageout causes the page to be written to the page's backing store and placed on the free list. When the page scanner scans memory, no distinction is made as to the origin of the page. The page might have come from a data file, or it might represent a page from an executable's text, data, or stack.

- As memory pressure on the system increases, the algorithm becomes more aggressive in the pages it will consider as candidates for reclamation and in how frequently the paging algorithm runs. (For more information, see [“fastscan” on page 46](#) and [“slowscan” on page 47](#).) As available memory falls between the range `lotsfree` and `minfree`, the system linearly increases the amount of memory scanned in each invocation of the pageout thread from the value specified by `slowscan` to the value specified by `fastscan`. The system uses the `desfree` parameter to control a number of decisions about resource usage and behavior.

The system initially constrains itself to use no more than 4 percent of one CPU for pageout operations. As memory pressure increases, the amount of CPU time consumed in support of pageout operations linearly increases until a maximum of 80 percent of one CPU is consumed. The algorithm looks through some amount of memory between `slowscan` and `fastscan`, then stops when one of the following occurs:

- Enough pages have been found to satisfy the memory shortfall.
- The planned number of pages have been looked at.
- Too much time has elapsed.

If a memory shortfall is still present when pageout finishes its scan, another scan is scheduled for 1/4 second in the future.

The configuration mechanism of the paging subsystem was changed. Instead of depending on a set of predefined values for `fastscan`, `slowscan`, and `handspreadpages`, the system determines the appropriate settings for these parameters at boot time. Setting any of these parameters in the `/etc/system` file can cause the system to use less than optimal values.



Caution – Remove all tuning of the VM system from the `/etc/system` file. Run with the default settings and determine if it is necessary to adjust any of these parameters. Do not set either `cachefree` or `priority_paging`.

Dynamic reconfiguration (DR) for CPU and memory is supported. A system in a DR operation that involves the addition or deletion of memory recalculates values for the relevant parameters, unless the parameter has been explicitly set in `/etc/system`. In that case, the value specified in `/etc/system` is used, unless a constraint on the value of the variable has been violated. In this case, the value is reset.

lotsfree

Description	Serves as the initial trigger for system paging to begin. When this threshold is crossed, the page scanner wakes up to begin looking for memory pages to reclaim.
Data Type	Unsigned long
Default	The greater of 1/64th of physical memory or 512 KB
Range	The minimum value is 512 KB or 1/64th of physical memory, whichever is greater, expressed as pages using the page size returned by <code>getpagesize</code> . For more information, see getpagesize(3C) . The maximum value is the number of physical memory pages. The maximum value should be no more than 30 percent of physical memory. The system does not enforce this range, other than that described in the Validation section.
Units	Pages
Dynamic?	Yes, but dynamic changes are lost if a memory-based DR operation occurs.
Validation	If <code>lotsfree</code> is greater than the amount of physical memory, the value is reset to the default.
Implicit	The relationship of <code>lotsfree</code> being greater than <code>desfree</code> , which is greater than <code>minfree</code> , should be maintained at all times.
When to Change	When demand for pages is subject to sudden sharp spikes, the memory algorithm might be unable to keep up with demand. One workaround is to start reclaiming memory at an earlier time. This solution gives the paging system some additional margin.

A rule of thumb is to set this parameter to 2 times what the system needs to allocate in a few seconds. This parameter is workload dependent. A DBMS server can probably work fine with the default settings. However, you might need to adjust this parameter for a system doing heavy file system I/O.

For systems with relatively static workloads and large amounts of memory, lower this value. The minimum acceptable value is 512 KB, expressed as pages using the page size returned by `getpagesize`.

Commitment Level Unstable

desfree

Description	Specifies the preferred amount of memory to be free at all times on the system.
Data Type	Unsigned integer
Default	<code>lotsfree / 2</code>
Range	The minimum value is 256 KB or 1/128th of physical memory, whichever is greater, expressed as pages using the page size returned by <code>getpagesize</code> . The maximum value is the number of physical memory pages. The maximum value should be no more than 15 percent of physical memory. The system does not enforce this range other than that described in the Validation section.
Units	Pages
Dynamic?	Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided in the <code>/etc/system</code> file or calculated from the new physical memory value.
Validation	If <code>desfree</code> is greater than <code>lotsfree</code> , <code>desfree</code> is set to <code>lotsfree / 2</code> . No message is displayed.
Implicit	The relationship of <code>lotsfree</code> being greater than <code>desfree</code> , which is greater than <code>minfree</code> , should be maintained at all times.
Side Effects	Several side effects can arise from increasing the value of this parameter. When the new value nears or exceeds the amount of available memory on the system, the following can occur:

- Asynchronous I/O requests are not processed, unless available memory exceeds `des f ree`. Increasing the value of `des f ree` can result in rejection of requests that otherwise would succeed.
- NFS asynchronous writes are executed as synchronous writes.
- The swapper is awakened earlier, and the behavior of the swapper is biased towards more aggressive actions.
- The system might not preload (prefault) as many executable pages as possible into the system. This side effect results in applications potentially running slower than they otherwise would.

When to Change	For systems with relatively static workloads and large amounts of memory, lower this value. The minimum acceptable value is 256 KB, expressed as pages using the page size returned by <code>getpagesize</code> .
Commitment Level	Unstable

minfree

Description	Specifies the minimum acceptable memory level. When memory drops below this number, the system biases allocations toward allocations necessary to successfully complete pageout operations or to swap processes completely out of memory. Either allocation denies or blocks other allocation requests.
Data Type	Unsigned integer
Default	<code>des f ree / 2</code>
Range	<p>The minimum value is 128 KB or 1/256th of physical memory, whichever is greater, expressed as pages using the page size returned by <code>getpagesize</code>.</p> <p>The maximum value is the number of physical memory pages. The maximum value should be no more than 7.5 percent of physical memory. The system does not enforce this range other than that described in the Validation section.</p>
Units	Pages
Dynamic?	Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided in the <code>/etc/system</code> file or calculated from the new physical memory value.
Validation	If <code>minfree</code> is greater than <code>des f ree</code> , <code>minfree</code> is set to <code>des f ree / 2</code> . No message is displayed.

Implicit	The relationship of <code>lotsfree</code> being greater than <code>desfree</code> , which is greater than <code>minfree</code> , should be maintained at all times.
When to Change	The default value is generally adequate. For systems with relatively static workloads and large amounts of memory, lower this value. The minimum acceptable value is 128 KB, expressed as pages using the page size returned by <code>getpagesize</code> .
Commitment Level	Unstable

throttlefree

Description	Specifies the memory level at which blocking memory allocation requests are put to sleep, even if the memory is sufficient to satisfy the request.
Data Type	Unsigned integer
Default	<code>minfree</code>
Range	<p>The minimum value is 128 KB or 1/256th of physical memory, whichever is greater, expressed as pages using the page size returned by <code>getpagesize</code>.</p> <p>The maximum value is the number of physical memory pages. The maximum value should be no more than 4 percent of physical memory. The system does not enforce this range other than that described in the Validation section.</p>
Units	Pages
Dynamic?	Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided in the <code>/etc/system</code> file or calculated from the new physical memory value.
Validation	If <code>throttlefree</code> is greater than <code>desfree</code> , <code>throttlefree</code> is set to <code>minfree</code> . No message is displayed.
Implicit	The relationship of <code>lotsfree</code> is greater than <code>desfree</code> , which is greater than <code>minfree</code> , should be maintained at all times.
When to Change	The default value is generally adequate. For systems with relatively static workloads and large amounts of memory, lower this value. The minimum acceptable value is 128 KB, expressed as pages using the page size returned by <code>getpagesize</code> . For more information, see getpagesize(3C) .

Commitment Level Unstable

pageout_reserve

Description	Specifies the number of pages reserved for the exclusive use of the pageout or scheduler threads. When available memory is less than this value, nonblocking allocations are denied for any processes other than pageout or the scheduler. Pageout needs to have a small pool of memory for its use so it can allocate the data structures necessary to do the I/O for writing a page to its backing store.
Data Type	Unsigned integer
Default	<code>throttelfree / 2</code>
Range	<p>The minimum value is 64 KB or 1/512th of physical memory, whichever is greater, expressed as pages using the page size returned by <code>getpagesize(3C)</code>.</p> <p>The maximum is the number of physical memory pages. The maximum value should be no more than 2 percent of physical memory. The system does not enforce this range, other than that described in the Validation section.</p>
Units	Pages
Dynamic?	Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided in the <code>/etc/system</code> file or calculated from the new physical memory value.
Validation	If <code>pageout_reserve</code> is greater than <code>throttelfree / 2</code> , <code>pageout_reserve</code> is set to <code>throttelfree / 2</code> . No message is displayed.
Implicit	The relationship of <code>lotsfree</code> being greater than <code>desfree</code> , which is greater than <code>minfree</code> , should be maintained at all times.
When to Change	The default value is generally adequate. For systems with relatively static workloads and large amounts of memory, lower this value. The minimum acceptable value is 64 KB, expressed as pages using the page size returned by <code>getpagesize</code> .
Commitment Level	Unstable

pages_pp_maximum

Description	Defines the number of pages that must be unlocked. If a request to lock pages would force available memory below this value, that request is refused.
Data Type	Unsigned long
Default	The greater of (<code>tune_t_minarmem + 100</code> and [4% of memory available at boot time + 4 MB])
Range	Minimum value enforced by the system is <code>tune_t_minarmem + 100</code> . The system does not enforce a maximum value.
Units	Pages
Dynamic?	Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided in the <code>/etc/system</code> file or was calculated from the new physical memory value.
Validation	If the value specified in the <code>/etc/system</code> file or the calculated default is less than <code>tune_t_minarmem + 100</code> , the value is reset to <code>tune_t_minarmem + 100</code> . No message is displayed if the value from the <code>/etc/system</code> file is increased. Validation is done only at boot time and during dynamic reconfiguration operations that involve adding or deleting memory.
When to Change	When memory-locking requests fail or when attaching to a shared memory segment with the <code>SHARE_MMU</code> flag fails, yet the amount of memory available seems to be sufficient. Excessively large values can cause memory locking requests (<code>mlock</code> , <code>mlockall</code> , and <code>mlock(3C)</code>) to fail unnecessarily. For more information, see mlock(3C) , mlockall(3C) , and memcntl(2) .
Commitment Level	Unstable

tune_t_minarmem

Description	Defines the minimum available resident (not swappable) memory to maintain necessary to avoid deadlock. Used to reserve a portion of memory for use by the core of the OS. Pages restricted in this way are not seen when the OS determines the maximum amount of memory available.
-------------	--

Data Type	Signed integer
Default	25
Range	1 to physical memory
Units	Pages
Dynamic?	No
Validation	None. Large values result in wasted physical memory.
When to Change	The default value is generally adequate. Consider increasing the default value if the system locks up and debugging information indicates that no memory was available.
Commitment Level	Unstable

fastscan

Description	Defines the maximum number of pages per second that the system looks at when memory pressure is highest.
Data Type	Signed integer
Default	The <code>fastscan</code> default value is set in one of the following ways: <ul style="list-style-type: none">▪ The <code>fastscan</code> value set in the <code>/etc/system</code> file is used.▪ The <code>maxfastscan</code> value set in the <code>/etc/system</code> file is used.▪ If neither <code>fastscan</code> nor <code>maxfastscan</code> is set in the <code>/etc/system</code> file, <code>fastscan</code> is set to 64 MB when the system is booted. Then, after the system is booted for a few minutes, the <code>fastscan</code> value is set to the number of pages that the scanner can scan in one second using 10% of a CPU. <p>In all three cases, if the derived value is more than half the memory in the system, the <code>fastscan</code> value is capped at the value of half the memory in the system.</p>
Range	64 MB to half the system's physical memory
Units	Pages
Dynamic?	Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided by <code>/etc/system</code> or calculated from the new physical memory value.
Validation	The maximum value is the lesser of 64 MB and 1/2 of physical memory.

When to Change When more aggressive scanning of memory is preferred during periods of memory shortfall, especially when the system is subject to periods of intense memory demand or when performing heavy file I/O.

Commitment Level Unstable

slowscan

Description Defines the minimum number of pages per second that the system looks at when attempting to reclaim memory.

Data Type Signed integer

Default The smaller of 1/20th of physical memory in pages and 100.

Range 1 to $\text{fastscan} / 2$

Units Pages

Dynamic? Yes, unless dynamic reconfiguration operations that add or delete memory occur. At that point, the value is reset to the value provided in the `/etc/system` file or calculated from the new physical memory value.

Validation If `slowscan` is larger than $\text{fastscan} / 2$, `slowscan` is reset to $\text{fastscan} / 2$. No message is displayed.

When to Change When more aggressive scanning of memory is preferred during periods of memory shortfall, especially when the system is subject to periods of intense memory demand.

Commitment Level Unstable

min_percent_cpu

Description Defines the minimum percentage of CPU that pageout can consume. This parameter is used as the starting point for determining the maximum amount of time that can be consumed by the page scanner.

Data Type Signed integer

Default 4

Range 1 to 80

Units Percentage

Dynamic? Yes

Validation	None
When to Change	Increasing this value on systems with multiple CPUs and lots of memory, which are subject to intense periods of memory demand, enables the pager to spend more time attempting to find memory.
Commitment Level	Unstable

handspreadpages

Description	The Oracle Solaris OS uses a two-handed clock algorithm to look for pages that are candidates for reclaiming when memory is low. The first hand of the clock walks through memory marking pages as unused. The second hand walks through memory some distance after the first hand, checking to see if the page is still marked as unused. If so, the page is subject to being reclaimed. The distance between the first hand and the second hand is <code>handspreadpages</code> .
Data Type	Unsigned long
Default	<code>fastscan</code>
Range	1 to maximum number of physical memory pages on the system
Units	Pages
Dynamic?	Yes. This parameter requires that the kernel <code>reset_hands</code> parameter also be set to a non-zero value. Once the new value of <code>handspreadpages</code> has been recognized, <code>reset_hands</code> is set to zero.
Validation	The value is set to the lesser of either the amount of physical memory and the <code>handspreadpages</code> <i>value</i> .
When to Change	When you want to increase the amount of time that pages are potentially resident before being reclaimed. Increasing this value increases the separation between the hands, and therefore, the amount of time before a page can be reclaimed.
Commitment Level	Unstable

pages_before_pager

Description	Defines part of a system threshold that immediately frees pages after an I/O completes instead of storing the pages for possible reuse. The
-------------	---

	threshold is <code>lots_free + pages_before_pager</code> . The NFS environment also uses this threshold to curtail its asynchronous activities as memory pressure mounts.
Data Type	Signed integer
Default	200
Range	1 to amount of physical memory
Units	Pages
Dynamic?	No
Validation	None
When to Change	You might change this parameter when the majority of I/O is done for pages that are truly read or written once and never referenced again. Setting this variable to a larger amount of memory keeps adding pages to the free list. You might also change this parameter when the system is subject to bursts of severe memory pressure. A larger value here helps maintain a larger cushion against the pressure.
Commitment Level	Unstable

maxpgio

Description	Defines the maximum number of page I/O requests that can be queued by the paging system. This number is divided by 4 to get the actual maximum number used by the paging system. This parameter is used to throttle the number of requests as well as to control process swapping.
Data Type	Signed integer
Default	40
Range	1 to a variable maximum that depends on the system architecture, but mainly by the I/O subsystem, such as the number of controllers, disks, and disk swap size
Units	I/Os
Dynamic?	No
Validation	None

Implicit	The maximum number of I/O requests from the pager is limited by the size of a list of request buffers, which is currently sized at 256.
When to Change	Increase this parameter to page out memory faster. A larger value might help to recover faster from memory pressure if more than one swap device is configured or if the swap device is a striped device. Note that the existing I/O subsystem should be able to handle the additional I/O load. Also, increased swap I/O could degrade application I/O performance if the swap partition and application files are on the same disk.
Commitment Level	Unstable

Swapping-Related Parameters

Swapping in the Oracle Solaris OS is accomplished by the `swapfs` pseudo file system. The combination of space on swap devices and physical memory is treated as the pool of space available to support the system for maintaining backing store for anonymous memory. The system attempts to allocate space from disk devices first, and then uses physical memory as backing store. When `swapfs` is forced to use system memory for backing store, limits are enforced to ensure that the system does not deadlock because of excessive consumption by `swapfs`.

swapfs_reserve

Description	Defines the amount of system memory that is reserved for use by system (UID = 0) processes.
Data Type	Unsigned long
Default	The smaller of 4 MB and 1/16th of physical memory
Range	The minimum value is 4 MB or 1/16th of physical memory, whichever is smaller, expressed as pages using the page size returned by <code>getpagesize</code> . The maximum value is the number of physical memory pages. The maximum value should be no more than 10 percent of physical memory. The system does not enforce this range, other than that described in the Validation section.
Units	Pages
Dynamic?	No

Validation	None
When to Change	Generally not necessary. Only change when recommended by a software provider, or when system processes are terminating because of an inability to obtain swap space. A much better solution is to add physical memory or additional swap devices to the system.
Commitment Level	Unstable

swapfs_minfree

Description	Defines the desired amount of physical memory to be kept free for the rest of the system. Attempts to reserve memory for use as swap space by any process that causes the system's perception of available memory to fall below this value are rejected. Pages reserved in this manner can only be used for locked-down allocations by the kernel or by user-level processes.
Data Type	Unsigned long
Default	The larger of 2 MB and 12.5% of physical memory
Range	1 to amount of physical memory
Units	Pages
Dynamic?	No
Validation	None
When to Change	<p>Consider reducing this parameter value when processes are failing because of an inability to obtain swap space, yet the system has memory available. For example, change this value to use no more than 6.25% of system memory, but do not reduce it below 5% of system memory.</p> <p>On SPARC systems, the value should be at least 2 times the value of <code>tsb_alloc_hiwater_factor</code>. For more information, see “tsb_alloc_hiwater_factor” on page 75.</p>
Commitment Level	Unstable
Change History	For information, see “swapfs_minfree” on page 179 .

Kernel Memory Allocator

The Oracle Solaris kernel memory allocator distributes chunks of memory for use by clients inside the kernel. The allocator creates a number of caches of varying size for use by its clients. Clients can also request the allocator to create a cache for use by that client (for example, to allocate structures of a particular size). Statistics about each cache that the allocator manages can be seen by using the `kstat -c kmem_cache` command.

Occasionally, systems might panic because of memory corruption. The kernel memory allocator supports a debugging interface (a set of flags), that performs various integrity checks on the buffers. The kernel memory allocator also collects information on the allocators. The integrity checks provide the opportunity to detect errors closer to where they actually occurred. The collected information provides additional data for support people when they try to ascertain the reason for the panic.

Use of the flags incurs additional overhead and memory usage during system operations. The flags should only be used when a memory corruption problem is suspected.

kmem_flags

Description

The Oracle Solaris kernel memory allocator has various debugging and test options.

Five supported flag settings are described here.

Flag	Setting	Description
AUDIT	0x1	The allocator maintains a log that contains recent history of its activity. The number of items logged depends on whether CONTENTS is also set. The log is a fixed size. When space is exhausted, earlier records are reclaimed.
TEST	0x2	The allocator writes a pattern into freed memory and checks that the pattern is unchanged when the buffer is next allocated. If some portion of the buffer is changed, then the memory was probably used by a client that had previously allocated and freed the buffer. If an overwrite is identified, the system panics.

Flag	Setting	Description
REDZONE	0x4	The allocator provides extra memory at the end of the requested buffer and inserts a special pattern into that memory. When the buffer is freed, the pattern is checked to see if data was written past the end of the buffer. If an overwrite is identified, the kernel panics.
CONTENTS	0x8	The allocator logs up to 256 bytes of buffer contents when the buffer is freed. This flag requires that AUDIT also be set. The numeric value of these flags can be logically added together and set by the <code>/etc/system</code> file.
LITE	0x100	Does minimal integrity checking when a buffer is allocated and freed. When enabled, the allocator checks that the redzone has not been written into, that a freed buffer is not being freed again, and that the buffer being freed is the size that was allocated. Do not combine this flag with any other flags.

Data Type	Signed integer
Default	0 (disabled)
Range	0 (disabled) or 1 - 15 or 256 (0x100)
Dynamic?	Yes. Changes made during runtime only affect new kernel memory caches. After system initialization, the creation of new caches is rare.
Validation	None
When to Change	When memory corruption is suspected
Commitment Level	Unstable

kmem_stackinfo

Description	If the <code>kmem_stackinfo</code> variable is enabled in the <code>/etc/system</code> file at kernel thread creation time, the kernel thread stack is filled with a specific pattern instead of filled with zeros. During kernel thread execution, this kernel thread stack pattern is progressively overwritten. A simple count from the stack top until the pattern is not found gives a high watermark value, which is the maximum kernel stack space used by a kernel thread. This mechanism allows the following features:
-------------	--

	<ul style="list-style-type: none"> ▪ Compute the percentage of kernel thread stack really used (a high watermark) for current kernel threads in the system ▪ When a kernel thread ends, the system logs the last kernel threads that have used the most of their kernel thread stacks before dying to a small circular memory buffer
Data Type	Unsigned integer
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
Validation	None
When to Change	When you want to monitor kernel thread stack usage. Keep in mind that when <code>kmem_stackinfo</code> is enabled, the performance of creating and deleting <code>kthreads</code> is decreased. For more information, see the Chapter 5, “Built-In Commands,” in <i>Oracle Solaris Modular Debugger Guide</i> .
Zone Configuration	This parameter must be set in the global zone.
Commitment Level	Unstable

General Driver Parameters

moddebug

Description	When this parameter is enabled, messages about various steps in the module loading process are displayed.
Data Type	Signed integer
Default	0 (messages off)
Range	Here are the most useful values: <ul style="list-style-type: none"> ▪ 0x80000000 – Prints <code>[un] loading... message</code>. For every module loaded, messages such as the following appear on the console and in the <code>/var/adm/messages</code> file:

```
Apr 20 17:18:04 neo genunix: [ID 943528 kern.notice] load 'sched/TS_DPTBL' id 15
loaded @ 0x7be1b2f8/0x19c8380 size 176/2096
Apr 20 17:18:04 neo genunix: [ID 131579 kern.notice] installing TS_DPTBL,
module id 15.
```

- `0x40000000` – Prints detailed error messages. For every module loaded, messages such as the following appear on the console and in the `/var/adm/messages` file:

```
Apr 20 18:30:00 neo unix: Errno = 2
Apr 20 18:30:00 neo unix: kobj_open: vn_open of /platform/sun4v/kernel/exec/sparcv9/intpexec fails
Apr 20 18:30:00 neo unix: Errno = 2
Apr 20 18:30:00 neo unix: kobj_open: '/kernel/exec/sparcv9/intpexec'
Apr 20 18:30:00 neo unix: vp = 6001577600
Apr 20 18:30:00 neo unix: kobj_close: 0x6001577600
Apr 20 18:30:00 neo unix: kobj_open: vn_open of /platform/SUNW,Sun-Fire-T200/kernel/exec/sparcv9
/intpexec fails,
Apr 20 18:30:00 neo unix: Errno = 2
Apr 20 18:30:00 neo unix: kobj_open: vn_open of /platform/sun4v/kernel/exec/sparcv9/intpexec fails
```

- `0x20000000` - Prints even more detailed messages. This value doesn't print any additional information beyond what the `0x40000000` flag does during system boot. However, this value does print additional information about releasing the module when the module is unloaded.

These values can be added together to set the final value.

Dynamic?	Yes
Validation	None
When to Change	When a module is either not loading as expected, or the system seems to hang while loading modules. Note that when <code>0x40000000</code> is set, system boot is slowed down considerably by the number of messages written to the console.
Commitment Level	Unstable

ddi_msix_alloc_limit

Description	x86 only: This parameter controls the number of Extended Message Signaled Interrupts (MSI-X) that a device instance can allocate. Due to an existing system limitation, the default value is 2. You can increase the number of MSI-X interrupts that a device instance can allocate by increasing the value of this parameter. This parameter can be set either by editing the <code>/etc/system</code> file or by setting it with <code>mdb</code> before the device driver attach occurs.
Data Type	Signed integer
Default	2
Range	1 to 16

Dynamic?	Yes
Validation	None
When to Change	To increase the number of MSI-X interrupts that a device instance can allocate. However, if you increase the number of MSI-X interrupts that a device instance can allocate, adequate interrupts might not be available to satisfy all allocation requests. If this happens, some devices might stop functioning or the system might fail to boot. Reduce the value or remove the parameter in this case.
Commitment Level	Unstable

Network Driver Parameters

igb Parameters

mr_enable

Description	This parameter enables or disables multiple receive and transmit queues that are used by the <code>igb</code> network driver. This parameter can be set by editing the <code>/etc/driver/drv/igb.conf</code> file before the <code>igb</code> driver attach occurs.
Data Type	Boolean
Default	1 (disable multiple queues)
Range	0 (enable multiple queues) or 1 (disable multiple queues)
Dynamic?	No
Validation	None
When to Change	To enable or disable multiple receive and transmit queues that are used by the <code>igb</code> network driver.
Commitment Level	Unstable

intr_force

Description	This parameter is used to force an interrupt type, such as MSI, MSI-X, or legacy, that is used by the <code>igb</code> network driver. This parameter can be set by editing the <code>/etc/driver/drv/igb.conf</code> file before the <code>igb</code> driver attach occurs.
-------------	--

Data Type	Unsigned integer
Default	0 (do not force an interrupt type)
Range	0 (do not force an interrupt type) 1 (force MSI-X interrupt type) 2 (force MSI interrupt type) 3 (force legacy interrupt type)
Dynamic?	No
Validation	None
When to Change	To force an interrupt type that is used by the igb network driver.
Commitment Level	Unstable

ixgbe Parameters

tx_queue_number

Description	This parameter controls the number of transmit queues that are used by the ixgbe network driver. You can increase the number of transmit queues by increasing the value of this parameter. This parameter can be set by editing the <code>/etc/driver/drv/ixgbe.conf</code> file before the ixgbe driver attach occurs.
Data Type	Unsigned integer
Default	8
Range	1 to 32
Dynamic?	No
Validation	None
When to Change	To change the number of transmit queues that are used by the ixgbe network driver.
Commitment Level	Unstable

rx_queue_number

Description	This parameter controls the number of receive queues that are used by the ixgbe network driver. You can increase the number of receive
-------------	--

queues by increasing the value of this parameter. This parameter can be set by editing the `/etc/driver/drv/ixgbe.conf` file before the `ixgbe` driver attach occurs.

Data Type	Unsigned integer
Default	8
Range	1 to 64
Dynamic?	No
Validation	None
When to Change	To change the number of receive queues that are used by the <code>ixgbe</code> network driver.
Commitment Level	Unstable

intr_throttling

Description This parameter controls the interrupt throttling rate of the `ixgbe` network driver. You can increase the rate of interrupt by decreasing the value of this parameter. This parameter can be set by editing the `/etc/driver/drv/ixgbe.conf` file before the `ixgbe` driver attach occurs.

Data Type	Unsigned integer
Default	200
Range	0 to 65535
Dynamic?	No
Validation	None
When to Change	To change the interrupt throttling rate that is used by the <code>ixgbe</code> network driver.
Commitment Level	Unstable

rx_limit_per_intr

Description This parameter controls the maximum number of receive queue buffer descriptors per interrupt that are used by the `ixgbe` network driver. You can increase the number of receive queue buffer descriptors by increasing the value of this parameter. This parameter can be set by editing the `/etc/driver/drv/ixgbe.conf` file before the `ixgbe` driver attach occurs.

Data Type	Unsigned integer
Default	256
Range	16 to 4096
Dynamic?	No
Validation	None
When to Change	To change the number of receive queue buffer descriptors that are handled per interrupt by the ixgbe network driver.
Commitment Level	Unstable

tx_ring_size

Description	This parameter controls the transmit queue size that is used by the ixgbe network driver. You can increase the transmit queue size by increasing the value of this parameter. This parameter can be set by editing the <code>/etc/driver/drv/ixgbe.conf</code> file before the ixgbe driver attach occurs.
Data Type	Unsigned integer
Default	1024
Range	64 to 4096
Dynamic?	No
Validation	None
When to Change	To change the transmit queue size that is used by the ixgbe network driver.
Commitment Level	Unstable

rx_ring_size

Description	This parameter controls the receive queue size that is used by the ixgbe network driver. You can increase the receive queue size by increasing the value of this parameter. This parameter can be set by editing the <code>/etc/driver/drv/ixgbe.conf</code> file before the ixgbe driver attach occurs.
Data Type	Unsigned integer
Default	1024
Range	64 to 4096

Dynamic?	No
Validation	None
When to Change	To change the receive queue size that is used by the ixgbe network driver.
Commitment Level	Unstable

tx_copy_threshold

Description	This parameter controls the transmit buffer copy threshold that is used by the ixgbe network driver. You can increase the transmit buffer copy threshold by increasing the value of this parameter. This parameter can be set by editing the <code>/etc/driver/drv/ixgbe.conf</code> file before the ixgbe driver attach occurs.
Data Type	Unsigned integer
Default	512
Range	0 to 9126
Dynamic?	No
Validation	None
When to Change	To change the transmit buffer copy threshold that is used by the ixgbe network driver.
Commitment Level	Unstable

rx_copy_threshold

Description	This parameter controls the receive buffer copy threshold that is used by the ixgbe network driver. You can increase the receive buffer copy threshold by increasing the value of this parameter. This parameter can be set by editing the <code>/etc/driver/drv/ixgbe.conf</code> file before the ixgbe driver attach occurs.
Data Type	Unsigned integer
Default	128
Range	0 to 9126
Dynamic?	No
Validation	None
When to Change	To change the receive buffer copy threshold that is used by the ixgbe network driver.

Commitment Level Unstable

General I/O Parameters

maxphys

Description	Defines the maximum size of physical I/O requests. If a driver encounters a request larger than this size, the driver breaks the request into maxphys sized chunks. File systems can and do impose their own limit.
Data Type	Signed integer
Default	131,072 (sun4u or sun4v) or 57,344 (x86). The sd driver uses the value of 1,048,576 if the drive supports wide transfers. The sdd driver uses 1,048,576 by default.
Range	Machine-specific page size to MAXINT
Units	Bytes
Dynamic?	Yes, but many file systems load this value into a per-mount point data structure when the file system is mounted. A number of drivers load the value at the time a device is attached to a driver-specific data structure.
Validation	None
When to Change	When doing I/O to and from raw devices in large chunks. Note that a DBMS doing OLTP operations issues large numbers of small I/Os. Changing maxphys does not result in any performance improvement in that case.
Commitment Level	Unstable

rlim_fd_max

Description	Specifies the “hard” limit on file descriptors that a single process might have open. Overriding this limit requires superuser privilege.
Data Type	Signed integer
Default	65,536
Range	1 to MAXINT

Units	File descriptors
Dynamic?	No
Validation	None
When to Change	<p>When the maximum number of open files for a process is not enough. Other limitations in system facilities can mean that a larger number of file descriptors is not as useful as it might be. For example:</p> <ul style="list-style-type: none">▪ A 32-bit program using standard I/O is limited to 256 file descriptors. A 64-bit program using standard I/O can use up to 2 billion descriptors. Specifically, standard I/O refers to the <code>stdio(3C)</code> functions in <code>libc(3LIB)</code>.▪ <code>select</code> is by default limited to 1024 descriptors per <code>fd_set</code>. For more information, see <code>select(3C)</code>. A 32-bit application code can be recompiled with a larger <code>fd_set</code> size (less than or equal to 65,536). A 64-bit application uses an <code>fd_set</code> size of 65,536, which cannot be changed. <p>An alternative to changing this on a system wide basis is to use the <code>plimit(1)</code> command. If a parent process has its limits changed by <code>plimit</code>, all children inherit the increased limit. This alternative is useful for daemons such as <code>inetd</code>.</p>
Commitment Level	Unstable

rlim_fd_cur

Description	Defines the “soft” limit on file descriptors that a single process can have open. A process might adjust its file descriptor limit to any value up to the “hard” limit defined by <code>rlim_fd_max</code> by using the <code>setrlimit()</code> call or by issuing the <code>limit</code> command in whatever shell it is running. You do not require superuser privilege to adjust the limit to any value less than or equal to the hard limit.
Data Type	Signed integer
Default	256
Range	1 to MAXINT
Units	File descriptors
Dynamic?	No
Validation	Compared to <code>rlim_fd_max</code> . If <code>rlim_fd_cur</code> is greater than <code>rlim_fd_max</code> , <code>rlim_fd_cur</code> is reset to <code>rlim_fd_max</code> .

When to Change	When the default number of open files for a process is not enough. Increasing this value means only that it might not be necessary for a program to use <code>setrlimit</code> to increase the maximum number of file descriptors available to it.
Commitment Level	Unstable

General File System Parameters

ncsize

Description	<p>Defines the number of entries in the directory name look-up cache (DNLC). This parameter is used by UFS, NFS, and ZFS to cache elements of path names that have been resolved.</p> <p>The DNLC also caches negative look-up information, which means it caches a name not found in the cache.</p>
Data Type	Signed integer
Default	$(4 \times (v.v_proc + \text{maxusers}) + 320) + (4 \times (v.v_proc + \text{maxusers}) + 320) / 100$
Range	0 to MAXINT
Units	DNLC entries
Dynamic?	No
Validation	None. Larger values cause the time it takes to unmount a file system to increase as the cache must be flushed of entries for that file system during the unmount process.
When to Change	<p>You can use the <code>kstat -n dnlcstats</code> command to determine when entries have been removed from the DNLC because it was too small. The sum of the <code>pick_heuristic</code> and the <code>pick_last</code> parameters represents otherwise valid entries that were reclaimed because the cache was too small.</p> <p>Excessive values of <code>ncsize</code> have an immediate impact on the system because the system allocates a set of data structures for the DNLC based on the value of <code>ncsize</code>. By default, a system allocates 64-byte structures for <code>ncsize</code>. The value has a further effect on UFS and NFS, unless <code>ufs_ninode</code> and <code>nfs:ninode</code> are explicitly set.</p>
Commitment Level	Unstable

dnlc_dir_enable

Description Enables large directory caching

Note – This parameter has no effect on NFS or ZFS file systems.

Data Type Unsigned integer

Default 1 (enabled)

Range 0 (disabled) or 1 (enabled)

Dynamic? Yes, but do not change this tunable dynamically. You can enable this parameter if it was originally disabled. Or, you can disable this parameter if it was originally enabled. However, enabling, disabling, and then enabling this parameter might lead to stale directory caches.

Validation No

When to Change Directory caching has no known problems. However, if problems occur, then set `dnlc_dir_enable` to 0 to disable caching.

Commitment Level Unstable

dnlc_dir_min_size

Description Specifies the minimum number of entries cached for one directory.

Note – This parameter has no effect on NFS or ZFS file systems.

Data Type Unsigned integer

Default 40

Range 0 to MAXUINT (no maximum)

Units Entries

Dynamic? Yes, this parameter can be changed at any time.

Validation None

When to Change If performance problems occur with caching small directories, then increase `dnlc_dir_min_size`. Note that individual file systems might have their own range limits for caching directories.

Commitment Level Unstable

dnlc_dir_max_size

Description Specifies the maximum number of entries cached for one directory.

Note – This parameter has no effect on NFS or ZFS file systems.

Data Type Unsigned integer

Default MAXUINT (no maximum)

Range 0 to MAXUINT

Dynamic? Yes, this parameter can be changed at any time.

Validation None

When to Change If performance problems occur with large directories, then decrease `dnlc_dir_max_size`.

Commitment Level Unstable

dnlc_dircache_percent

Description Calculates the maximum percentage of physical memory that the DNLC directory cache can consume.

Data Type Integer

Default 100

Range 0 to 100

Units Percentage

Dynamic? No

Validation At boot time, the value range is checked and default value is enforced.

When to Change When the system experiences a memory shortage and high kernel memory consumption, consider lowering this value. If performance issues are seen with the default value, consider increasing the value.

Note – The DNLC is used by UFS and ZFS file systems and NFS clients. Setting this tunable might be considered for better performance when there are memory shortages and high kernel memory consumption or when a memory is needed by the ARC or other kernel caches.

Commitment Level	Unstable
Change History	For information, see “ dnlc_dircache_percent ” on page 179.

TMPFS Parameters

tmpfs:tmpfs_maxkmem

Description	Defines the maximum amount of kernel memory that TMPFS can use for its data structures (tmpnodes and directory entries).
Data Type	Unsigned long
Default	One page or 4 percent of physical memory, whichever is greater.
Range	Number of bytes in one page (8192 for sun4u or sun4v systems, 4096 for all other systems) to 25 percent of the available kernel memory at the time TMPFS was first used.
Units	Bytes
Dynamic?	Yes
Validation	None
When to Change	Increase if the following message is displayed on the console or written in the messages file: <pre>tmp_memalloc: tmpfs over memory limit</pre> <p>The current amount of memory used by TMPFS for its data structures is held in the <code>tmp_kmemspace</code> field. This field can be examined with a kernel debugger.</p>
Commitment Level	Unstable

tmpfs:tmpfs_minfree

Description	Defines the minimum amount of swap space that TMPFS leaves for the rest of the system.
Data Type	Signed long
Default	512
Range	0 to maximum swap space size
Units	Pages
Dynamic?	Yes
Validation	None
When to Change	To maintain a reasonable amount of swap space on systems with large amounts of TMPFS usage, you can increase this number. The limit has been reached when the console or messages file displays the following message: <i>fs-name: File system full, swap space limit exceeded</i>
Commitment Level	Unstable

Pseudo Terminals

Pseudo terminals, ptys, are used for two purposes in Oracle Solaris software:

- Supporting remote logins by using the `telnet`, `rlogin`, or `rsh` commands
- Providing the interface through which the X Window system creates command interpreter windows

The default number of pseudo-terminals is sufficient for a desktop workstation. So, tuning focuses on the number of ptys available for remote logins.

The default number of ptys is now based on the amount of memory on the system. This default should be changed only to restrict or increase the number of users who can log in to the system.

Three related variables are used in the configuration process:

- `pt_cnt` – Default maximum number of ptys.
- `pt_pctofmem` – Percentage of kernel memory that can be dedicated to pty support structures. A value of zero means that no remote users can log in to the system.
- `pt_max_pty` – Hard maximum for number of ptys.

`pt_cnt` has a default value of zero, which tells the system to limit logins based on the amount of memory specified in `pt_pctofmem`, unless `pt_max_pty` is set. If `pt_cnt` is non-zero, ptys are allocated until this limit is reached. When that threshold is crossed, the system looks at `pt_max_pty`. If `pt_max_pty` has a non-zero value, it is compared to `pt_cnt`. The pty allocation is allowed if `pt_cnt` is less than `pt_max_pty`. If `pt_max_pty` is zero, `pt_cnt` is compared to the number of ptys supported based on `pt_pctofmem`. If `pt_cnt` is less than this value, the pty allocation is allowed. Note that the limit based on `pt_pctofmem` only comes into play if both `pt_cnt` and `ptms_ptymax` have default values of zero.

To put a hard limit on ptys that is different than the maximum derived from `pt_pctofmem`, set `pt_cnt` and `ptms_ptymax` in `/etc/system` to the preferred number of ptys. The setting of `ptms_pctofmem` is not relevant in this case.

To dedicate a different percentage of system memory to pty support and let the operating system manage the explicit limits, do the following:

- Do not set `pt_cnt` or `ptms_ptymax` in `/etc/system`.
- Set `pt_pctofmem` in `/etc/system` to the preferred percentage. For example, set `pt_pctofmem=10` for a 10 percent setting.

Note that the memory is not actually allocated until it is used in support of a pty. Once memory is allocated, it remains allocated.

pt_cnt

Description	The number of available <code>/dev/pts</code> entries is dynamic up to a limit determined by the amount of physical memory available on the system. <code>pt_cnt</code> is one of three variables that determines the minimum number of logins that the system can accommodate. The default maximum number of <code>/dev/pts</code> devices the system can support is determined at boot time by computing the number of pty structures that can fit in a percentage of system memory (see <code>pt_pctofmem</code>). If <code>pt_cnt</code> is zero, the system allocates up to that maximum. If <code>pt_cnt</code> is non-zero, the system allocates to the greater of <code>pt_cnt</code> and the default maximum.
Data Type	Unsigned integer
Default	0
Range	0 to <code>maxpid</code>
Units	Logins/windows
Dynamic?	No
Validation	None

When to Change When you want to explicitly control the number of users who can remotely log in to the system.

Commitment Level Unstable

pt_pctofmem

Description Specifies the maximum percentage of physical memory that can be consumed by data structures to support /dev/pts entries. A system consumes 176 bytes per /dev/pts entry.

Data Type Unsigned integer

Default 5

Range 0 to 100

Units Percentage

Dynamic? No

Validation None

When to Change When you want to either restrict or increase the number of users who can log in to the system. A value of zero means that no remote users can log in to the system.

Commitment Level Unstable

pt_max_pty

Description Defines the maximum number of ptys the system offers

Data Type Unsigned integer

Default 0 (Uses system-defined maximum)

Range 0 to MAXUINT

Units Logins/windows

Dynamic? Yes

Validation None

Implicit Should be greater than or equal to pt_cnt. Value is not checked until the number of ptys allocated exceeds the value of pt_cnt.

When to Change	When you want to place an absolute ceiling on the number of logins supported, even if the system could handle more based on its current configuration values.
Commitment Level	Unstable

STREAMS Parameters

nstrpush

Description	Specifies the number of modules that can be inserted into (pushed onto) a STREAM.
Data Type	Signed integer
Default	9
Range	9 to 16
Units	Modules
Dynamic?	Yes
Validation	None
When to Change	At the direction of your software vendor. No messages are displayed when a STREAM exceeds its permitted push count. A value of EINVAL is returned to the program that attempted the push.
Commitment Level	Unstable

strmsgsz

Description	Specifies the maximum number of bytes that a single system call can pass to a STREAM to be placed in the data part of a message. Any <code>write</code> exceeding this size is broken into multiple messages. For more information, see write(2) .
Data Type	Signed integer
Default	65,536
Range	0 to 262,144
Units	Bytes

Dynamic?	Yes
Validation	None
When to Change	When <code>putmsg</code> calls return <code>ERANGE</code> . For more information, see putmsg(2) .
Commitment Level	Unstable

strctlsz

Description	Specifies the maximum number of bytes that a single system call can pass to a <code>STREAM</code> to be placed in the control part of a message
Data Type	Signed integer
Default	1024
Range	0 to <code>MAXINT</code>
Units	Bytes
Dynamic?	Yes
Validation	None
When to Change	At the direction of your software vendor. <code>putmsg(2)</code> calls return <code>ERANGE</code> if they attempt to exceed this limit.
Commitment Level	Unstable

System V Message Queues

System V message queues provide a message-passing interface that enables the exchange of messages by queues created in the kernel. Interfaces are provided in the Oracle Solaris environment to enqueue and dequeue messages. Messages can have a type associated with them. Enqueueing places messages at the end of a queue. Dequeueing removes the first message of a specific type from the queue or the first message if no type is specified.

For detailed information on tuning these system resources, see [Chapter 6, “Resource Controls \(Overview\)”](#), in *Oracle Solaris Administration: Oracle Solaris Zones, Oracle Solaris 10 Zones, and Resource Management*.

System V Semaphores

System V semaphores provide counting semaphores in the Oracle Solaris OS. A *semaphore* is a counter used to provide access to a shared data object for multiple processes. In addition to the standard set and release operations for semaphores, System V semaphores can have values that are incremented and decremented as needed (for example, to represent the number of resources available). System V semaphores also provide the ability to do operations on a group of semaphores simultaneously as well as to have the system undo the last operation by a process if the process dies.

System V Shared Memory

System V shared memory allows the creation of a segment by a process. Cooperating processes can attach to the memory segment (subject to access permissions on the segment) and gain access to the data contained in the segment. This capability is implemented as a loadable module. Entries in the `/etc/system` file must contain the `shmsys:` prefix.

A special kind of shared memory known as *intimate shared memory* (ISM) is used by DBMS vendors to maximize performance. When a shared memory segment is made into an ISM segment, the memory for the segment is locked. This feature enables a faster I/O path to be followed and improves memory usage. A number of kernel resources describing the segment are then shared between all processes that attach to the segment in ISM mode.

segspt_minfree

Description	Identifies pages of system memory that cannot be allocated for ISM shared memory.
Data Type	Unsigned long
Default	5 percent of available system memory when the first ISM segment is created
Range	0 to 50 percent of physical memory
Units	Pages
Dynamic?	Yes
Validation	None. Values that are too small can cause the system to hang or performance to severely degrade when memory is consumed with ISM segments.
When to Change	On database servers with large amounts of physical memory using ISM, the value of this parameter can be decreased. If ISM segments are not

used, this parameter has no effect. A maximum value of 128 MB (0x4000) is almost certainly sufficient on large memory machines.

Commitment Level Unstable

Scheduling

disp_rechoose_interval

Description

Similar to the previous `rechoose_interval` parameter, this parameter specifies the amount of time before a process is deemed to have lost all affinity for the last CPU it ran on. However, this parameter is set in more granular time increments. This parameter should be used instead of the deprecated `rechoose_interval` parameter, but the `rechoose_interval` parameter is still accepted if it is set in the `/etc/system` file.

After this interval expires, any CPU is considered a candidate for scheduling a thread. This parameter does not apply to threads in the real-time class, but applies to threads in all other scheduling classes.

Use `mdb` if you want to change the value of this parameter by using the following steps:

1. Convert nanoseconds to unscaled time. For example, to convert a 5000000 nanosecond based value to unscaled time, use the following syntax:

```
# mdb -kw
.
.
.
> 0t5000000::time -u
0xb6a444
```

2. Set `disp_rechoose_interval` to the unscaled time value. For example, provide the value that was returned in preceding step.

```
> disp_rechoose_interval /Z 0xb6a444
disp_rechoose_interval: 0x447d998 = 0xb6a444
```

3. Verify that `disp_rechoose_interval` has been set to the right value. For example:

```
> disp_rechoose_interval::print
0xb6a444
```

Data Type Signed integer

Default	3
Range	0 to MAXINT
Dynamic?	Yes
Validation	None
When to Change	When caches are large, or when the system is running a critical process or a set of processes that seem to suffer from excessive cache misses not caused by data access patterns. Consider using the processor set capabilities or processor binding before changing this parameter. For more information, see psrset(1M) or pbind(1M) .
Commitment Level	Unstable
Change History	For information, see “ disp_rechoose_interval ” on page 178.

Timers

hires_tick

Description	When set, this parameter causes the Oracle Solaris OS to use a system clock rate of 1000 instead of the default value of 100.
Data Type	Signed integer
Default	0
Range	0 (disabled) or 1 (enabled)
Dynamic?	No. Causes new system timing variable to be set at boot time. Not referenced after boot.
Validation	None
When to Change	When you want timeouts with a resolution of less than 10 milliseconds, and greater than or equal to 1 millisecond.
Commitment Level	Unstable

timer_max

Description	Specifies the number of POSIX timers available.
-------------	---

Data Type	Signed integer
Default	1000
Range	0 to MAXINT
Dynamic?	No. Increasing the value can cause a system crash.
Validation	None
When to Change	When the default number of timers offered by the system is inadequate. Applications receive an EAGAIN error when executing <code>timer_create</code> system calls.
Commitment Level	Unstable

SPARC System Specific Parameters

The following parameters apply to sun4v and SPARC M-Series sun4u platforms.

tsb_alloc_hiwater_factor

Description Initializes `tsb_alloc_hiwater` to impose an upper limit on the amount of physical memory that can be allocated for translation storage buffers (TSBs) as follows:

$$\text{tsb_alloc_hiwater} = \text{physical memory (bytes)} / \text{tsb_alloc_hiwater_factor}$$

When the memory that is allocated to TSBs is equal to the value of `tsb_alloc_hiwater`, the TSB memory allocation algorithm attempts to reclaim TSB memory as pages are unmapped.

Exercise caution when using this factor to increase the value of `tsb_alloc_hiwater`. To prevent system hangs, the resulting high water value must be considerably lower than the value of `swapfs_minfree` and `segspt_minfree`.

Data Type	Integer
Default	32
Range	1 to MAXINIT

Note that a factor of 1 makes all physical memory available for allocation to TSBs, which could cause the system to hang. A factor that is too high will not leave memory available for allocation to TSBs, decreasing system performance.

Dynamic?	Yes
Validation	None
When to Change	Change the value of this parameter if the system has many processes that attach to very large shared memory segments. Under most circumstances, tuning of this variable is not necessary.
Commitment Level	Unstable

default_tsb_size

Description	Selects size of the initial translation storage buffers (TSBs) allocated to all processes.
Data Type	Integer
Default	Default is 0 (8 KB), which corresponds to 512 entries
Range	Possible values are:

Value	Description
0	8 KB
1	16 KB
3	32 KB
4	128 KB
5	256 KB
6	512 KB
7	1 Mbyte

Dynamic?	Yes
Validation	None

When to Change Generally, you do not need to change this value. However, doing so might provide some advantages if the majority of processes on the system have a larger than average working set, or if resident set size (RSS) sizing is disabled.

Commitment Level Unstable

enable_tsb_rss_sizing

Description Enables a resident set size (RSS) based TSB sizing heuristic.

Data Type Boolean

Default 1 (TSBs can be resized)

Range 0 (TSBs remain at `tsb_default_size`) or 1 (TSBs can be resized)

If set to 0, then `tsb_rss_factor` is ignored.

Dynamic? Yes

Validation Yes

When to Change Can be set to 0 to prevent growth of the TSBs. Under most circumstances, this parameter should be left at the default setting.

Commitment Level Unstable

tsb_rss_factor

Description Controls the RSS to TSB span ratio of the RSS sizing heuristic. This factor divided by 512 yields the percentage of the TSB span which must be resident in memory before the TSB is considered as a candidate for resizing.

Data Type Integer

Default 384, resulting in a value of 75%. Thus, when the TSB is 3/4 full, its size will be increased. Note that some virtual addresses typically map to the same slot in the TSB. Therefore, conflicts can occur before the TSB is at 100% full.

Range 0 to 512

Dynamic? Yes

Validation None

When to Change	<p>If the system is experiencing an excessive number of traps due to TSB misses, for example, due to virtual address conflicts in the TSB, you might consider decreasing this value toward 0.</p> <p>For example, changing <code>tsb_rss_factor</code> to 256 (effectively, 50%) instead of 384 (effectively, 75%) might help eliminate virtual address conflicts in the TSB in some cases, but will use more kernel memory, particularly on a heavily loaded system.</p> <p>TSB activity can be monitored with the <code>trapstat -T</code> command.</p>
Commitment Level	Unstable

Locality Group Parameters

This section provides generic memory tunables, which apply to any SPARC or x86 system that uses a Non-Uniform Memory Architecture (NUMA).

lpg_alloc_prefer

Description	<p>Controls a heuristic for allocation of large memory pages when the requested page size is not immediately available in the local memory group, but could be satisfied from a remote memory group.</p> <p>By default, the Oracle Solaris OS allocates a remote large page if local free memory is fragmented, but remote free memory is not. Setting this parameter to 1 indicates that additional effort should be spent attempting to allocate larger memory pages locally, potentially moving smaller pages around to coalesce larger pages in the local memory group.</p>
Data Type	Boolean
Default	0 (Prefer remote allocation if local free memory is fragmented and remote free memory is not)
Range	0 (Prefer remote allocation if local free memory is fragmented and remote free memory is not) 1 (Prefer local allocation whenever possible, even if local free memory is fragmented and remote free memory is not)
Dynamic?	No
Validation	None

When to Change	<p>This parameter might be set to 1 if long-running programs on the system tend to allocate memory that is accessed by a single program, or if memory that is accessed by a group of programs is known to be running in the same locality group (lgroup). In these circumstances, the extra cost of page coalesce operations can be amortized over the long run of the programs.</p> <p>This parameter might be left at the default value (0) if multiple programs tend to share memory across different locality groups, or if pages tend to be used for short periods of time. In these circumstances, quick allocation of the requested size tends to be more important than allocation in a particular location.</p> <p>TLB miss activity might be observed by using the <code>trapstat -T</code> command.</p>
Commitment Level	Uncommitted

lgrp_mem_pset_aware

Description	<p>If a process is running within a user processor set, this variable determines whether <i>randomly</i> placed memory for the process is selected from among all the lgroups in the system or only from those lgroups that are spanned by the processors in the processor set.</p> <p>For more information about creating processor sets, see psrset(1M).</p>
Data Type	Boolean
Default	0, the Oracle Solaris OS selects memory from all the lgroups in the system
Range	<ul style="list-style-type: none"> ▪ 0, the Oracle Solaris OS selects memory from all the lgroups in the system (default) ▪ 1, try selecting memory only from those lgroups that are spanned by the processors in the processor set. If the first attempt fails, memory can be allocated in any lgroup.
Dynamic?	No
Validation	None
When to Change	Setting this value to a value of one (1) might lead to more reproducible performance when processor sets are used to isolate applications from one another.

Commitment Level Uncommitted

Oracle Solaris ZFS Tunable Parameters

This chapter describes ZFS tunable parameters that might need consideration, depending on your system and application requirements. In addition, tunable recommendations for using ZFS with database products are provided.

- “Tuning ZFS Considerations” on page 82
- “ZFS ARC Parameters” on page 82
- “ZFS File-Level Prefetch” on page 83
- “ZFS Device I/O Queue Depth” on page 84
- “Tuning ZFS When Using Flash Storage” on page 85
- “Tuning ZFS for Database Products” on page 89

Where to Find Tunable Parameter Information

Tunable Parameter	For Information
Oracle Solaris kernel tunable parameters	Chapter 2, “Oracle Solaris Kernel Tunable Parameters”
NFS tunable parameters	Chapter 4, “NFS Tunable Parameters”
Internet Protocol Suite tunable parameters	Chapter 5, “Internet Protocol Suite Tunable Parameters”

Tuning ZFS Considerations

Review the following considerations before tuning ZFS:

- Default values are generally the best value. If a better value exists, it should be the default. While alternative values might help a given workload, it could quite possibly degrade some other aspects of performance. Occasionally, catastrophically so.
- The ZFS best practices should be followed before ZFS tuning is applied. These practices are a set of recommendations that have been shown to work in different environments and are expected to keep working in the foreseeable future. So, before turning to tuning, make sure you've read and understood the best practices. For more information, see [Chapter 12, “Recommended Oracle Solaris ZFS Practices,”](#) in *Oracle Solaris 11.1 Administration: ZFS File Systems*.
- Unless noted otherwise, the tunable parameters are global and impact ZFS behavior across the system.

ZFS ARC Parameters

This section describes parameters related to ZFS ARC behavior.

zfs_arc_min

Description	Determines the minimum size of the ZFS Adaptive Replacement Cache (ARC). See also “ zfs_arc_max ” on page 83.
Data Type	Unsigned Integer (64-bit)
Default	64 MB
Range	64 MB to zfs_arc_max
Units	Bytes
Dynamic?	No
Validation	Yes, the range is validated.
When to Change	When a system's workload demand for memory fluctuates, the ZFS ARC caches data at a period of weak demand and then shrinks at a period of strong demand. However, ZFS does not shrink below the value of <code>zfs_arc_min</code> . Generally, you do not need to change the default value.
Commitment Level	Unstable
Change History	For information, see “ zfs_arc_min ” on page 177.

zfs_arc_max

Description	Determines the maximum size of the ZFS Adaptive Replacement Cache (ARC). See also “ zfs_arc_min ” on page 82.
Data Type	Unsigned Integer (64-bit)
Default	75% of memory on systems with less than 4 GB of memory physmem minus 1 GB on systems with greater than 4 GB of memory
Range	64 MB to physmem
Units	Bytes
Dynamic?	No
Validation	Yes, the range is validated.
When to Change	If a future memory requirement is significantly large and well defined, you might consider reducing the value of this parameter to cap the ARC so that it does not compete with the memory requirement. For example, if you know that a future workload requires 20% of memory, it makes sense to cap the ARC such that it does not consume more than the remaining 80% of memory.
Commitment Level	Unstable
Change History	For information, see “ zfs_arc_max ” on page 177.

ZFS File-Level Prefetch

zfs_prefetch_disable

Description	<p>This parameter determines a file-level prefetching mechanism called <code>zfetch</code>. This mechanism looks at the patterns of reads to files and anticipates on some reads, thereby reducing application wait times. The current behavior suffers from two drawbacks:</p> <ul style="list-style-type: none"> Sequential read patterns made of small reads very often hit in the cache. In this case, the current behavior consumes a significant amount of CPU time trying to find the next I/O to issue, whereas performance is governed more by the CPU availability. The <code>zfetch</code> code has been observed to limit scalability of some loads. CPU profiling can be done by using the <code>lockstat -I</code> command or <code>er_kernel</code> as described here:
-------------	---

<http://developers.sun.com/prodtech/cc/articles/perftools.html>

You can disable prefetching by setting `zfs_prefetch_disable` in the `/etc/system` file.

Device-level prefetching is disabled when `zfs_vdev_cache_size` is disabled. This means that tuning `vdev_cache_shift` is no longer necessary if `zfs_vdev_cache_size` is disabled.

Data Type	Boolean
Default	0 (enabled)
Range	0 (enabled) or 1 (disabled)
Dynamic?	Yes
Validation	No
When to Change	If the results of <code>er_kernel</code> show significant time in <code>zfetch_*</code> functions, or if lock profiling with <code>lockstat</code> shows contention around <code>zfetch</code> locks, then disabling file level prefetching should be considered.
Commitment Level	Unstable

ZFS Device I/O Queue Depth

`zfs_vdev_max_pending`

Description	This parameter controls the maximum number of concurrent I/Os pending to each device.
Data Type	Integer
Default	10
Range	0 to MAXINT
Dynamic?	Yes
Validation	No
When to Change	In a storage array where LUNs are made of a large number of disk drives, the ZFS queue can become a limiting factor on read IOPS. This behavior is one of the underlying reasoning for the best practice of presenting as many LUNs as there are backing spindles to the ZFS storage pool. That is, if you create LUNs from a 10 disk-wide array level

raid-group, then using 5 to 10 LUNs to build a storage pool allows ZFS to manage enough of an I/O queue without the need to set this specific tunable.

However, when no separate intent log is in use and the pool is made of JBOD disks, using a small `zfs_vdev_max_pending` value, such as 10, can improve the synchronous write latency as those are competing for the disk resource. Using separate intent log devices can alleviate the need to tune this parameter for loads that are synchronously write intensive since those synchronous writes are not competing with a deep queue of non-synchronous writes.

Tuning this parameter is not expected to be effective for NVRAM-based storage arrays in the case where volumes are made of small number of spindles. However, when ZFS is presented with a volume made of a large (greater than 10) number of spindles, then this parameter can limit the read throughput obtained on the volume. The reason is that with a maximum of 10 or 35 queued I/Os per LUN, this can translate into less than 1 I/O per storage spindle, which is not enough for individual disks to deliver their IOPS. This issue would appear in `iostat actv queue` output approaching the value of `zfs_vdev_max_pending`.

Device drivers may also limit the number of outstanding I/Os per LUN. If you are using LUNs on storage arrays that can handle large numbers of concurrent IOPS, then the device driver constraints can limit concurrency. Consult the configuration for the drivers your system uses. For example, the limit for the QLogic ISP2200, ISP2300, and SP212 family FCI HBA (qlc) driver is described as the `execution-throttle` parameter in `/kernel/drv/qlc.conf`.

Commitment Level Unstable

Tuning ZFS When Using Flash Storage

The following information applies to Flash SSDs, F20 PCIe Accelerator Card, F40 PCIe Accelerator Card, and F5100 Flash Storage Array.

Review the following general comments when using ZFS with Flash storage:

- Consider using LUNs or low latency disks that are managed by a controller with persistent memory, if available, for the ZIL (ZFS intent log). This option can be considerably more cost effective than using flash for low latency commits. The size of the log devices must only be

large enough to hold 10 seconds of maximum write throughput. Examples would include a storage array based LUN, or a disk connected to an HBA with a battery protected write cache.

If no such device is available, segment a separate pool of flash devices for use as log devices in a ZFS storage pool.

- The F40 and F20 Flash Accelerator cards contain and export 4 independent flash modules to the OS. The F5100 contains up to 80 independent flash modules. Each flash module appears to the operating system as a single device. SSDs are viewed as a single device by the OS. Flash devices may be used as ZFS log devices to reduce commit latency, particularly if used in an NFS server. For example, a single flash module of a flash device used as a ZFS log device can reduce latency of single lightly threaded operations by 10x. More flash devices can be striped together to achieve higher throughput for large amounts of synchronous operations.
- Log devices should be mirrored for reliability. For maximum protection, the mirrors should be created on separate flash devices. In the case of F20 and F40 PCIe accelerator cards, maximum protection is achieved by ensuring that mirrors reside on different physical PCIe cards. Maximum protection with the F5100 storage array is obtained by placing mirrors on separate F5100 devices.
- Flash devices that are not used as log devices may be used as second level cache devices. This serves to both offload IOPS from primary disk storage and also to improve read latency for commonly used data.

Adding Flash Devices as ZFS Log or Cache Devices

Review the following recommendations when adding flash devices as ZFS log or cache devices.

- A ZFS log or cache device can be added to an existing ZFS storage pool by using the `zpool add` command. Be very careful with `zpool add` commands. Mistakenly adding a log device as a normal pool device is a mistake that will require you to destroy and restore the pool from scratch. Individual log devices themselves can be removed from a pool.
- Familiarize yourself with the `zpool add` command before attempting this operation on active storage. You can use the `zpool add -n` option to preview the configuration without creating the configuration. For example, the following incorrect `zpool add` preview syntax attempts to add a device as a log device:

```
# zpool add -n tank c4t1d0
vdev verification failed: use -f to override the following errors:
mismatched replication level: pool uses mirror and new vdev is disk
Unable to build pool from specified devices: invalid vdev configuration
```

This is the correct `zpool add` preview syntax for adding a log device to an existing pool:

```
# zpool add -n tank log c4t1d0
would update 'tank' to the following configuration:
    tank
```

```

mirror
  c4t0d0
  c5t0d0
logs
  c4t1d0

```

If multiple devices are specified, they are striped together. For more information, see the examples below or [zpool\(1M\)](#).

A flash device, `c4t1d0`, can be added as a ZFS log device:

```
# zpool add pool log c4t1d0
```

If 2 flash devices are available, you can add mirrored log devices:

```
# zpool add pool log mirror c4t1d0 c4t2d0
```

Available flash devices can be added as a cache device for reads.

```
# zpool add pool cache c4t3d0
```

You can't mirror cache devices, they will be striped together.

```
# zpool add pool cache c4t3d0 c4t4d0
```

Ensuring Proper Cache Flush Behavior for Flash and NVRAM Storage Devices

ZFS is designed to work with storage devices that manage a disk-level cache. ZFS commonly asks the storage device to ensure that data is safely placed on stable storage by requesting a cache flush. For JBOD storage, this works as designed and without problems. For many NVRAM-based storage arrays, a performance problem might occur if the array takes the cache flush request and actually does something with it, rather than ignoring it. Some storage arrays flush their large caches despite the fact that the NVRAM protection makes those caches as good as stable storage.

ZFS issues infrequent flushes (every 5 second or so) after the uberblock updates. The flushing infrequency is fairly inconsequential so no tuning is warranted here. ZFS also issues a flush every time an application requests a synchronous write (`O_DSYNC`, `fsync`, NFS commit, and so on). The completion of this type of flush is waited upon by the application and impacts performance. Greatly so, in fact. From a performance standpoint, this neutralizes the benefits of having an NVRAM-based storage.

Cache flush tuning was recently shown to help flash device performance when used as log devices. When all LUNs exposed to ZFS come from NVRAM-protected storage array and procedures ensure that no unprotected LUNs will be added in the future, ZFS can be tuned to

not issue the flush requests by setting `zfs_nocacheflush`. If some LUNs exposed to ZFS are not protected by NVRAM, then this tuning can lead to data loss, application level corruption, or even pool corruption. In some NVRAM-protected storage arrays, the cache flush command is a no-op, so tuning in this situation makes no performance difference.

A recent OS change is that the flush request semantic has been qualified to instruct storage devices to ignore the requests if they have the proper protection. This change requires a fix to our disk drivers and for the NVRAM device to support the updated semantics. If the NVRAM device does not recognize this improvement, use these instructions to tell the Solaris OS not to send any synchronize cache commands to the array. If you use these instructions, make sure all targeted LUNS are indeed protected by NVRAM.

Occasionally, flash and NVRAM devices do not properly advertise to the OS that they are non-devices, and that caches do not need to be flushed. Cache flushing is an expensive operation. Unnecessary flushes can drastically impede performance in some cases.

Review the following `zfs_nocacheflush` syntax restrictions before applying the tuning entries below:

- The tuning syntax below can be included in `sd.conf` but there must be only a single `sd-config-list` entry per vendor/product.
- If multiple devices entries are desired, multiple pairs of vendor IDs and `sd` tuning strings can be specified on the same line by using the following syntax:

```
#           "012345670123456789012345", "tuning      ",
sd-config-list="|-VID1-||-----PID1-----|", "param1:val1, param2:val2",
               "|-VIDN-||-----PIDN-----|", "param1:val1, param3:val3";
```

Make sure the vendor ID (VID) string is padded to 8 characters and the product ID (PID) string is padded to 16 characters as described in the preceding example.



Caution – All cache sync commands are ignored by the device. Use at your own risk.

1. Use the `format` utility to run the `inquiry` subcommand on a LUN from the storage array.

For example:

```
# format
.
.
.
Specify disk (enter its number): x
format> inquiry
Vendor:   ATA
Product:  Marvell
Revision: XXXX
format>
```

2. Select one of the following based on your architecture:

- For F40 flash devices, add the following entry to `/kernel/drv/sd.conf`. In the entry below, ensure that "ATA " is padded to 8 characters, and "3E128-TS2-550B01" contains 16 characters. Total string length is 24.


```
sd-config-list="ATA      3E128-TS2-550B01","disksort:false, cache-non:true";
```
- For F20 and F5100 flash devices, choose one of the following based on your architecture. In the entries below, "ATA " is padded to 8 characters, and "MARVELL SD88SA02" contains 16 characters. The total string length is 24.
- Many SPARC architectures – Add the following entry to `/kernel/drv/ssd.conf`:

```
ssd-config-list = "ATA      MARVELL SD88SA02","throttle-max:32, disksort:false, cache-non:true";
```

- x64 and a few SPARC drivers – Add the following entry to `/kernel/drv/sd.conf`

```
ssd-config-list="ATA      MARVELL SD88SA02","throttle-max:32, disksort:false, cache-non:true";
```

3. Carefully add whitespace to make the vendor ID (VID) 8 characters long (here ATA) and Product ID (PID) 16 characters long (here MARVELL) in the `sd-config-list` entry as illustrated.
4. Reboot the system.

You can tune `zfs_nocacheflush` back to its default value (0) with no adverse effect on performance.

SCSI Unmap Considerations for Flash Devices

The Solaris 11.1 OS has a problem that resulted in excessive calls to SCSI unmap routines. This particular problem is detrimental to flash performance. The workaround is to disable the unmap feature as follows:

- Include the following entry in the `/etc/system` file.


```
set zfs:zfs_unmap_ignore_size=0
```
- Reboot the system.

Tuning ZFS for Database Products

Review the following considerations when using ZFS with a database product.

- If the database uses a fixed disk block or record size for I/O, set the ZFS `recordsize` property to match it. You can do this on a per-file system basis, even though multiple file systems might share a single pool.
- With ZFS's copy-on-write design, tuning down the `recordsize` is a way to improve OLTP performance at the expense of batch reporting queries.

- ZFS checksums every block stored on disk. This alleviates the need for the database layer to checksum data an additional time. If checksums are computed by ZFS instead of at the database layer, any discrepancy can be caught and fixed before the data is returned to the application.
- UFS direct I/O is used to overcome some of the design deficiencies of UFS and to eliminate double buffering of data. In ZFS, the UFS design deficiencies do not exist and ZFS uses the `primarycache` and `secondarycache` properties to manage buffering data in the ARC. Note that using the `secondarycache` (`L2ARC`) property to improve random reads also requires the `primarycache` property to be enabled.
- Keep pool space under 90% utilization to maintain pool performance.

Tuning ZFS for an Oracle Database

ZFS is recommended for any Oracle database version in single instance mode. ZFS can be used with an Oracle RAC database when it is available as a NFS-shared file system.

Review the following recommendations below for tuning ZFS for an Oracle database:

- **Verify that you are running the latest Solaris release**
Start with the latest Solaris 10 or Solaris 11 release, with the Solaris 10 9/10 release as a minimum starting point.
- **Create LUNs for your ZFS storage pools, if needed**
Use your storage array tools to create LUNs that will be presented to the ZFS storage pool. Or, consider using whole disks for your mirrored ZFS storage pools. For more information, see [Chapter 3, “Managing Oracle Solaris ZFS Storage Pools,”](#) in *Oracle Solaris 11.1 Administration: ZFS File Systems*.
- **Create a storage pool for data files for tables, index, undo and temp data**
Consider creating a mirrored storage pool to provide a higher level of data redundancy. For example:

```
# zpool status dbpool
pool: dbpool
state: ONLINE
scan: none requested
config:
```

NAME	STATE	READ	WRITE	CKSUM
dbpool	ONLINE	0	0	0
mirror-0	ONLINE	0	0	0
c0t5000C500335F95E3d0	ONLINE	0	0	0
c0t5000C500335F907Fd0	ONLINE	0	0	0
mirror-1	ONLINE	0	0	0
c0t5000C500335BD117d0	ONLINE	0	0	0
c0t5000C500335DC60Fd0	ONLINE	0	0	0

errors: No known data errors

For databases with high redo log activity, such as a typical OLTP database with many commits, use a separate LUN for a separate log device.

- **Create a storage pool for the archivelog**

If available, a system's internal disk can handle this type of load. The archivelog file system can also be a file system in the dbpool.

```
# zpool create archivepool c0t5000C500335E106Bd0
```

- **Create the ZFS file systems and set the specific file system properties by using the following guidelines.**

Create separate file systems for redo, archive, undo, and temp database components using the default record size of 128 KB. The general rule is to set the file system `recordsize = db_block_size` for the file systems that contains Oracle data files. For table data and index components, create a file system with an 8 KB record size. Also consider providing metadata caching hints for your database file systems by using the `primarycache` property. For more information about ZFS file system properties, see [“Introducing ZFS Properties” in Oracle Solaris 11.1 Administration: ZFS File Systems](#).

- Create file systems for the table data files and index data files with an 8 KB `recordsize`. Use the default value for `primarycache`.

```
# zfs create -o recordsize=8k -o mountpoint=/my_db_path/index dbpool/index
# zfs set logbias=throughput dbpool/index
# zfs get primarycache,recordsize,logbias dbpool/index
NAME          PROPERTY  VALUE      SOURCE
dbpool/index  primarycache  all        default
dbpool/index  recordsize   8K         local
dbpool/index  logbias     throughput  local
```

- Create file systems for temporary and undo table spaces, using the default `recordsize` and `primarycache` values.

```
# zfs create -o mountpoint=/my_db_path/temp dbpool/temp
# zfs set logbias=throughput dbpool/temp
# zfs create -o mountpoint=/my_db_path/undo dbpool/undo
# zfs set logbias=throughput dbpool/undo
```

- Create a storage pool for redo logs with a separate log device. For databases with high redo log activity, such as a typical OLTP database with many commits, use a separate log device LUN.

Partition the disk into two slices, a small slice, `s0`, in the 64 to 150 MB range, for the separate log device. The `s1` slice contains the remaining disk space for the redo log.

```
# zpool create redopool c0t50015179594B6F11d0s1 log c0t50015179594B6F11d0s0
# zpool status redopool
  pool: redopool
  state: ONLINE
  scan: none requested
  config:
```

NAME	STATE	READ	WRITE	CKSUM
redopool	ONLINE	0	0	0
c0t50015179594B6F11d0s1	ONLINE	0	0	0
logs				
c0t50015179594B6F11d0s0	ONLINE	0	0	0

errors: No known data errors

- Create a file system for redo logs in the redo pool. Use default file system values for recordsize and primarycache.

```
# zfs create -o mountpoint=/my_db_path/redo redopool/redo
# zfs set logbias=latency redopool/redo
```

- Create a file system for archivelog files in the archive pool, enabling compression, use the default value for recordsize and set primarycache to metadata.

```
# zfs create -o compression=on -o primarycache=metadata -o mountpoint=
/my_db_admin_path/archive archivepool/archive
# zfs get primarycache,recordsize,compressratio,compression,available,
used,quota archivepool/archive
```

NAME	PROPERTY	VALUE	SOURCE
archivepool/archive	primarycache	metadata	local
archivepool/archive	recordsize	128K	default
archivepool/archive	compressratio	1.32x	-
archivepool/archive	compression	on	local
archivepool/archive	available	40.0G	-
archivepool/archive	used	10.0G	-
archivepool/archive	quota	50G	local

- **Consider tuning storage array I/O queues (for systems with HDS or EMC storage arrays)**

ZFS aggregates read and write I/O and manages the priority of I/O before sending it to the driver level, which handles the device. The `zfs_vdev_max_pending` parameter defines the maximum number of I/Os that ZFS sends to any storage pool device.

In a legacy storage environment, the `ssd_max_throttle` and `sd_max_throttle` parameters define the maximum number of concurrent I/Os that the driver can send to the storage. By setting the `zfs_vdev_max_pending` default value equal to the value of the `[s]sd_max_throttle` parameter, we prevent ZFS from queuing I/O to yet another unnecessary SD layer.

If you have `ssd:ssd_max_throttle` or `sd:sd_max_throttle` in the `/etc/system` file in your existing environment, then set `zfs:zfs_vdev_max_pending` at the same value. For example, if the storage array administrator asked for the following setting:

```
set ssd:ssd_max_throttle=20
```

Then, also set this parameter as follows:

```
set ssd:ssd_max_throttle=20
set zfs:zfs_vdev_max_pending=20
```

Setting this parameter allows ZFS to control each LUN queue. This means that the total number of pending I/Os in the storage can grow as follows:

number of LUNs * ZFS_VDEV_MAX_PENDING

- **Allocate sufficient memory and swap resources**

You can reduce ZFS memory consumption by tuning the `zfs_arc_max` parameter to a low value, but we still recommend provisioning enough memory to cache metadata for the actively used portion of the database, which is estimated at 1.5% with an 8 KB ZFS record size and proportionately less or more with larger or smaller records. The file system that hold index files is the one that has the largest benefit from file system caching because it is the last one to invalidate in case of lack of memory. The `zfs_arc_max` parameter is in bytes and accepts decimal or hexadecimal values. The following example sets this parameter to 2 GB:

```
set zfs:zfs_arc_max=2147483648
or
set zfs:zfs_arc_max=0x80000000
```

To prevent applications from failing due to lack of memory, you must configure some amount of swap space. The amount of swap equivalent to all of system memory is always enough for this purpose. This swap space is not expected to be used, but is needed as a reservation area. For information about increasing swap space, see “Managing Your ZFS Swap and Dump Devices” in *Oracle Solaris 11.1 Administration: ZFS File Systems*.

- Additional Oracle database configuration recommendations

- *Configuring Your Oracle Database on ZFS File Systems* in the following white paper:

<http://www.oracle.com/technetwork/server-storage/solaris/config-solaris-zfs-wp-167894.pdf>

- *Dynamic SGA Tuning of Oracle Database on Oracle Solaris with DISM* white paper:

<http://www.oracle.com/technetwork/articles/systems-hardware-architecture/using-dynamic-intimate-memory-sparc-168402.pdf>

- Oracle 11g Installation Guides

- Oracle Database Quick Installation Guide 11g Release 2 (11.2) for Oracle Solaris on SPARC (64-Bit)

http://docs.oracle.com/cd/E11882_01/install.112/e24349/toc.htm

- Oracle Database Quick Installation Guide 11g Release 2 (11.2) for Oracle Solaris on x86-64 (64-Bit)

http://docs.oracle.com/cd/E11882_01/install.112/e24351/toc.htm

Using ZFS with MySQL Considerations

Review the following considerations when using ZFS with MySQL.

- **ZFS recordsize**

Match the ZFS `recordsize` property to the storage engine block size for better OLTP performance.

- **InnoDB**
 - With a known application memory footprint, such as for a database application, you might cap the ARC size so that the application will not need to reclaim its necessary memory from the ZFS cache.
 - Create a separate pool for the logs.
 - Set a different path for data and log in the `my.cnf` file.
 - Set the ZFS `recordsize` property to 16K for the InnoDB data files, and use the default `recordsize` value for InnoDB logs, prior to creating data files.

NFS Tunable Parameters

This section describes the NFS tunable parameters.

- [“Tuning the NFS Environment” on page 95](#)
- [“NFS Module Parameters” on page 96](#)
- [“rpcmod Module Parameters” on page 123](#)

Where to Find Tunable Parameter Information

Tunable Parameter	For Information
Oracle Solaris kernel tunable parameters	Chapter 2, “Oracle Solaris Kernel Tunable Parameters”
Oracle Solaris ZFS tunable parameters	Chapter 3, “Oracle Solaris ZFS Tunable Parameters”
Internet Protocol Suite tunable parameters	Chapter 5, “Internet Protocol Suite Tunable Parameters”

Tuning the NFS Environment

You can define NFS parameters in the `/etc/system` file, which is read during the boot process. Each parameter includes the name of its associated kernel module. For more information, see [“Tuning an Oracle Solaris System” on page 19](#).



Caution – The names of the parameters, the modules that they reside in, and the default values can change between releases. Check the documentation for the version of the active SunOS release before making changes or applying values from previous releases.

NFS Module Parameters

This section describes parameters related to the NFS kernel module.

nfs:nfs3_pathconf_disable_cache

Description	Controls the caching of pathconf information for NFS Version 3 mounted file systems.
Data Type	Integer (32-bit)
Default	0 (caching enabled)
Range	0 (caching enabled) or 1 (caching disabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	The pathconf information is cached on a per file basis. However, if the server can change the information for a specific file dynamically, use this parameter to disable caching. There is no mechanism for the client to validate its cache entry.
Commitment Level	Unstable

nfs:nfs4_pathconf_disable_cache

Description	Controls the caching of pathconf information for NFS Version 4 mounted file systems.
Data Type	Integer (32-bit)
Default	0 (caching enabled)
Range	0 (caching enabled) or 1 (caching disabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	The pathconf information is cached on a per file basis. However, if the server can change the information for a specific file dynamically, use this parameter to disable caching. There is no mechanism for the client to validate its cache entry.

Commitment Level Unstable

nfs:nfs_allow_preepoch_time

Description	<p>Controls whether files with incorrect or <i>negative</i> time stamps should be made visible on the client.</p> <p>Historically, neither the NFS client nor the NFS server would do any range checking on the file times being returned. The over-the-wire timestamp values are unsigned and 32-bits long. So, all values have been legal.</p> <p>The timestamp values on the 64-bit Solaris kernel are signed and 64-bits long. It is impossible to determine whether a time field represents a full 32-bit time or a negative time, that is, a time prior to January 1, 1970.</p> <p>It is impossible to determine whether to sign extend a time value when converting from 32 bits to 64 bits. The time value should be sign extended if the time value is truly a negative number. However, the time value should not be sign extended if it does truly represent a full 32-bit time value. This problem is resolved by simply disallowing full 32-bit time values.</p>
Data Type	Integer (32-bit)
Default	0 (32-bit time stamps disabled)
Range	0 (32-bit time stamps disabled) or 1 (32-bit time stamps enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	Even during normal operation, it is possible for the timestamp values on some files to be set very far in the future or very far in the past. If access to these files is preferred using NFS mounted file systems, set this parameter to 1 to allow the timestamp values to be passed through unchecked.
Commitment Level	Unstable

nfs:nfs_cots_timeo

Description	Controls the default RPC timeout for NFS version 2 mounted file systems using connection-oriented transports such as TCP for the transport protocol.
Data Type	Signed integer (32-bit)
Default	600 (60 seconds)
Range	0 to $2^{31} - 1$
Units	10th of seconds
Dynamic?	Yes, but the RPC timeout for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None
When to Change	<p>TCP does a good job ensuring requests and responses are delivered appropriately. However, if the round-trip times are very large in a particularly slow network, the NFS version 2 client might time out prematurely.</p> <p>Increase this parameter to prevent the client from timing out incorrectly. The range of values is very large, so increasing this value too much might result in situations where a retransmission is not detected for long periods of time.</p>
Commitment Level	Unstable

nfs:nfs3_cots_timeo

Description	Controls the default RPC timeout for NFS version 3 mounted file systems using connection-oriented transports such as TCP for the transport protocol.
Data Type	Signed integer (32-bit)
Default	600 (60 seconds)
Range	0 to $2^{31} - 1$
Units	10th of seconds
Dynamic?	Yes, but the RPC timeout for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.

Validation	None
When to Change	TCP does a good job ensuring requests and responses are delivered appropriately. However, if the round-trip times are very large in a particularly slow network, the NFS version 3 client might time out prematurely. Increase this parameter to prevent the client from timing out incorrectly. The range of values is very large, so increasing this value too much might result in situations where a retransmission is not detected for long periods of time.
Commitment Level	Unstable

nfs:nfs4_cots_timeo

Description	Controls the default RPC timeout for NFS version 4 mounted file systems using connection-oriented transports such as TCP for the transport protocol. The NFS Version 4 protocol specification disallows retransmission over the same TCP connection. Thus, this parameter primarily controls how quickly the client responds to certain events, such as detecting a forced unmount operation or detecting how quickly the server fails over to a new server.
Data Type	Signed integer (32-bit)
Default	600 (60 seconds)
Range	0 to $2^{31} - 1$
Units	10th of seconds
Dynamic?	Yes, but this parameter is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None
When to Change	TCP does a good job ensuring requests and responses are delivered appropriately. However, if the round-trip times are very large in a particularly slow network, the NFS version 4 client might time out prematurely.

Increase this parameter to prevent the client from timing out incorrectly. The range of values is very large, so increasing this value too much might result in situations where a retransmission is not detected for long periods of time.

Commitment Level Unstable

nfs:nfs_do_symlink_cache

Description	Controls whether the contents of symbolic link files are cached for NFS version 2 mounted file systems.
Data Type	Integer (32-bit)
Default	1 (caching enabled)
Range	0 (caching disabled) or 1 (caching enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	If a server changes the contents of a symbolic link file without updating the modification timestamp on the file or if the granularity of the timestamp is too large, then changes to the contents of the symbolic link file might not be visible on the client for extended periods. In this case, use this parameter to disable the caching of symbolic link contents. Doing so makes the changes immediately visible to applications running on the client.
Commitment Level	Unstable

nfs:nfs3_do_symlink_cache

Description	Controls whether the contents of symbolic link files are cached for NFS version 3 mounted file systems.
Data Type	Integer (32-bit)
Default	1 (caching enabled)
Range	0 (caching disabled) or 1 (caching enabled)
Units	Boolean values
Dynamic?	Yes

Validation	None
When to Change	If a server changes the contents of a symbolic link file without updating the modification timestamp on the file or if the granularity of the timestamp is too large, then changes to the contents of the symbolic link file might not be visible on the client for extended periods. In this case, use this parameter to disable the caching of symbolic link contents. Doing so makes the changes immediately visible to applications running on the client.
Commitment Level	Unstable

nfs:nfs4_do_symlink_cache

Description	Controls whether the contents of symbolic link files are cached for NFS version 4 mounted file systems.
Data Type	Integer (32-bit)
Default	1 (caching enabled)
Range	0 (caching disabled) or 1 (caching enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	If a server changes the contents of a symbolic link file without updating the modification timestamp on the file or if the granularity of the timestamp is too large, then changes to the contents of the symbolic link file might not be visible on the client for extended periods. In this case, use this parameter to disable the caching of symbolic link contents. Doing so makes the changes immediately visible to applications running on the client.
Commitment Level	Unstable

nfs:nfs_dynamic

Description	Controls whether a feature known as <i>dynamic retransmission</i> is enabled for NFS version 2 mounted file systems using connectionless transports such as UDP. This feature attempts to reduce retransmissions by monitoring server response times and then adjusting RPC timeouts and read- and write- transfer sizes.
-------------	---

Data Type	Integer (32-bit)
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None
When to Change	Do not change this parameter.
Commitment Level	Unstable

nfs:nfs3_dynamic

Description	Controls whether a feature known as <i>dynamic retransmission</i> is enabled for NFS version 3 mounted file systems using connectionless transports such as UDP. This feature attempts to reduce retransmissions by monitoring server response times and then adjusting RPC timeouts and read- and write- transfer sizes.
Data Type	Integer (32-bit)
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Units	Boolean values
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None
When to Change	Do not change this parameter.
Commitment Level	Unstable

nfs:nfs_lookup_neg_cache

Description	Controls whether a negative name cache is used for NFS version 2 mounted file systems. This negative name cache records file names that
-------------	---

	were looked up, but not found. The cache is used to avoid over-the-network look-up requests made for file names that are already known to not exist.
Data Type	Integer (32-bit)
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	For the cache to perform correctly, negative entries must be strictly verified before they are used. This consistency mechanism is relaxed slightly for read-only mounted file systems. It is assumed that the file system on the server is not changing or is changing very slowly, and that it is okay for such changes to propagate slowly to the client. The consistency mechanism becomes the normal attribute cache mechanism in this case. If file systems are mounted read-only on the client, but are expected to change on the server and these changes need to be seen immediately by the client, use this parameter to disable the negative cache. If you disable the <code>nfs:nfs_disable_rmdir_cache</code> parameter, you should probably also disable this parameter. For more information, see “nfs:nfs_disable_rmdir_cache” on page 113 .
Commitment Level	Unstable

nfs:nfs3_lookup_neg_cache

Description	Controls whether a negative name cache is used for NFS version 3 read-only mounted file systems. This negative name cache records file names that were looked up, but were not found. The cache is used to avoid over-the-network look-up requests made for file names that are already known to not exist.
Data Type	Integer (32-bit)
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Units	Boolean values

Dynamic?	Yes
Validation	None
When to Change	<p>For the cache to perform correctly, negative entries must be strictly verified before they are used. This consistency mechanism is relaxed slightly for read-only mounted file systems. It is assumed that the file system on the server is not changing or is changing very slowly, and that it is okay for such changes to propagate slowly to the client. The consistency mechanism becomes the normal attribute cache mechanism in this case.</p> <p>Negative cache entries are used only for read-only mounted file system. By assuming that the file system on the server is not changing or is changing very slowly, it is okay for such changes to propagate slowly to the client. The consistency mechanism becomes the normal attribute cache mechanism in this case.</p> <p>If file systems are mounted read-only on the client, but are expected to change on the server and these changes need to be seen immediately by the client, use this parameter to disable the negative cache.</p> <p>If you disable the <code>nfs:nfs_disable_rmdir_cache</code> parameter, you should probably also disable this parameter. For more information, see “nfs:nfs_disable_rmdir_cache” on page 113.</p>
Commitment Level	Unstable

nfs:nfs4_lookup_neg_cache

Description	Controls whether a negative name cache is used for NFS version 4 mounted file systems. This negative name cache records file names that were looked up, but were not found. The cache is used to avoid over-the-network look-up requests made for file names that are already known to not exist.
Data Type	Integer (32-bit)
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None

When to Change	<p>For the cache to perform correctly, negative entries must be strictly verified before they are used. This consistency mechanism is relaxed slightly for read-only mounted file systems. It is assumed that the file system on the server is not changing or is changing very slowly, and that it is okay for such changes to propagate slowly to the client. The consistency mechanism becomes the normal attribute cache mechanism in this case.</p> <p>If file systems are mounted read-only on the client, but are expected to change on the server and these changes need to be seen immediately by the client, use this parameter to disable the negative cache.</p> <p>If you disable the <code>nfs:nfs_disable_rmdir_cache</code> parameter, you should probably also disable this parameter. For more information, see “nfs:nfs_disable_rmdir_cache” on page 113.</p>
Commitment Level	Unstable

nfs:nfs_max_threads

Description	<p>Controls the number of kernel threads that perform asynchronous I/O for the NFS version 2 client. Because NFS is based on RPC and RPC is inherently synchronous, separate execution contexts are required to perform NFS operations that are asynchronous from the calling thread.</p> <p>The operations that can be executed asynchronously are read for read-ahead, readdir for readdir read-ahead, write for putpage and pageio operations, commit, and inactive for cleanup operations that the client performs when it stops using a file.</p>
Data Type	Unsigned short
Default	8
Range	0 to $2^{15} - 1$
Units	Threads
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None
When to Change	To increase or reduce the number of simultaneous I/O operations that are outstanding at any given time. For example, for a very low bandwidth network, you might want to decrease this value so that the

NFS client does not overload the network. Alternately, if the network is very high bandwidth, and the client and server have sufficient resources, you might want to increase this value. Doing so can more effectively utilize the available network bandwidth, and the client and server resources.

Commitment Level Unstable

nfs:nfs3_max_threads

Description Controls the number of kernel threads that perform asynchronous I/O for the NFS version 3 client. Because NFS is based on RPC and RPC is inherently synchronous, separate execution contexts are required to perform NFS operations that are asynchronous from the calling thread.

The operations that can be executed asynchronously are read for read-ahead, readdir for readdir read-ahead, write for putpage and pageio requests, and commit.

Data Type Unsigned short

Default 8

Range 0 to $2^{15} - 1$

Units Threads

Dynamic? Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.

Validation None

When to Change To increase or reduce the number of simultaneous I/O operations that are outstanding at any given time. For example, for a very low bandwidth network, you might want to decrease this value so that the NFS client does not overload the network. Alternately, if the network is very high bandwidth, and the client and server have sufficient resources, you might want to increase this value. Doing so can more effectively utilize the available network bandwidth, and the client and server resources.

Commitment Level Unstable

nfs:nfs4_max_threads

Description	Controls the number of kernel threads that perform asynchronous I/O for the NFS version 4 client. Because NFS is based on RPC and RPC is inherently synchronous, separate execution contexts are required to perform NFS operations that are asynchronous from the calling thread. The operations that can be executed asynchronously are read for read-ahead, write-behind, directory read-ahead, and cleanup operations that the client performs when it stops using a file.
Data Type	Unsigned short
Default	8
Range	0 to $2^{15} - 1$
Units	Threads
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None
When to Change	To increase or reduce the number of simultaneous I/O operations that are outstanding at any given time. For example, for a very low bandwidth network, you might want to decrease this value so that the NFS client does not overload the network. Alternately, if the network is very high bandwidth, and the client and server have sufficient resources, you might want to increase this value. Doing so can more effectively utilize the available network bandwidth, and the client and server resources.
Commitment Level	Unstable

nfs:nfs_nra

Description	Controls the number of read-ahead operations that are queued by the NFS version 2 client when sequential access to a file is discovered. These read-ahead operations increase concurrency and read throughput. Each read-ahead request is generally for one logical block of file data.
Data Type	Integer (32-bit)
Default	4

Range	0 to $2^{31} - 1$
Units	Logical blocks.
Dynamic?	Yes
Validation	None
When to Change	To increase or reduce the number of read-ahead requests that are outstanding for a specific file at any given time. For example, for a very low bandwidth network or on a low memory client, you might want to decrease this value so that the NFS client does not overload the network or the system memory. Alternately, if the network is very high bandwidth, and the client and server have sufficient resources, you might want to increase this value. Doing so can more effectively utilize the available network bandwidth, and the client and server resources.
Commitment Level	Unstable

nfs:nfs3_nra

Description	Controls the number of read-ahead operations that are queued by the NFS version 3 client when sequential access to a file is discovered. These read-ahead operations increase concurrency and read throughput. Each read-ahead request is generally for one logical block of file data.
Data Type	Integer (32-bit)
Default	4
Range	0 to $2^{31} - 1$
Units	Logical blocks. (See “ nfs:nfs3_bsize ” on page 114.)
Dynamic?	Yes
Validation	None
When to Change	To increase or reduce the number of read-ahead requests that are outstanding for a specific file at any given time. For example, for a very low bandwidth network or on a low memory client, you might want to decrease this value so that the NFS client does not overload the network or the system memory. Alternately, if the network is very high bandwidth and the client and server have sufficient resources, you might want to increase this value. Doing so can more effectively utilize the available network bandwidth, and the client and server resources.
Commitment Level	Unstable

nfs:nfs4_nra

Description	Controls the number of read-ahead operations that are queued by the NFS version 4 client when sequential access to a file is discovered. These read-ahead operations increase concurrency and read throughput. Each read-ahead request is generally for one logical block of file data.
Data Type	Integer (32-bit)
Default	4
Range	0 to $2^{31} - 1$
Units	Logical blocks. (See “ nfs:nfs4_bsize ” on page 114.)
Dynamic?	Yes
Validation	None
When to Change	To increase or reduce the number of read-ahead requests that are outstanding for a specific file at any given time. For example, for a very low bandwidth network or on a low memory client, you might want to decrease this value so that the NFS client does not overload the network or the system memory. Alternately, if the network is very high bandwidth, and the client and server have sufficient resources, you might want to increase this value. Doing so can more effectively utilize the available network bandwidth, and the client and server resources.
Commitment Level	Unstable

nfs:nrnode

Description	Controls the size of the rnode cache on the NFS client. The rnode, used by both NFS version 2, 3, and 4 clients, is the central data structure that describes a file on the NFS client. The rnode contains the file handle that identifies the file on the server. The rnode also contains pointers to various caches used by the NFS client to avoid network calls to the server. Each rnode has a one-to-one association with a vnode. The vnode caches file data. The NFS client attempts to maintain a minimum number of rnodes to attempt to avoid destroying cached data and metadata. When an rnode is reused or freed, the cached data and metadata must be destroyed.
Data Type	Integer (32-bit)

Default	The default setting of this parameter is 0, which means that the value of <code>nnode</code> should be set to the value of the <code>ncsize</code> parameter. Actually, any non positive value of <code>nnode</code> results in <code>nnode</code> being set to the value of <code>ncsize</code> .
Range	1 to $2^{31} - 1$
Units	rnodes
Dynamic?	No. This value can only be changed by adding or changing the parameter in the <code>/etc/system</code> file, and then rebooting the system.
Validation	The system enforces a maximum value such that the <code>rnode</code> cache can only consume 25 percent of available memory.
When to Change	<p>Because <code>rnodes</code> are created and destroyed dynamically, the system tends to settle upon a <code>nnode</code>-size cache, automatically adjusting the size of the cache as memory pressure on the system increases or as more files are simultaneously accessed. However, in certain situations, you could set the value of <code>nnode</code> if the mix of files being accessed can be predicted in advance. For example, if the NFS client is accessing a few very large files, you could set the value of <code>nnode</code> to a small number so that system memory can cache file data instead of <code>rnodes</code>. Alternately, if the client is accessing many small files, you could increase the value of <code>nnode</code> to optimize for storing file metadata to reduce the number of network calls for metadata.</p> <p>Although it is not recommended, the <code>rnode</code> cache can be effectively disabled by setting the value of <code>nnode</code> to 1. This value instructs the client to only cache 1 <code>rnode</code>, which means that it is reused frequently.</p>
Commitment Level	Unstable

nfs:nfs_shrinkreaddir

Description	<p>Some older NFS servers might incorrectly handle NFS version 2 REaddir requests for more than 1024 bytes of directory information. This problem is due to a bug in the server implementation. However, this parameter contains a workaround in the NFS version 2 client.</p> <p>When this parameter is enabled, the client does not generate a REaddir request for larger than 1024 bytes of directory information. If this parameter is disabled, then the over-the-wire size is set to the lesser of either the size passed in by using the <code>getdents</code> system call or by using <code>NFS_MAXDATA</code>, which is 8192 bytes. For more information, see getdents(2).</p>
-------------	---

Data Type	Integer (32-bit)
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	Examine the value of this parameter if an older NFS version 2 only server is used and interoperability problems occur when the server tries to read directories. Enabling this parameter might cause a slight decrease in performance for applications that read directories.
Commitment Level	Unstable

nfs:nfs3_shrinkreaddir

Description	<p>Some older NFS servers might incorrectly handle NFS version 3 READDIR requests for more than 1024 bytes of directory information. This problem is due to a bug in the server implementation. However, this parameter contains a workaround in the NFS version 3 client.</p> <p>When this parameter is enabled, the client does not generate a READDIR request for larger than 1024 bytes of directory information. If this parameter is disabled, then the over-the-wire size is set to the minimum of either the size passed in by using the <code>getdents</code> system call or by using <code>MAXBSIZE</code>, which is 8192 bytes. For more information, see getdents(2).</p>
Data Type	Integer (32-bit)
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	Examine the value of this parameter if an older NFS version 3 only server is used and interoperability problems occur when the server tries to read directories. Enabling this parameter might cause a slight decrease in performance for applications that read directories.

Commitment Level Unstable

nfs:nfs_write_error_interval

Description	Controls the time duration in between logging ENOSPC and EDQUOT write errors received by the NFS client. This parameter affects NFS version 2, 3, and 4 clients.
Data Type	Long integer (64-bit)
Default	5 seconds
Range	0 to $2^{63} - 1$
Units	Seconds
Dynamic?	Yes
Validation	None
When to Change	Increase or decrease the value of this parameter in response to the volume of messages being logged by the client. Typically, you might want to increase the value of this parameter to decrease the number of out of space messages being printed when a full file system on a server is being actively used.
Commitment Level	Unstable

nfs:nfs_write_error_to_cons_only

Description	Controls whether NFS write errors are logged to the system console and <code>syslog</code> or to the system console only. This parameter affects messages for NFS version 2, 3, and 4 clients.
Data Type	Integer (32-bit)
Default	0 (system console and <code>syslog</code>)
Range	0 (system console and <code>syslog</code>) or 1 (system console)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	Examine the value of this parameter to avoid filling up the file system containing the messages logged by the <code>syslogd</code> daemon. When this

parameter is enabled, messages are printed on the system console only and are not copied to the `syslog` messages file.

Commitment Level Unstable

nfs:nfs_disable_rddir_cache

Description	Controls the use of a cache to hold responses from READDIR and READDIRPLUS requests. This cache avoids over-the-wire calls to the server to retrieve directory information.
Data Type	Integer (32-bit)
Default	0 (caching enabled)
Range	0 (caching enabled) or 1 (caching disabled)
Units	Boolean values
Dynamic?	Yes
Validation	None
When to Change	<p>Examine the value of this parameter if interoperability problems develop due to a server that does not update the modification time on a directory when a file or directory is created in it or removed from it. The symptoms are that new names do not appear in directory listings after they have been added to the directory or that old names do not disappear after they have been removed from the directory.</p> <p>This parameter controls the caching for NFS version 2, 3, and 4 mounted file systems. This parameter applies to all NFS mounted file systems, so caching cannot be disabled or enabled on a per file system basis.</p> <p>If you disable this parameter, you should also disable the following parameters to prevent bad entries in the DNLC negative cache:</p> <ul style="list-style-type: none"> ▪ “<code>nfs:nfs_lookup_neg_cache</code>” on page 102 ▪ “<code>nfs:nfs3_lookup_neg_cache</code>” on page 103 ▪ “<code>nfs:nfs4_lookup_neg_cache</code>” on page 104
Commitment Level	Unstable

nfs:nfs3_bsize

Description	Controls the logical block size used by the NFS version 3 client. This block size represents the amount of data that the client attempts to read from or write to the server when it needs to do an I/O.
Data Type	Unsigned integer (32-bit)
Default	32,768 (32 KB)
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, but the block size for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None. Setting this parameter too low or too high might cause the system to malfunction. Do not set this parameter to anything less than PAGESIZE for the specific platform. Do not set this parameter too high because it might cause the system to hang while waiting for memory allocations to be granted.
When to Change	Examine the value of this parameter when attempting to change the maximum data transfer size. Change this parameter in conjunction with the <code>nfs:nfs3_max_transfer_size</code> parameter. If larger transfers are preferred, increase both parameters. If smaller transfers are preferred, then just reducing this parameter should suffice.
Commitment Level	Unstable

nfs:nfs4_bsize

Description	Controls the logical block size used by the NFS version 4 client. This block size represents the amount of data that the client attempts to read from or write to the server when it needs to do an I/O.
Data Type	Unsigned integer (32-bit)
Default	32,768 (32 KB)
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, but the block size for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.

Validation	None. Setting this parameter too low or too high might cause the system to malfunction. Do not set this parameter to anything less than <code>PAGESIZE</code> for the specific platform. Do not set this parameter too high because it might cause the system to hang while waiting for memory allocations to be granted.
When to Change	Examine the value of this parameter when attempting to change the maximum data transfer size. Change this parameter in conjunction with the <code>nfs:nfs4_max_transfer_size</code> parameter. If larger transfers are preferred, increase both parameters. If smaller transfers are preferred, then just reducing this parameter should suffice.
Commitment Level	Unstable

nfs:nfs_async_clusters

Description	<p>Controls the mix of asynchronous requests that are generated by the NFS version 2 client. The four types of asynchronous requests are read-ahead, putpage, pageio, and readdir-ahead. The client attempts to round-robin between these different request types to attempt to be fair and not starve one request type in favor of another.</p> <p>However, the functionality in some NFS version 2 servers such as write gathering depends upon certain behaviors of existing NFS Version 2 clients. In particular, this functionality depends upon the client sending out multiple WRITE requests at about the same time. If one request is taken out of the queue at a time, the client would be defeating this server functionality designed to enhance performance for the client.</p> <p>Thus, use this parameter to control the number of requests of each request type that are sent out before changing types.</p>
Data Type	Unsigned integer (32-bit)
Default	1
Range	0 to $2^{31} - 1$
Units	Asynchronous requests
Dynamic?	Yes, but the cluster setting for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.

Validation	None. However, setting the value of this parameter to 0 causes all of the queued requests of a particular request type to be processed before moving on to the next type. This effectively disables the fairness portion of the algorithm.
When to Change	To increase the number of each type of asynchronous request that is generated before switching to the next type. Doing so might help with server functionality that depends upon clusters of requests coming from the client.
Commitment Level	Unstable

nfs:nfs3_async_clusters

Description	<p>Controls the mix of asynchronous requests that are generated by the NFS version 3 client. The five types of asynchronous requests are read-ahead, putpage, pageio, readdir-ahead, and commit. The client attempts to round-robin between these different request types to attempt to be fair and not starve one request type in favor of another.</p> <p>However, the functionality in some NFS version 3 servers such as write gathering depends upon certain behaviors of existing NFS version 3 clients. In particular, this functionality depends upon the client sending out multiple WRITE requests at about the same time. If one request is taken out of the queue at a time, the client would be defeating this server functionality designed to enhance performance for the client.</p> <p>Thus, use this parameter to control the number of requests of each request type that are sent out before changing types.</p>
Data Type	Unsigned integer (32-bit)
Default	1
Range	0 to $2^{31} - 1$
Units	Asynchronous requests
Dynamic?	Yes, but the cluster setting for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None. However, setting the value of this parameter to 0 causes all of the queued requests of a particular request type to be processed before moving on to the next type. This value effectively disables the fairness portion of the algorithm.

When to Change	To increase the number of each type of asynchronous operation that is generated before switching to the next type. Doing so might help with server functionality that depends upon clusters of operations coming from the client.
Commitment Level	Unstable

nfs:nfs4_async_clusters

Description	<p>Controls the mix of asynchronous requests that are generated by the NFS version 4 client. The six types of asynchronous requests are read-ahead, putpage, pageio, readdir-ahead, commit, and inactive. The client attempts to round-robin between these different request types to attempt to be fair and not starve one request type in favor of another.</p> <p>However, the functionality in some NFS version 4 servers such as write gathering depends upon certain behaviors of existing NFS version 4 clients. In particular, this functionality depends upon the client sending out multiple WRITE requests at about the same time. If one request is taken out of the queue at a time, the client would be defeating this server functionality designed to enhance performance for the client.</p> <p>Thus, use this parameter to control the number of requests of each request type that are sent out before changing types.</p>
Data Type	Unsigned integer (32-bit)
Default	1
Range	0 to $2^{31} - 1$
Units	Asynchronous requests
Dynamic?	Yes, but the cluster setting for a file system is set when the file system is mounted. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None. However, setting the value of this parameter to 0 causes all of the queued requests of a particular request type to be processed before moving on to the next type. This effectively disables the fairness portion of the algorithm.
When to Change	To increase the number of each type of asynchronous request that is generated before switching to the next type. Doing so might help with server functionality that depends upon clusters of requests coming from the client.

Commitment Level Unstable

nfs:nfs_async_timeout

Description	Controls the duration of time that threads, which execute asynchronous I/O requests, sleep with nothing to do before exiting. When there are no more requests to execute, each thread goes to sleep. If no new requests come in before this timer expires, the thread wakes up and exits. If a request does arrive, a thread is woken up to execute requests until there are none again. Then, the thread goes back to sleep waiting for another request to arrive, or for the timer to expire.
Data Type	Integer (32-bit)
Default	6000 (1 minute expressed as 60 sec * 100Hz)
Range	0 to $2^{31} - 1$
Units	Hz. (Typically, the clock runs at 100Hz.)
Dynamic?	Yes
Validation	None. However, setting this parameter to a non positive value causes these threads exit as soon as there are no requests in the queue for them to process.
When to Change	If the behavior of applications in the system is known precisely and the rate of asynchronous I/O requests can be predicted, it might be possible to tune this parameter to optimize performance slightly in either of the following ways: <ul style="list-style-type: none">▪ By making the threads expire more quickly, thus freeing up kernel resources more quickly▪ By making the threads expire more slowly, thus avoiding thread create and destroy overhead
Commitment Level	Unstable

nfs:nacache

Description	Tunes the number of hash queues that access the file access cache on the NFS client. The file access cache stores file access rights that users have with respect to files that they are trying to access. The cache itself is dynamically allocated. However, the hash queues used to index into the cache are statically allocated. The algorithm assumes that there is one access cache entry per active file and four of these access cache
-------------	---

	entries per hash bucket. Thus, by default, the value of this parameter is set to the value of the <code>nnode</code> parameter.
Data Type	Integer (32-bit)
Default	The default setting of this parameter is 0. This value means that the value of <code>nacache</code> should be set to the value of the <code>nnode</code> parameter.
Range	1 to $2^{31} - 1$
Units	Access cache entries
Dynamic?	No. This value can only be changed by adding or changing the parameter in the <code>/etc/system</code> file, and then rebooting system.
Validation	None. However, setting this parameter to a negative value will probably cause the system to try to allocate a very large set of hash queues. While trying to do so, the system is likely to hang.
When to Change	Examine the value of this parameter if the basic assumption of one access cache entry per file would be violated. This violation could occur for systems in a timesharing mode where multiple users are accessing the same file at about the same time. In this case, it might be helpful to increase the expected size of the access cache so that the hashed access to the cache stays efficient.
Commitment Level	Unstable

nfs:nfs3_jukebox_delay

Description	Controls the duration of time that the NFS version 3 client waits to transmit a new request after receiving the <code>NFS3ERR_JUKEBOX</code> error from a previous request. The <code>NFS3ERR_JUKEBOX</code> error is generally returned from the server when the file is temporarily unavailable for some reason. This error is generally associated with hierarchical storage, and CD or tape jukeboxes.
Data Type	Long integer (64-bit)
Default	1000 (10 seconds expressed as $10 \text{ sec} * 100\text{Hz}$)
Range	0 to $2^{63} - 1$ on 64-bit platforms
Units	Hz. (Typically, the clock runs at 100Hz.)
Dynamic?	Yes
Validation	None

When to Change	Examine the value of this parameter and perhaps adjust it to match the behaviors exhibited by the server. Increase this value if the delays in making the file available are long in order to reduce network overhead due to repeated retransmissions. Decrease this value to reduce the delay in discovering that the file has become available.
Commitment Level	Unstable

nfs:nfs3_max_transfer_size

Description	Controls the maximum size of the data portion of an NFS version 3 READ, WRITE, REaddir, or REaddirplus request. This parameter controls both the maximum size of the request that the server returns as well as the maximum size of the request that the client generates.
Data Type	Integer (32-bit)
Default	1,048,576 (1 Mbyte)
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None. However, setting the maximum transfer size on the server to 0 is likely to cause clients to malfunction or just decide not to attempt to talk to the server. There is also a limit on the maximum transfer size when using NFS over the UDP transport. UDP has a hard limit of 64 KB per datagram. This 64 KB must include the RPC header as well as other NFS information, in addition to the data portion of the request. Setting the limit too high might result in errors from UDP and communication problems between the client and the server.
When to Change	To tune the size of data transmitted over the network. In general, the <code>nfs:nfs3_bsize</code> parameter should also be updated to reflect changes in this parameter. For example, when you attempt to increase the transfer size beyond 32 KB, update <code>nfs:nfs3_bsize</code> to reflect the increased value. Otherwise, no change in the over-the-wire request size is observed. For more information, see “nfs:nfs3_bsize” on page 114 .

If you want to use a smaller transfer size than the default transfer size, use the mount command's `-wsize` or `-rsize` option on a per-file system basis.

Commitment Level Unstable

nfs:nfs4_max_transfer_size

Description	Controls the maximum size of the data portion of an NFS version 4 READ, WRITE, REaddir, or REaddirPLUS request. This parameter controls both the maximum size of the request that the server returns as well as the maximum size of the request that the client generates.
Data Type	Integer (32-bit)
Default	32,768 (32 KB)
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None. However, setting the maximum transfer size on the server to 0 is likely to cause clients to malfunction or just decide not to attempt to talk to the server. There is also a limit on the maximum transfer size when using NFS over the UDP transport. For more information on the maximum for UDP, see “ nfs:nfs3_max_transfer_size ” on page 120.
When to Change	To tune the size of data transmitted over the network. In general, the <code>nfs:nfs4_bsize</code> parameter should also be updated to reflect changes in this parameter. For example, when you attempt to increase the transfer size beyond 32 KB, update <code>nfs:nfs4_bsize</code> to reflect the increased value. Otherwise, no change in the over-the-wire request size is observed. For more information, see “ nfs:nfs4_bsize ” on page 114. If you want to use a smaller transfer size than the default transfer size, use the mount command's <code>-wsize</code> or <code>-rsize</code> option on a per-file system basis.
Commitment Level	Unstable

nfs:nfs3_max_transfer_size_clts

Description	Controls the maximum size of the data portion of an NFS version 3 READ, WRITE, REaddir, or REaddirPLUS request over UDP. This parameter controls both the maximum size of the request that the server returns as well as the maximum size of the request that the client generates.
Data Type	Integer (32-bit)
Default	32,768 (32 KB)
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.
Validation	None. However, setting the maximum transfer size on the server to 0 is likely to cause clients to malfunction or just decide not to attempt to talk to the server.
When to Change	Do not change this parameter.
Commitment Level	Unstable

nfs:nfs3_max_transfer_size_cots

Description	Controls the maximum size of the data portion of an NFS version 3 READ, WRITE, REaddir, or REaddirPLUS request over TCP. This parameter controls both the maximum size of the request that the server returns as well as the maximum size of the request that the client generates.
Data Type	Integer (32-bit)
Default	1,048,576 bytes
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, but this parameter is set per file system at mount time. To affect a particular file system, unmount and mount the file system after changing this parameter.

Validation	None. However, setting the maximum transfer size on the server to 0 is likely to cause clients to malfunction or just decide not to attempt to talk to the server.
When to Change	Do not change this parameter unless transfer sizes larger than 1 Mbyte are preferred.
Commitment Level	Unstable

rpcmod Module Parameters

This section describes NFS parameters for the `rpcmod` module.

rpcmod:clnt_max_conns

Description	Controls the number of TCP connections that the NFS client uses when communicating with each NFS server. The kernel RPC is constructed so that it can multiplex RPCs over a single connection. However, multiple connections can be used, if preferred.
Data Type	Integer (32-bit)
Default	1
Range	1 to $2^{31} - 1$
Units	Connections
Dynamic?	Yes
Validation	None
When to Change	In general, one connection is sufficient to achieve full network bandwidth. However, if TCP cannot utilize the bandwidth offered by the network in a single stream, then multiple connections might increase the throughput between the client and the server. Increasing the number of connections doesn't come without consequences. Increasing the number of connections also increases kernel resource usage needed to keep track of each connection.
Commitment Level	Unstable

rpcmod:clnt_idle_timeout

Description	Controls the duration of time on the client that a connection between the client and server is allowed to remain idle before being closed.
Data Type	Long integer (64-bit)
Default	300,000 milliseconds (5 minutes)
Range	0 to $2^{63} - 1$
Units	Milliseconds
Dynamic?	Yes
Validation	None
When to Change	Use this parameter to change the time that idle connections are allowed to exist on the client before being closed. You might want to close connections at a faster rate to avoid consuming system resources.
Commitment Level	Unstable

rpcmod:svc_idle_timeout

Description	Controls the duration of time on the server that a connection between the client and server is allowed to remain idle before being closed.
Data Type	Long integer (64-bit)
Default	360,000 milliseconds (6 minutes)
Range	0 to $2^{63} - 1$
Units	Milliseconds
Dynamic?	Yes
Validation	None
When to Change	Use this parameter to change the time that idle connections are allowed to exist on the server before being closed. You might want to close connections at a faster rate to avoid consuming system resources.
Commitment Level	Unstable

rpcmod:svc_default_stksize

Description	Sets the size of the kernel stack for kernel RPC service threads.
-------------	---

Data Type	Integer (32-bit)
Default	The default value is 0. This value means that the stack size is set to the system default.
Range	0 to $2^{31} - 1$
Units	Bytes
Dynamic?	Yes, for all new threads that are allocated. The stack size is set when the thread is created. Therefore, changes to this parameter do not affect existing threads but are applied to all new threads that are allocated.
Validation	None
When to Change	Very deep call depths can cause the stack to overflow and cause red zone faults. The combination of a fairly deep call depth for the transport, coupled with a deep call depth for the local file system, can cause NFS service threads to overflow their stacks. Set this parameter to a multiple of the hardware page size on the platform.
Commitment Level	Unstable

rpcmod:maxdupreqs

Description	Controls the size of the duplicate request cache that detects RPC- level retransmissions on connectionless transports. This cache is indexed by the client network address and the RPC procedure number, program number, version number, and transaction ID. This cache avoids processing retransmitted requests that might not be idempotent.
Data Type	Integer (32-bit)
Default	8192
Range	1 to $2^{31} - 1$
Units	Requests
Dynamic?	The cache is dynamically sized, but the hash queues that provide fast access to the cache are statically sized. Making the cache very large might result in long search times to find entries in the cache. Do not set the value of this parameter to 0. This value prevents the NFS server from handling non idempotent requests.
Validation	None

When to Change	<p>Examine the value of this parameter if false failures are encountered by NFS clients. For example, if an attempt to create a directory fails, but the directory is actually created, perhaps that retransmitted MKDIR request was not detected by the server.</p> <p>The size of the cache should match the load on the server. The cache records non idempotent requests and so only needs to track a portion of the total requests. The cache does need to hold the information long enough to be able to detect a retransmission by the client. Typically, the client timeout for connectionless transports is relatively short, starting around 1 second and increasing to about 20 seconds.</p>
Commitment Level	Unstable

rpcmod:cotsmaxdupreqs

Description	<p>Controls the size of the duplicate request cache that detects RPC- level retransmissions on connection-oriented transports. This cache is indexed by the client network address and the RPC procedure number, program number, version number, and transaction ID. This cache avoids processing retransmitted requests that might not be idempotent.</p>
Data Type	Integer (32-bit)
Default	8192
Range	1 to $2^{31} - 1$
Units	Requests
Dynamic?	Yes
Validation	<p>The cache is dynamically sized, but the hash queues that provide fast access to the cache are statically sized. Making the cache very large might result in long search times to find entries in the cache.</p> <p>Do not set the value of this parameter to 0. It prevents the NFS server from handling non-idempotent requests.</p>
When to Change	<p>Examine the value of this parameter if false failures are encountered by NFS clients. For example, if an attempt to create a directory fails, but the directory is actually created, it is possible that a retransmitted MKDIR request was not detected by the server.</p> <p>The size of the cache should match the load on the server. The cache records non-idempotent requests and so only needs to track a portion</p>

of the total requests. It does need to hold the information long enough to be able to detect a retransmission on the part of the client. Typically, the client timeout for connection oriented transports is very long, about 1 minute. Thus, entries need to stay in the cache for fairly long times.

Commitment Level Unstable

Internet Protocol Suite Tunable Parameters

This chapter describes various Internet Protocol suite properties.

- “IP Tunable Parameters” on page 130
- “TCP Tunable Parameters” on page 136
- “UDP Tunable Parameters” on page 154
- “IPQoS Tunable Parameter” on page 156
- “SCTP Tunable Parameters” on page 157
- “Per-Route Metrics” on page 168

Where to Find Tunable Parameter Information

Tunable Parameter	For Information
Oracle Solaris kernel tunable parameters	Chapter 2, “Oracle Solaris Kernel Tunable Parameters”
Oracle Solaris ZFS tunable parameters	Chapter 3, “Oracle Solaris ZFS Tunable Parameters”
NFS tunable parameters	Chapter 4, “NFS Tunable Parameters”

Overview of Tuning IP Suite Parameters

You can set all of the tuning parameters described in this chapter by using the `ipadm` command except for the following parameters:

- “`ipcl_conn_hash_size`” on page 149
- “`ip_queue_worker_wait`” on page 149
- “`ip_queue_fanout`” on page 134

These parameters can only be set in the `/etc/system` file.

Use the following syntax to set TCP/IP parameters by using the `ipadm` command:

```
# ipadm set-prop -p parameter ip|ipv4|ipv6|tcp|udp|sctp
```

For example:

```
# ipadm set-prop -p extra_priv_ports=1047 tcp
# ipadm show-prop -p extra_priv_ports tcp
PROTO PROPERTY          PERM CURRENT      PERSISTENT  DEFAULT      POSSIBLE
tcp  extra_priv_ports    rw   1047            1047         2049,4045    1-65535
```

For more information, see [ipadm\(1M\)](#).

Use the following syntax to set TCP/IP parameters by using the `ndd` command:

```
# ndd -set driver parameter value
```

For more information, see [ndd\(1M\)](#).

IP Suite Parameter Validation

All parameters described in this section are checked to verify that they fall in the parameter range. The parameter's range is provided with the description for each parameter.

Internet Request for Comments (RFCs)

Internet protocol and standard specifications are described in RFC documents. You can review RFCs from the following site:

<https://www.ietf.org/rfc.html>

At this site, you can browse RFC topics by entering an RFC number or an internet-draft file name in the IETF Repository Retrieval search field.

IP Tunable Parameters

`_icmp_err_interval` and `_icmp_err_burst`

Description Controls the rate of IP in generating ICMP error messages. IP generates only up to `_icmp_err_burst` IP error messages in any `_icmp_err_interval`.

The `_icmp_err_interval` parameter protects IP from denial of service attacks. Setting this parameter to 0 disables rate limiting. It does not disable the generation of error messages.

Default	100 milliseconds for <code>_icmp_err_interval</code> 10 error messages for <code>_icmp_err_burst</code>
Range	0 – 99,999 milliseconds for <code>_icmp_err_interval</code> 1 – 99,999 error messages for <code>_icmp_err_burst</code>
Dynamic?	Yes
When to Change	If you need a higher error message generation rate for diagnostic purposes.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

`_respond_to_echo_broadcast` and `_respond_to_echo_multicast (ipv4 or ipv6)`

Description	Controls whether IP responds to a broadcast ICMPv4 echo request or a IPv6 multicast ICMPv6 echo request.
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	If you do not want this behavior for security reasons, disable it.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

`_send_redirects (ipv4 or ipv6)`

Description	Controls whether IPv4 or IPv6 sends out ICMPv4 or ICMPv6 redirect messages.
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	If you do not want this behavior for security reasons, disable it.

Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

forwarding (ipv4 or ipv6)

Description	Controls whether IPv4 or IPv6 forwards packets with source IPv4 routing options or IPv6 routing headers.
Default	0 (disabled or off)
Range	0 (disabled or off) or 1 (enabled or on)
Dynamic?	Yes
When to Change	Keep this parameter disabled to prevent denial of service attacks.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

ttl

Description	Controls the time to live (TTL) value in the IPv4 header for the outbound IPv4 packets on an IP association.
Default	255
Range	1 to 255
Dynamic?	Yes
When to Change	Generally, you do not need to change this value.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

hoplimit (ipv6)

Description	Sets the value of the hop limit in the IPv6 header for the outbound IPv6 packets on an IP association.
Default	255

Range	0 to 255
Dynamic?	Yes
When to Change	Generally, you do not need to change this value.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

`_addr_per_if`

Description	Defines the maximum number of logical IP interfaces associated with a real interface.
Default	256
Range	1 to 8,192
Dynamic?	Yes
When to Change	Do not change the value. If more logical interfaces are required, you might consider increasing the value. However, recognize that this change might have a negative impact on IP's performance.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

`hostmodel (ipv4 or ipv6)`

Description	Controls send and receive behavior for IPv4 or IPv6 packets on a multi-homed system. This property can have the following values: <code>weak</code> , <code>strong</code> , and <code>src-priority</code> . The default value is <code>weak</code> .
Default	<code>weak</code>
Range	<code>weak</code> , <code>strong</code> , or <code>src-priority</code> <ul style="list-style-type: none"> ▪ <code>weak</code> <ul style="list-style-type: none"> ▪ Outgoing packets - The source address of the packet going out need not match the address configured on the outgoing interface.

- Incoming packets - The destination address of the incoming packet need not match the address configured on the incoming interface.
- strong
 - Outgoing packets - The source address of the packet going out must match the address configured on the outgoing interface.
 - Incoming packets - The destination address of the incoming packet must match the address configured on the incoming interface.
- src-priority
 - Outgoing packets - If multiple routes for the IP destination in the packet are available, the system prefers routes where the IP source address in the packet is configured on the outgoing interface.

If no such route is available, the system falls back to selecting the *best* route, as with the weak ES case.
 - Incoming packets - The destination address of the incoming packet must be configured on any one of the host's interface.

Dynamic? Yes

When to Change If a machine has interfaces that cross strict networking domains (for example, a firewall or a VPN node), set this parameter to strong.

Commitment Level Unstable

Change History For information, see [“IP Parameter Name Changes \(Oracle Solaris 11\)”](#) on page 180.

ip_squeue_fanout

Description Determines the mode of associating TCP/IP connections with squeues.

A value of 0 associates a new TCP/IP connection with the CPU that creates the connection. A value of 1 associates the connection with multiple squeues that belong to different CPUs.

Default 1

Range 0 or 1

Dynamic? Yes

When to Change	Consider setting this parameter to 1 to spread the load across all CPUs in certain situations. For example, when the number of CPUs exceed the number of NICs, and one CPU is not capable of handling the network load of a single NIC, change this parameter to 1. This property can only be set in the <code>/etc/system</code> file.
Zone Configuration	This parameter can only be set in the global zone.
Commitment Level	Unstable

IP Tunable Parameters With Additional Cautions

Changing the following parameters is not recommended.

`_pathmtu_interval`

Description	Specifies the interval in milliseconds when IP flushes the path maximum transfer unit (PMTU) discovery information, and tries to rediscover PMTU. Refer to RFC 1191 on PMTU discovery.
Default	1200 milliseconds (20 minutes)
Range	2-999999999
Dynamic?	Yes
When to Change	Do not change this value.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180 .

`_icmp_return_data_bytes (ipv4 or ipv6)`

Description	When IPv4 or IPv6 sends an ICMPv4 or ICMPv6 error message, it includes the IP header of the packet that caused the error message. This parameter controls how many extra bytes of the packet beyond the IPv4 or IPv6 header are included in the ICMPv4 or ICMPv6 error message.
Default	64 for IPv4 1280 for IPv6

Range	8-6,636 for IPv4 8-1,280 for IPv6
Dynamic?	Yes
When to Change	Do not change the value. Including more information in an ICMP error message might help in diagnosing network problems. If this feature is needed, increase the value.
Commitment Level	Unstable
Change History	For information, see “IP Parameter Name Changes (Oracle Solaris 11)” on page 180.

TCP Tunable Parameters

_deferred_ack_interval

Description	Specifies the time-out value for the TCP-delayed acknowledgment (ACK) timer for hosts that are not directly connected. Refer to RFC 1122, 4.2.3.2.
Default	100 milliseconds
Range	1 millisecond to 1 minute
Dynamic?	Yes
When to Change	Do not increase this value to more than 500 milliseconds. Increase the value under the following circumstances: <ul style="list-style-type: none">▪ Slow network links (less than 57.6 Kbps) with greater than 512 bytes maximum segment size (MSS)▪ The interval for receiving more than one TCP segment is short
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

`_local_dack_interval`

Description	Specifies the time-out value for TCP-delayed acknowledgment (ACK) timer for hosts that are directly connected. Refer to RFC 1122, 4.2.3.2.
Default	50 milliseconds
Range	10 milliseconds to 500 milliseconds
Dynamic?	Yes
When to Change	Do not increase this value to more than 500 milliseconds. Increase the value under the following circumstances: <ul style="list-style-type: none"> ▪ Slow network links (less than 57.6 Kbps) with greater than 512 bytes maximum segment size (MSS) ▪ The interval for receiving more than one TCP segment is short
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

`_deferred_acks_max`

Description	Specifies the maximum number of TCP segments received from remote destinations (not directly connected) before an acknowledgment (ACK) is generated. TCP segments are measured in units of maximum segment size (MSS) for individual connections. If set to 0 or 1, no ACKs are delayed, assuming all segments are 1 MSS long. The actual number is dynamically calculated for each connection. The value is the default maximum.
Default	2
Range	0 to 16
Dynamic?	Yes
When to Change	Do not change the value. In some circumstances, when the network traffic becomes very bursty because of the delayed ACK effect, decrease the value. Do not decrease this value below 2.
Commitment Level	Unstable

Change History For information, see [“TCP Parameter Name Changes \(Oracle Solaris 11\)” on page 181](#).

_local_dacks_max

Description	Specifies the maximum number of TCP segments received from directly connected destinations before an acknowledgment (ACK) is generated. TCP segments are measured in units of maximum segment size (MSS) for individual connections. If set to 0 or 1, it means no ACKs are delayed, assuming all segments are 1 MSS long. The actual number is dynamically calculated for each connection. The value is the default maximum.
Default	8
Range	0 to 16
Dynamic?	Yes
When to Change	Do not change the value. In some circumstances, when the network traffic becomes very bursty because of the delayed ACK effect, decrease the value. Do not decrease this value below 2.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

_wscale_always

Description	When this parameter is enabled, which is the default setting, TCP always sends a SYN segment with the window scale option, even if the window scale option value is 0. Note that if TCP receives a SYN segment with the window scale option, even if the parameter is disabled, TCP responds with a SYN segment with the window scale option. In addition, the option value is set according to the receive window size. Refer to RFC 1323 for the window scale option.
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes

When to Change	If there is an interoperability problem with an old TCP stack that does not support the window scale option, disable this parameter.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

`_tstamp_always`

Description	If set to 1, TCP always sends a SYN segment with the timestamp option. Note that if TCP receives a SYN segment with the timestamp option, TCP responds with a SYN segment with the timestamp option even if the parameter is set to 0.
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	If getting an accurate measurement of round-trip time (RTT) and TCP sequence number wraparound is a problem, enable this parameter. Refer to RFC 1323 for more reasons to enable this option.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

`send_buf`

Description	Defines the default send window size in bytes. Refer to “Per-Route Metrics” on page 168 for a discussion of setting a different value on a per-route basis. See also “max_buf” on page 140 .
Default	49,152
Range	4,096 to the current value of “max_buf” on page 140
Dynamic?	Yes
When to Change	An application can use setsockopt(3XNET) <code>SO_SNDBUF</code> to change the individual connection's send buffer.
Commitment Level	Unstable

Change History For information, see [“TCP Parameter Name Changes \(Oracle Solaris 11\)” on page 181](#).

recv_buf

Description Defines the default receive window size in bytes. Refer to [“Per-Route Metrics” on page 168](#) for a discussion of setting a different value on a per-route basis. See also [“max_buf” on page 140](#) and [“_recv_hiwat_minmss” on page 153](#).

Default 128,000

Range 2,048 to the current value of [“max_buf” on page 140](#)

Dynamic? Yes

When to Change An application can use [setsockopt\(3XNET\)](#) SO_RCVBUF to change the individual connection's receive buffer.

Commitment Level Unstable

Change History For information, see [“TCP Parameter Name Changes \(Oracle Solaris 11\)” on page 181](#).

max_buf

Description Defines the maximum send and receive buffer size in bytes. This parameter controls how large the send and receive buffers are set to by an application that uses [setsockopt\(3XNET\)](#).

Default 1,048,576

Range 128,000 to 1,073,741,824

Dynamic? Yes

When to Change If TCP connections are being made in a high-speed network environment, increase the value to match the network link speed.

Commitment Level Unstable

Change History For information, see [“TCP Parameter Name Changes \(Oracle Solaris 11\)” on page 181](#).

_cwnd_max

Description	Defines the maximum value of the TCP congestion window (cwnd) in bytes. For more information on the TCP congestion window, refer to RFC 1122 and RFC 2581.
Default	1,048,576
Range	128 to 1,073,741,824
Dynamic?	Yes
When to Change	Even if an application uses <code>setsockopt(3XNET)</code> to change the window size to a value higher than <code>_cwnd_max</code> , the actual window used can never grow beyond <code>_cwnd_max</code> . Thus, <code>_max_buf</code> should be greater than <code>_cwnd_max</code> .
Commitment Level	Unstable
Change History	For information, see “ TCP Parameter Name Changes (Oracle Solaris 11) ” on page 181.

_slow_start_initial

Description	Defines the maximum initial congestion window (cwnd) size in the maximum segment size (MSS) of a TCP connection. Refer to RFC 2414 on how the initial congestion window size is calculated.
Default	10
Range	1 to 10
Dynamic?	Yes
When to Change	Do not change the value. If the initial cwnd size causes network congestion under special circumstances, decrease the value.
Commitment Level	Unstable
Change History	For information, see “ TCP Parameter Name Changes (Oracle Solaris 11) ” on page 181.

`_local_slow_start_initial`

Description	Defines the initial congestion window (cwnd) size in the maximum segment size (MSS) of a TCP connection between directly connected hosts.
Default	10
Range	1 to 16,384
Dynamic?	Yes
When to Change	Consider increasing this parameter value if applications would benefit from a larger initial window.
Commitment Level	Unstable
Change History	For information, see “ _local_slow_start_initial ” on page 180.

`_slow_start_after_idle`

Description	The congestion window size in the maximum segment size (MSS) of a TCP connection after it has been idled (no segment received) for a period of one retransmission timeout (RTO). Refer to RFC 2414 on how the initial congestion window size is calculated.
Default	4
Range	1 to 16,384
Dynamic?	Yes
When to Change	For more information, see “ _slow_start_initial ” on page 141.
Commitment Level	Unstable
Change History	For information, see “ TCP Parameter Name Changes (Oracle Solaris 11) ” on page 181.

sack

Description	If set to 2, TCP always sends a SYN segment with the selective acknowledgment (SACK) permitted option. If TCP receives a SYN segment with a SACK-permitted option and this parameter is set to 1, TCP responds with a SACK-permitted option. If the parameter is set to
-------------	---

	0, TCP does not send a SACK-permitted option, regardless of whether the incoming segment contains the SACK permitted option.
	Refer to RFC 2018 for information on the SACK option.
Default	2 (active enabled)
Range	0 (disabled), 1 (passive enabled), or 2 (active enabled)
Dynamic?	Yes
When to Change	SACK processing can improve TCP retransmission performance so it should be actively enabled. Sometimes, the other side can be confused with the SACK option actively enabled. If this confusion occurs, set the value to 1 so that SACK processing is enabled only when incoming connections allow SACK processing.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

_rev_src_routes

Description	If set to 0, TCP does not reverse the IP source routing option for incoming connections for security reasons. If set to 1, TCP does the normal reverse source routing.
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	If IP source routing is needed for diagnostic purposes, enable it.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

_time_wait_interval

Description	Specifies the time in milliseconds that a TCP connection stays in TIME-WAIT state.
	For more information, refer to RFC 1122, 4.2.2.13.

Default	60,000 (60 seconds)
Range	1 second to 10 minutes
Dynamic?	Yes
When to Change	Do not set the value lower than 60 seconds. For information on changing this parameter, refer to RFC 1122, 4.2.2.13.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181 .

ecn

Description	<p>Controls Explicit Congestion Notification (ECN) support.</p> <p>If this parameter is set to 0, TCP does not negotiate with a peer that supports the ECN mechanism.</p> <p>If this parameter is set to 1 when initiating a connection, TCP does not tell a peer that it supports ECN mechanism.</p> <p>However, TCP tells a peer that it supports ECN mechanism when accepting a new incoming connection request if the peer indicates that it supports ECN mechanism in the SYN segment.</p> <p>If this parameter is set to 2, in addition to negotiating with a peer on the ECN mechanism when accepting connections, TCP indicates in the outgoing SYN segment that it supports the ECN mechanism when TCP makes active outgoing connections.</p> <p>Refer to RFC 3168 for information on ECN.</p>
Default	1 (passive enabled)
Range	0 (disabled), 1 (passive enabled), or 2 (active enabled)
Dynamic?	Yes
When to Change	ECN can help TCP better handle congestion control. However, there are existing TCP implementations, firewalls, NATs, and other network devices that are confused by this mechanism. These devices do not comply to the IETF standard.

Because of these devices, the default value of this parameter is set to 1. In rare cases, passive enabling can still cause problems. Set the parameter to 0 only if absolutely necessary.

Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

`_conn_req_max_q`

Description	Specifies the default maximum number of pending TCP connections for a TCP listener waiting to be accepted by <code>accept(3SOCKET)</code> . See also “<code>_conn_req_max_q0</code>” on page 146.
Default	128
Range	1 to 4,294,967,295
Dynamic?	Yes
When to Change	For applications such as web servers that might receive several connection requests, the default value might be increased to match the incoming rate.

Do not increase the parameter to a very large value. The pending TCP connections can consume excessive memory. Also, if an application cannot handle that many connection requests fast enough because the number of pending TCP connections is too large, new incoming requests might be denied.

Note that increasing `_conn_req_max_q` does not mean that applications can have that many pending TCP connections. Applications can use `listen(3SOCKET)` to change the maximum number of pending TCP connections for each socket. This parameter is the maximum an application can use `listen()` to set the number to. Thus, even if this parameter is set to a very large value, the actual maximum number for a socket might be much less than `_conn_req_max_q`, depending on the value used in `listen()`.

Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

`_conn_req_max_q0`

Description	<p>Specifies the default maximum number of incomplete (three-way handshake not yet finished) pending TCP connections for a TCP listener.</p> <p>For more information on TCP three-way handshake, refer to RFC 793. See also “_conn_req_max_q” on page 145.</p>
Default	1,024
Range	0 to 4,294,967,295
Dynamic?	Yes
When to Change	<p>For applications such as web servers that might receive excessive connection requests, you can increase the default value to match the incoming rate.</p> <p>The following explains the relationship between <code>_conn_req_max_q0</code> and the maximum number of pending connections for each socket.</p> <p>When a connection request is received, TCP first checks if the number of pending TCP connections (three-way handshake is done) waiting to be accepted exceeds the maximum (N) for the listener. If the connections are excessive, the request is denied. If the number of connections is allowable, then TCP checks if the number of incomplete pending TCP connections exceeds the sum of N and <code>_conn_req_max_q0</code>. If it does not, the request is accepted. Otherwise, the oldest incomplete pending TCP request is dropped.</p>
Commitment Level	Unstable
Change History	For information, see “ TCP Parameter Name Changes (Oracle Solaris 11) ” on page 181.

`_conn_req_min`

Description	<p>Specifies the default minimum value for the maximum number of pending TCP connection requests for a listener waiting to be accepted. This is the lowest maximum value of <code>listen(3SOCKET)</code> that an application can use.</p>
Default	1
Range	1 to 1,024
Dynamic?	Yes

When to Change	This parameter can be a solution for applications that use listen(3SOCKET) to set the maximum number of pending TCP connections to a value too low. Increase the value to match the incoming connection request rate.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

`_rst_sent_rate_enabled`

Description	If this parameter is set to 1, the maximum rate of sending a RST segment is controlled by the <code>ipadm</code> parameter, <code>_rst_sent_rate</code> . If this parameter is set to 0, no rate control when sending a RST segment is available.
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	This tunable helps defend against denial of service attacks on TCP by limiting the rate by which a RST segment is sent out. The only time this rate control should be disabled is when strict conformance to RFC 793 is required.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

`_rst_sent_rate`

Description	Sets the maximum number of RST segments that TCP can send out per second.
Default	40
Range	0 to 4,294,967,295
Dynamic?	Yes
When to Change	In a TCP environment, there might be a legitimate reason to generate more RSTs than the default value allows. In this case, increase the default value of this parameter.

Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

smallest_anon_port

Description	This parameter controls the smallest port number TCP can select as an ephemeral port. An application can use an ephemeral port when it creates a connection with a specified protocol and it does not specify a port number. Ephemeral ports are not associated with a specific application. When the connection is closed, the port number can be reused by a different application.
Unit	Port number
Default	32,768
Range	1,024 to 65,535
Dynamic?	Yes
When to Change	When a larger ephemeral port range is required.
Commitment Level	Unstable
Change History	For information, see “[tcp, sctp, udp]_smallest_anon_port and [tcp, sctp, udp]_largest_anon_port” on page 180.

largest_anon_port

Description	This parameter controls the largest port number TCP can select as an ephemeral port. An application can use an ephemeral port when it creates a connection with a specified protocol and it does not specify a port number. Ephemeral ports are not associated with a specific application. When the connection is closed, the port number can be reused by a different application.
Unit	Port number
Default	65,535
Range	32,768 to 65,535
Dynamic?	Yes
When to Change	When a larger ephemeral port range is required.
Commitment Level	Unstable

Change History For information, see “[tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port” on page 180.

TCP/IP Parameters Set in the /etc/system File

The following parameters can be set only in the /etc/system file. After the file is modified, reboot the system.

For example, the following entry sets the `ipcl_conn_hash_size` parameter:

```
set ip:ipcl_conn_hash_size=value
```

ipcl_conn_hash_size

Description	Controls the size of the connection hash table used by IP. The default value of 0 means that the system automatically sizes an appropriate value for this parameter at boot time, depending on the available memory.
Data Type	Unsigned integer
Default	0
Range	0 to 82,500
Dynamic?	No. The parameter can only be changed at boot time.
When to Change	If the system consistently has tens of thousands of TCP connections, the value can be increased accordingly. Increasing the hash table size means that more memory is wired down, thereby reducing available memory to user applications.
Commitment Level	Unstable

ip_queue_worker_wait

Description	Governs the maximum delay in waking up a worker thread to process TCP/IP packets that are enqueued on a queue. An <i>queue</i> is a serialization queue that is used by the TCP/IP kernel code to process TCP/IP packets.
Default	10 milliseconds
Range	0 – 50 milliseconds
Dynamic?	Yes

When to Change	Consider tuning this parameter if latency is an issue, and network traffic is light. For example, if the machine serves mostly interactive network traffic. The default value usually works best on a network file server, a web server, or any server that has substantial network traffic.
Zone Configuration	This parameter can only be set in the global zone.
Commitment Level	Unstable

TCP Parameters With Additional Cautions

Changing the following parameters is not recommended.

keepalive_interval

Description	<p>This <code>ipadm</code> parameter sets a probe interval that is first sent out after a TCP connection is idle on a system-wide basis.</p> <p>Solaris supports the TCP keep-alive mechanism as described in RFC 1122. This mechanism is enabled by setting the <code>SO_KEEPALIVE</code> socket option on a TCP socket.</p> <p>If <code>SO_KEEPALIVE</code> is enabled for a socket, the first keep-alive probe is sent out after a TCP connection is idle for two hours, the default value of the <code>tcp_keepalive_interval</code> parameter. If the peer does not respond to the probe after eight minutes, the TCP connection is aborted. For more information, refer to “_rexmit_interval_initial” on page 151.</p> <p>You can also use the <code>TCP_KEEPALIVE_THRESHOLD</code> socket option on individual applications to override the default interval so that each application can have its own interval on each socket. The option value is an unsigned integer in milliseconds. See also tcp(7P).</p>
Default	2 hours
Range	10 seconds to 10 days
Units	Unsigned integer (milliseconds)
Dynamic?	Yes
When to Change	Do not change the value. Lowering it may cause unnecessary network traffic and might also increase the chance of premature termination of the connection because of a transient network problem.

Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

_ip_abort_interval

Description	Specifies the default total retransmission timeout value for a TCP connection. For a given TCP connection, if TCP has been retransmitting for <code>_ip_abort_interval</code> period of time and it has not received any acknowledgment from the other endpoint during this period, TCP closes this connection. For TCP retransmission timeout (RTO) calculation, refer to RFC 1122, 4.2.3. See also “_rexmit_interval_max” on page 152.
Default	5 minutes
Range	500 milliseconds to 1193 hours
Dynamic?	Yes
When to Change	Do not change this value. See “_rexmit_interval_max” on page 152 for exceptions.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

_rexmit_interval_initial

Description	Specifies the default initial retransmission timeout (RTO) value for a TCP connection. Refer to “Per-Route Metrics” on page 168 for a discussion of setting a different value on a per-route basis.
Default	1000 milliseconds
Range	1 millisecond to 20000 milliseconds
Dynamic?	Yes
When to Change	Do not change this value. Lowering the value can result in unnecessary retransmissions.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

`_rexmit_interval_max`

Description	Defines the default maximum retransmission timeout value (RTO). The calculated RTO for all TCP connections cannot exceed this value. See also “ _ip_abort_interval ” on page 151.
Default	6000 milliseconds
Range	1 millisecond to 7200000 milliseconds
Dynamic?	Yes
When to Change	Do not change the value in a normal network environment. If, in some special circumstances, the round-trip time (RTT) for a connection is about 10 seconds, you can increase this value. If you change this value, you should also change the <code>_ip_abort_interval</code> parameter. Change the value of <code>_ip_abort_interval</code> to at least four times the value of <code>_rexmit_interval_max</code> .
Commitment Level	Unstable
Change History	For information, see “ TCP Parameter Name Changes (Oracle Solaris 11) ” on page 181.

`_rexmit_interval_min`

Description	Specifies the default minimum retransmission time out (RTO) value. The calculated RTO for all TCP connections cannot be lower than this value. See also “ _rexmit_interval_max ” on page 152.
Default	200 milliseconds
Range	1 millisecond to 7200000 milliseconds
Dynamic?	Yes
When to Change	Do not change the value in a normal network environment. TCP's RTO calculation should cope with most RTT fluctuations. If, in some very special circumstances, the round-trip time (RTT) for a connection is about 10 seconds, increase this value. If you change this value, you should change the <code>_rexmit_interval_max</code> parameter. Change the value of <code>_rexmit_interval_max</code> to at least eight times the value of <code>_rexmit_interval_min</code> .
Commitment Level	Unstable
Change History	For information, see “ TCP Parameter Name Changes (Oracle Solaris 11) ” on page 181.

_rexmit_interval_extra

Description	Specifies a constant added to the calculated retransmission time out value (RTO).
Default	0 milliseconds
Range	0 to 7200000 milliseconds
Dynamic?	Yes
When to Change	Do not change the value. When the RTO calculation fails to obtain a good value for a connection, you can change this value to avoid unnecessary retransmissions.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

_tstamp_if_wscale

Description	If this parameter is set to 1, and the window scale option is enabled for a connection, TCP also enables the <code>timestamp</code> option for that connection.
Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	Do not change this value. In general, when TCP is used in high-speed network, protection against sequence number wraparound is essential. Thus, you need the <code>timestamp</code> option.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

_recv_hiwat_minmss

Description	Controls the default minimum receive window size. The minimum is <code>_recv_hiwat_minmss</code> times the size of maximum segment size (MSS) of a connection.
Default	8
Range	1 to 65,536

Dynamic?	Yes
When to Change	Do not change the value. If changing it is necessary, do not change the value lower than 4.
Commitment Level	Unstable
Change History	For information, see “TCP Parameter Name Changes (Oracle Solaris 11)” on page 181.

UDP Tunable Parameters

send_buf

Description	Defines the default send buffer size for a UDP socket. For more information, see “max_buf” on page 155.
Default	57,344 bytes
Range	1,024 to the current value of “max_buf” on page 155
Dynamic?	Yes
When to Change	Note that an application can use <code>setsockopt(3XNET)</code> <code>SO_SNDBUF</code> to change the size for an individual socket. In general, you do not need to change the default value.
Commitment Level	Unstable
Change History	For information, see “UDP Parameter Name Changes (Oracle Solaris 11)” on page 183.

recv_buf

Description	Defines the default receive buffer size for a UDP socket. For more information, see “max_buf” on page 155.
Default	57,344 bytes
Range	128 to the current value of “max_buf” on page 155
Dynamic?	Yes
When to Change	Note that an application can use <code>setsockopt(3XNET)</code> <code>SO_RCVBUF</code> to change the size for an individual socket. In general, you do not need to change the default value.

Commitment Level	Unstable
Change History	For information, see “UDP Parameter Name Changes (Oracle Solaris 11)” on page 183.

max_buf

Description	Defines the maximum send and receive buffer size for a UDP socket. It controls how large the send and receive buffers are set to by an application that uses <code>getsockopt(3SOCKET)</code> .
Default	2,097,152
Range	65,536 to 1,073,741,824
Dynamic?	Yes
When to Change	Increase the value of this parameter to match the network link speed if associations are being made in a high-speed network environment.
Commitment Level	Unstable
Change History	For information, see “UDP Parameter Name Changes (Oracle Solaris 11)” on page 183.

smallest_anon_port

Description	This parameter controls the smallest port number UDP can select as an ephemeral port. An application can use an ephemeral port when it creates a connection with a specified protocol and it does not specify a port number. Ephemeral ports are not associated with a specific application. When the connection is closed, the port number can be reused by a different application.
Unit	Port number
Default	32,768
Range	1,024 to 65,535
Dynamic?	Yes
When to Change	When a larger ephemeral port range is required.
Commitment Level	Unstable
Change History	For information, see “[tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port” on page 180.

largest_anon_port

Description	This parameter controls the largest port number UDP can select as an ephemeral port. An application can use an ephemeral port when it creates a connection with a specified protocol and it does not specify a port number. Ephemeral ports are not associated with a specific application. When the connection is closed, the port number can be reused by a different application.
Unit	Port number
Default	65,535
Range	32,768 to 65,535
Dynamic?	Yes
When to Change	When a larger ephemeral port range is required.
Commitment Level	Unstable
Change History	For information, see “[tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port” on page 180.

IPQoS Tunable Parameter

_policy_mask

Description Enables or disables IPQoS processing in any of the following callout positions: forward outbound, forward inbound, local outbound, and local inbound. This parameter is a bitmask as follows:

Not Used	Not Used	Not Used	Not Used	Forward Outbound	Forward Inbound	Local Outbound	Local Inbound
X	X	X	X	0	0	0	0

A 1 in any of the position masks or disables IPQoS processing in that particular callout position. For example, a value of 0x01 disables IPQoS processing for all the local inbound packets.

Default The default value is 0, meaning that IPQoS processing is enabled in all the callout positions.

Range	0 (0x00) to 15 (0x0F). A value of 15 indicates that IPQoS processing is disabled in all the callout positions.
Dynamic?	Yes
When to Change	If you want to enable or disable IPQoS processing in any of the callout positions.
Commitment Level	Unstable

SCTP Tunable Parameters

`_max_init_retr`

Description	Controls the maximum number of attempts an SCTP endpoint should make at resending an INIT chunk. The SCTP endpoint can use the SCTP initiation structure to override this value.
Default	8
Range	0 to 128
Dynamic?	Yes
When to Change	The number of INIT retransmissions depend on “ _pa_max_retr ” on page 157 . Ideally, <code>_max_init_retr</code> should be less than or equal to <code>_pa_max_retr</code> .
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183 .

`_pa_max_retr`

Description	Controls the maximum number of retransmissions (over all paths) for an SCTP association. The SCTP association is aborted when this number is exceeded.
Default	10
Range	1 to 128
Dynamic?	Yes
When to Change	The maximum number of retransmissions over all paths depend on the number of paths and the maximum number of retransmission over

each path. Ideally, `sctp_pa_max_retr` should be set to the sum of “`_pp_max_retr`” on page 158 over all available paths. For example, if there are 3 paths to the destination and the maximum number of retransmissions over each of the 3 paths is 5, then `_pa_max_retr` should be set to less than or equal to 15. (See the Note in Section 8.2, RFC 2960.)

Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

`_pp_max_retr`

Description	Controls the maximum number of retransmissions over a specific path. When this number is exceeded for a path, the path (destination) is considered unreachable.
Default	5
Range	1 to 128
Dynamic?	Yes
When to Change	Do not change this value to less than 5.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

`_cwnd_max`

Description	Controls the maximum value of the congestion window for an SCTP association.
Default	1,048,576
Range	128 to 1,073,741,824
Dynamic?	Yes
When to Change	Even if an application uses <code>setsockopt(3XNET)</code> to change the window size to a value higher than <code>_cwnd_max</code> , the actual window used can never grow beyond <code>_cwnd_max</code> . Thus, “ <code>max_buf</code> ” on page 163 should be greater than <code>_cwnd_max</code> .
Commitment Level	Unstable

Change History For information, see [“SCTP Parameter Name Changes \(Oracle Solaris 11\)” on page 183.](#)

`_ipv4_ttl`

Description Controls the time to live (TTL) value in the IP version 4 header for the outbound IPv4 packets on an SCTP association.

Default 64

Range 1 to 255

Dynamic? Yes

When to Change Generally, you do not need to change this value.

Commitment Level Unstable

Change History For information, see [“SCTP Parameter Name Changes \(Oracle Solaris 11\)” on page 183.](#)

`_ipv6_hoplimit`

Description Sets the value of the hop limit in the IPv6 header for the outbound IPv6 packets on an SCTP association.

Default 60

Range 0 to 255

Dynamic? Yes

When to Change Generally, you do not need to change this value.

Commitment Level Unstable

Change History For information, see [“SCTP Parameter Name Changes \(Oracle Solaris 11\)” on page 183.](#)

`_heartbeat_interval`

Description Computes the interval between HEARTBEAT chunks to an idle destination, that is allowed to heartbeat.

An SCTP endpoint periodically sends an HEARTBEAT chunk to monitor the reachability of the idle destinations transport addresses of its peer.

Default 30 seconds

Range	0 to 86,400 seconds
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 8.3.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

_new_secret_interval

Description	Determines when a new secret needs to be generated. The generated secret is used to compute the MAC for a cookie.
Default	2 minutes
Range	0 to 1,440 minutes
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 5.1.3.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

_initial_mtu

Description	Determines the initial maximum send size for an SCTP packet including the length of the IP header.
Default	1500 bytes
Range	68 to 65,535
Dynamic?	Yes
When to Change	Increase this parameter if the underlying link supports frame sizes that are greater than 1500 bytes.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

`_deferred_ack_interval`

Description	Sets the time-out value for SCTP delayed acknowledgment (ACK) timer in milliseconds.
Default	100 milliseconds
Range	1 to 60,000 milliseconds
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 6.2.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

`_ignore_path_mtu`

Description	Enables or disables path MTU discovery.
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	Enable this parameter if you want to ignore MTU changes along the path. However, doing so might result in IP fragmentation if the path MTU decreases.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

`_initial_ssthresh`

Description	Sets the initial slow start threshold for a destination address of the peer.
Default	1,048,576
Range	1,024 to 4,294,967,295
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 7.2.1.
Commitment Level	Unstable

Change History For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

send_buf

Description Defines the default send buffer size in bytes. See also “[max_buf](#)” on page 163.

Default 102,400

Range 8,192 to the current value of “[max_buf](#)” on page 163

Dynamic? Yes

When to Change An application can use `setsockopt(3XNET)` `SO_SNDBUF` to change the individual connection's send buffer.

Commitment Level Unstable

Change History For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

_xmit_lowat

Description Controls the lower limit on the send window size.

Default 8,192

Range 8,192 to 1,073,741,824

Dynamic? Yes

When to Change Generally, you do not need to change this value. This parameter sets the minimum size required in the send buffer for the socket to be marked writable. If required, consider changing this parameter in accordance with “[send_buf](#)” on page 162.

Commitment Level Unstable

Change History For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

recv_buf

Description Defines the default receive buffer size in bytes. See also “[max_buf](#)” on page 163.

Default	102,400
Range	8,192 to the current value of “ max_buf ” on page 163
Dynamic?	Yes
When to Change	An application can use setsockopt(3XNET) <code>SO_RCVBUF</code> to change the individual connection's receive buffer.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

max_buf

Description	Controls the maximum send and receive buffer size in bytes. It controls how large the send and receive buffers are set to by an application that uses getsockopt(3SOCKET) .
Default	1,048,576
Range	102,400 to 1,073,741,824
Dynamic?	Yes
When to Change	Increase the value of this parameter to match the network link speed if associations are being made in a high-speed network environment.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

_rto_min

Description	Sets the lower bound for the retransmission timeout (RTO) in milliseconds for all the destination addresses of the peer.
Default	1,000
Range	500 to 60,000
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 6.3.1.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

`_rto_max`

Description	Controls the upper bound for the retransmission timeout (RTO) in milliseconds for all the destination addresses of the peer.
Default	60,000
Range	1,000 to 60,000,000
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 6.3.1.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183 .

`_rto_initial`

Description	Controls the initial retransmission timeout (RTO) in milliseconds for all the destination addresses of the peer.
Default	3,000
Range	1,000 to 60,000,000
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 6.3.1.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183 .

`_cookie_life`

Description	Sets the lifespan of a cookie in milliseconds.
Default	60,000
Range	10 to 60,000,000
Dynamic?	Yes
When to Change	Generally, you do not need to change this value. This parameter might be changed in accordance with “<code>_rto_max</code>” on page 164 .
Commitment Level	Unstable

Change History For information, see [“SCTP Parameter Name Changes \(Oracle Solaris 11\)” on page 183.](#)

_max_in_streams

Description	Controls the maximum number of inbound streams permitted for an SCTP association.
Default	32
Range	1 to 65,535
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 5.1.1.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

_initial_out_streams

Description	Controls the maximum number of outbound streams permitted for an SCTP association.
Default	32
Range	1 to 65,535
Dynamic?	Yes
When to Change	Refer to RFC 2960, section 5.1.1.
Commitment Level	Unstable
Change History	For information, see “SCTP Parameter Name Changes (Oracle Solaris 11)” on page 183.

_shutack_wait_bound

Description	Controls the maximum time, in milliseconds, to wait for a SHUTDOWN ACK after having sent a SHUTDOWN chunk.
Default	60,000
Range	0 to 300,000
Dynamic?	Yes

When to Change	Generally, you do not need to change this value. This parameter might be changed in accordance with “ _rto_max ” on page 164.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

`_maxburst`

Description	Sets the limit on the number of segments to be sent in a burst.
Default	4
Range	2 to 8
Dynamic?	Yes
When to Change	You do not need to change this parameter. You might change it for testing purposes.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

`_addip_enabled`

Description	Enables or disables SCTP dynamic address reconfiguration.
Default	0 (disabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	The parameter can be enabled if dynamic address reconfiguration is needed. Due to security implications, enable this parameter only for testing purposes.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

`_prsctp_enabled`

Description	Enables or disables the partial reliability extension (RFC 3758) to SCTP.
-------------	---

Default	1 (enabled)
Range	0 (disabled) or 1 (enabled)
Dynamic?	Yes
When to Change	Disable this parameter if partial reliability is not supported in your SCTP environment.
Commitment Level	Unstable
Change History	For information, see “ SCTP Parameter Name Changes (Oracle Solaris 11) ” on page 183.

smallest_anon_port

Description	This parameter controls the smallest port number SCTP can select as an ephemeral port. An application can use an ephemeral port when it creates a connection with a specified protocol and it does not specify a port number. Ephemeral ports are not associated with a specific application. When the connection is closed, the port number can be reused by a different application.
Unit	Port number
Default	32,768
Range	1,024 to 65,535
Dynamic?	Yes
When to Change	When a larger ephemeral port range is required.
Commitment Level	Unstable
Change History	For information, see “ [tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port ” on page 180.

largest_anon_port

Description	This parameter controls the largest port number SCTP can select as an ephemeral port. An application can use an ephemeral port when it creates a connection with a specified protocol and it does not specify a port number. Ephemeral ports are not associated with a specific application. When the connection is closed, the port number can be reused by a different application.
Unit	Port number

Default	65,535
Range	32,768 to 65,535
Dynamic?	Yes
When to Change	When a larger ephemeral port range is required.
Commitment Level	Unstable
Change History	For information, see “[tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port” on page 180.

Per-Route Metrics

You can use per-route metrics to associate some properties with IPv4 and IPv6 routing table entries.

For example, a system has two different network interfaces, a fast Ethernet interface and a gigabit Ethernet interface. The system default `recv_maxbuf` is 128,000 bytes. This default is sufficient for the fast Ethernet interface, but may not be sufficient for the gigabit Ethernet interface.

Instead of increasing the system's default for `recv_maxbuf`, you can associate a different default TCP receive window size to the gigabit Ethernet interface routing entry. By making this association, all TCP connections going through the route will have the increased receive window size.

For example, the following is in the routing table (`netstat -rn`), assuming IPv4:

```
192.123.123.0      192.123.123.4      U      1      4 hme0
192.123.124.0      192.123.124.4      U      1      4 ge0
default          192.123.123.1      UG     1      8
```

In this example, do the following:

```
# route change -net 192.123.124.0 -recvpipe x
```

Then, all connections going to the `192.123.124.0` network, which is on the `ge0` link, use the receive buffer size `x`, instead of the default `128,000` receive window size.

If the destination is in the `a.b.c.d` network, and no specific routing entry exists for that network, you can add a prefix route to that network and change the metric. For example:

```
# route add -net a.b.c.d 192.123.123.1 -netmask w.x.y.z
# route change -net a.b.c.d -recvpipe y
```

Note that the prefix route's gateway is the default router. Then, all connections going to that network use the receive buffer size y . If you have more than one interface, use the `-i fp` argument to specify which interface to use. This way, you can control which interface to use for specific destinations. To verify the metric, use the `route(1M)` get command.

System Facility Parameters

This chapter describes most of the parameters default values for various system facilities.

- “autofs” on page 172
- “cron” on page 172
- “devfsadm” on page 172
- “dhcpageant” on page 172
- “fs” on page 173
- “ftp” on page 173
- “inetinit” on page 173
- “init” on page 173
- “ipsec” on page 174
- “kbd” on page 174
- “keyserv” on page 174
- “login” on page 175
- “mpathd” on page 175
- “nfs” on page 175
- “nfslogd” on page 175
- “nss” on page 175
- “passwd” on page 175
- “su” on page 176
- “syslog” on page 176
- “tar” on page 176
- “telnetd” on page 176
- “utmpd” on page 176

System Default Parameters

The functioning of various system facilities is governed by a set of values that are read by each facility on startup. The values for each facility might be stored in a file for the facility located in the `/etc/default` directory, or in properties of a service instance in the Service Management Facility (SMF) configuration repository. For more information on SMF services and properties, see “Managing SMF Services” in *Managing Services and Faults in Oracle Solaris 11.1*.

For information about setting power management properties, see Chapter 5, “Managing the System Console, Terminal Devices, and Power Services (Tasks),” in *Managing System Information, Processes, and Performance in Oracle Solaris 11.1*.

autofs

You can display or configure SMF autofs properties by using the `sharectl` command. For example:

```
# sharectl get autofs
timeout=600
automount_verbose=false
automountd_verbose=false
nobrowse=false
trace=0
environment=
# sharectl set -p timeout=200 autofs
```

For details, see [sharectl\(1M\)](#).

cron

This facility enables you to disable or enable cron logging.

devfsadm

This file is not currently used.

dhcpgent

Client usage of DHCP is provided by the `dhcpgent` daemon. When `ipadm` is used to create a DHCP address object, or when `ipadm` identifies an interface that has been configured to receive its network configuration from DHCP, `dhcpgent` is started to manage an address on that interface.

For more information, see the `/etc/default/dhccpagent` information in the FILES section of `dhccpagent(1M)`.

fs

File system administrative commands have a generic and file system-specific portion. If the file system type is not explicitly specified with the `-F` option, a default is applied. The value is specified in this file. For more information, see the Description section of `default_fs(4)`.

ftp

This facility enables you to set the `ls` command behavior to the RFC 959 NLST command. The default `ls` behavior is the same as in the previous Solaris release.

For details, see `ftp(4)`.

inetinit

This facility enables you to configure TCP sequence numbers and to enable or disable support for 6to4 relay routers.

init

System initialization properties are now part of the following SMF service:

```
svc:/system/environment:init
```

You can display and configure system initialization properties, such as TZ and LANG, by using similar syntax:

```
# svccfg -s svc:/system/environment:init
svc:/system/environment:init> setprop
Usage: setprop pg/name = [type:] value
       setprop pg/name = [type:] ([value...])
```

Set the `pg/name` property of the currently selected entity. Values may be enclosed in double-quotes. Value lists may span multiple lines.

```
svc:/system/environment:init> listprop
umask                application
umask/value_authorization  astring                solaris.smf.value.environment
umask/umask          astring                022
upgrade             application
upgrade/skip_init_upgrade  boolean                false
```

upgrade/value_authorization	astring	solaris.smf.value.environment
environment	application	
environment/LANG	astring	C
.		
.		
.		

For more information, see the FILES section of [init\(1M\)](#).

ipsec

This facility enables you to configure parameters, such as IKE daemon debugging information and the `ikeadm` privilege level.

kbd

Keyboard configuration properties are now part of the following SMF service:

```
svc:/system/keymap:default
```

You display and configure keyboard properties by using similar syntax:

```
# svccfg -s svc:/system/keymap:default
svc:/system/keymap:default> setprop
Usage:  setprop pg/name = [type:] value
        setprop pg/name = [type:] ([value...])
```

Set the `pg/name` property of the currently selected entity. Values may be enclosed in double-quotes. Value lists may span multiple lines.

```
svc:/system/keymap:default> listprop
general                               framework
general/complete                      astring
general/enabled                       boolean    false
keymap                                 system
keymap/console_beeper_freq            integer    900
keymap/kbd_beeper_freq                integer    2000
keymap/keyboard_abort                 astring    enable
keymap/keyclick                       boolean    false
.
```

For more information, see [kbd\(1\)](#).

keyserv

For details, see the `/etc/default/keyserv` information in the FILES section of [keyserv\(1M\)](#).

login

For details, see the `/etc/default/login` information in the FILES section of [login\(1\)](#).

mpathd

This facility enables you to set in `.mpathd` configuration parameters.

For details, see [in.mpathd\(1M\)](#).

nfs

You can display or configure SMF NFS properties by using the `sharectl` command. For example:

```
# sharectl get nfs
servers=1024
lockd_listen_backlog=32
lockd_servers=1024
lockd_retransmit_timeout=5
grace_period=90
server_versmin=2
server_versmax=4
client_versmin=2
client_versmax=4
server_delegation=on
nfsmapid_domain=
# sharectl set -p grace_period=60 nfs
```

For details, see [nfs\(4\)](#).

nfslogd

For details, see the Description section of [nfslogd\(1M\)](#).

nss

This facility enables you to configure `initgroups(3C)` lookup parameters.

For details, see [nss\(4\)](#).

passwd

For details, see the `/etc/default/passwd` information in the FILES section of [passwd\(1\)](#).

su

For details, see the `/etc/default/su` information in the FILES section of [su\(1M\)](#).

syslog

For details, see the `/etc/default/syslogd` information in the FILES section of [syslogd\(1M\)](#).

tar

For a description of the `-f` function modifier, see [tar\(1\)](#).

If the TAPE environment variable is not present and the value of one of the arguments is a number and `-f` is not specified, the number matching the `archiveN` string is looked up in the `/etc/default/tar` file. The value of the `archiveN` string is used as the output device with the blocking and size specifications from the file.

For example:

```
% tar -c 2 /tmp/*
```

This command writes the output to the device specified as `archive2` in the `/etc/default/tar` file.

telnetd

This file identifies the default BANNER that is displayed upon a telnet connection.

utmpd

The `utmpd` daemon monitors `/var/adm/utmpx` (and `/var/adm/utmp` in earlier Solaris versions) to ensure that `utmp` entries inserted by non-root processes by `pututxline(3C)` are cleaned up on process termination.

Two entries in `/etc/default/utmpd` are supported:

- `SCAN_PERIOD` – The number of seconds that `utmpd` sleeps between checks of `/proc` to see if monitored processes are still alive. The default is 300.
- `MAX_FDS` – The maximum number of processes that `utmpd` attempts to monitor. The default value is 4096 and should never need to be changed.

Tunable Parameters Change History

This chapter describes the change history of specific tunable parameters. If a parameter is in this section, it has changed from a previous release. Parameters whose functionality has been removed are listed also.

- “Kernel Parameters” on page 177
- “TCP/IP Tunable Parameters (Oracle Solaris 11 and 11.1)” on page 180
- “Parameters That Are Obsolete or Have Been Removed (Oracle Solaris 11 and 11.1)” on page 184

Kernel Parameters

General Kernel and Memory Parameters (Oracle Solaris 11 and 11.1)

zfs_arc_min

This parameter description is newly documented in the Oracle Solaris 11 release. For more information, see “[zfs_arc_min](#)” on page 82.

In Oracle Solaris 11.1, the `zfs_arc_min` parameter information has moved to [Chapter 3](#), “Oracle Solaris ZFS Tunable Parameters.”

zfs_arc_max

This parameter description is newly documented in the Oracle Solaris 11 release. For more information, see “[zfs_arc_max](#)” on page 83.

In Oracle Solaris 11.1, the `zfs_arc_max` parameter information has moved to [Chapter 3](#), “Oracle Solaris ZFS Tunable Parameters.”

disp_rechoose_interval

This parameter is new in the Oracle Solaris 11 release. For more information, see [“disp_rechoose_interval” on page 73](#).

Paging-Related Parameters (Oracle Solaris 11)

fastscan

The default value of `fastscan` was clarified. For more information, see [“fastscan” on page 46](#).

Process-Sizing Tunables (Oracle Solaris 11.1)

ngroups_max

This parameter is newly documented in the Oracle Solaris 11 release.

In the Oracle Solaris 11.1 release, the When to Change description was revised to include interaction information for `AUTH_SYS` authentication when this parameter is set to greater than 16. For more information, see [“ngroups_max” on page 38](#).

max_nprocs

The default value of this parameter was updated in the Oracle Solaris 11.1 release. For more information, see [“max_nprocs” on page 37](#).

maxusers

The default value of this parameter was updated in the Oracle Solaris 11.1 release. For more information, see [“maxusers” on page 34](#).

pidmax

The default value of this parameter was updated in the Oracle Solaris 11.1 release. For more information, see [“pidmax” on page 36](#).

segkpsize

The default value of this parameter was updated in the Oracle Solaris 11.1 release. For more information, see [“segkpsize” on page 29](#).

Swapping-Related Parameter (Oracle Solaris 11.1)

swapfs_minfree

The When to Change section of this parameter information was updated in the Oracle Solaris 11.1 release. For more information, see [“swapfs_minfree” on page 51](#).

General File System Parameter (Oracle Solaris 11.1)

dnlc_dircache_percent

This parameter is new in the Oracle Solaris 11.1 release. For more information, see [“dnlc_dircache_percent” on page 65](#).

General Driver Parameter (Oracle Solaris 11)

ddi_msix_alloc_limit

This parameter is newly documented in the Oracle Solaris 11 release. For more information, see [“ddi_msix_alloc_limit” on page 55](#).

Network Driver Parameters (Oracle Solaris 11)

igb Parameters

The igb network driver parameters are provided in the Oracle Solaris 11 release. For more information, see [“igb Parameters” on page 56](#).

ixgbe Parameters

The ixgbe network driver parameters are provided in the Oracle Solaris 11 release. For more information, see [“ixgbe Parameters” on page 57](#).

TCP/IP Tunable Parameters (Oracle Solaris 11 and 11.1)

[tcp,sctp,udp]_smallest_anon_port and [tcp,sctp,udp]_largest_anon_port

These parameters are newly documented in the Oracle Solaris 11 release.

- “smallest_anon_port” on page 167
- “largest_anon_port” on page 167
- “smallest_anon_port” on page 148
- “largest_anon_port” on page 148
- “smallest_anon_port” on page 155
- “largest_anon_port” on page 156

_local_slow_start_initial

This parameter is newly documented in the Oracle Solaris 11.1 release. For more information, see “_local_slow_start_initial” on page 142.

IP Parameter Name Changes (Oracle Solaris 11)

The following IP parameters have been renamed to IP properties.

You can set an IP property by using syntax similar to the following:

```
# ipadm set-prop -p _icmp_err_interval=100 ip
```

You can display IP property information by using syntax similar to the following:

```
# ipadm show-prop -p _icmp_err_interval ip
PROTO PROPERTY          PERM CURRENT    PERSISTENT  DEFAULT  POSSIBLE
ip   _icmp_err_interval  rw   100          100         100      0-99999
```

TABLE A-1 IP Parameter Name Changes

Previous IP Parameter Name	IP Property Name
ip_addrs_per_if	_addrs_per_if
ip_forwarding	forwarding (IPv4)
ip6_forwarding	forwarding (IPv6)
ip_forward_src_routed	_forward_src_routed (IPv4)

TABLE A-1 IP Parameter Name Changes (Continued)

Previous IP Parameter Name	IP Property Name
ip6_forward_src_routed	_forward_src_routed (IPv6)
ip_icmp_err_interval	_icmp_err_interval
ip_icmp_err_burst	_icmp_err_burst
ip_icmp_return_data_bytes	_icmp_return_data_bytes (IPv4)
ip6_icmp_return_data_bytes	_icmp_return_data_bytes (IPv6)
ip_ire_pathmtu_interval	_pathmtu_interval
ip_respond_to_echo_broadcast	_respond_to_echo_broadcast (IPv4)
ip6_respond_to_echo_broadcast	_respond_to_echo_broadcast (IPv6)
ip_respond_to_echo_multicast	_respond_to_echo_multicast (IPv4)
ip6_respond_to_echo_multicast	_respond_to_echo_multicast (IPv6)
ip_send_redirects	_send_redirects (IPv4)
ip6_send_redirects	_send_redirects (IPv6)
ip_strict_dst_multihoming	hostmodel

TCP Parameter Name Changes (Oracle Solaris 11)

The following TCP parameters have been renamed to TCP properties.

You can set a TCP property by using syntax similar to the following:

```
# ipadm set-prop -p _deferred_ack_interval=100 tcp
```

You can display TCP property information by using syntax similar to the following:

```
# ipadm show-prop -p _deferred_ack_interval tcp
PROTO PROPERTY          PERM CURRENT  PERSISTENT  DEFAULT  POSSIBLE
tcp   _deferred_ack_interval rw  100         --          100      1-60000
```

TABLE A-2 TCP Parameter Name Changes

Previous TCP Parameter Name	TCP Property Name
tcp_deferred_ack_interval	_deferred_ack_interval
tcp_local_dack_interval	_local_dack_interval
tcp_deferred_acks_max	_deferred_acks_max
tcp_local_dacks_max	_local_dacks_max

TABLE A-2 TCP Parameter Name Changes (Continued)

Previous TCP Parameter Name	TCP Property Name
tcp_wscale_always	_wscale_always
tcp_tstamp_always	_tstamp_always
tcp_xmit_hiwat	send_buf
tcp_rcv_hiwat	recv_buf
tcp_max_buf	max_buf
tcp_cwnd_max	_cwnd_max
tcp_slow_start_initial	_slow_start_initial
tcp_slow_start_after_idle	_slow_start_after_idle
tcp_sack_permitted	sack
tcp_rev_src_routes	_rev_src_routes
tcp_time_wait_interval	_time_wait_interval
tcp_ecn_permitted	ecn
tcp_conn_req_max_q	_conn_req_max_q
tcp_conn_req_max_q0	_conn_req_max_q0
tcp_conn_req_min	_conn_req_min
tcp_rst_sent_rate_enabled	_rst_sent_rate_enabled
tcp_rst_sent_rate	_rst_sent_rate
tcp_keepalive_interval	_keepalive_interval
tcp_ip_abort_interval	_ip_abort_interval
tcp_rexmit_interval_initial	_rexmit_interval_initial
tcp_rexmit_interval_max	_rexmit_interval_max
tcp_rexmit_interval_min	_rexmit_interval_min
tcp_rexmit_interval_extra	_rexmit_interval_extra
tcp_tstamp_if_wscale	_tstamp_if_wscale
tcp_rcv_hiwat_minmss	_recv_hiwat_minmss

UDP Parameter Name Changes (Oracle Solaris 11)

The following UDP parameters have been renamed to UDP properties.

You can set a UDP property by using syntax similar to the following:

```
# ipadm set-prop -p send_buf=57344 udp
```

You can display UDP property information by using syntax similar to the following:

```
# ipadm show-prop -p send_buf udp
PROTO PROPERTY          PERM CURRENT    PERSISTENT  DEFAULT  POSSIBLE
udp   send_buf            rw   57344         57344       57344     1024-2097152
```

TABLE A-3 UDP Parameter Name Changes

Previous UDP Parameter Name	UDP Property Name
udp_max_buf	max_buf
udp_xmit_hiwat	send_buf
udp_recv_hiwat	recv_buf

SCTP Parameter Name Changes (Oracle Solaris 11)

The following SCTP parameters have been renamed to SCTP properties.

You can set an SCTP property by using syntax similar to the following:

```
# ipadm set-prop -p _max_init_retr=8 sctp
```

You can display SCTP property information by using syntax similar to the following:

```
# ipadm show-prop -p _max_init_retr sctp
PROTO PROPERTY          PERM CURRENT    PERSISTENT  DEFAULT  POSSIBLE
sctp  _max_init_retr      rw   8             8           8         0-128
```

TABLE A-4 SCTP Parameter Name Changes

Previous SCTP Parameter Name	SCTP Property Name
sctp_max_init_retr	_max_init_retr
sctp_pa_max_retr	_pa_max_retr
sctp_pp_max_retr	_pp_max_retr
sctp_cwnd_max	_cwnd_max
sctp_ipv4_ttl	_ipv4_ttl

TABLE A-4 SCTP Parameter Name Changes (Continued)

Previous SCTP Parameter Name	SCTP Property Name
sctp_heartbeat_interval	_heartbeat_interval
sctp_new_secret_interval	_new_secret_interval
sctp_initial_mtu	_initial_mtu
sctp_deferred_ack_interval	_deferred_ack_interval
sctp_ignore_path_mtu	_ignore_path_mtu
sctp_initial_ssthresh	_initial_ssthresh
sctp_ipv6_hoplimit	_ipv6_hoplimit
sctp_xmit_lowat	_xmit_lowat
sctp_xmit_hiwat	send_buf
sctp_rcv_hiwat	recv_buf
sctp_max_buf	max_buf
sctp_rto_min	_rto_min
sctp_rto_max	_rto_max
sctp_rto_initial	_rto_initial
sctp_cookie_life	_cookie_life
sctp_max_in_streams	_max_in_streams
sctp_initial_out_streams	_initial_out_streams
sctp_shutack_wait_bound	_shutack_wait_bound
sctp_maxburst	_maxburst
sctp_addip_enabled	_addip_enabled
sctp_prsctp_enabled	_prsctp_enabled

Parameters That Are Obsolete or Have Been Removed (Oracle Solaris 11 and 11.1)

The following section describes parameters that are obsolete or have been removed from more recent Oracle Solaris releases.

NCA Parameters

The network cache accelerator (NCA) parameter information is obsolete and has been removed.

consistent_coloring

This parameter is obsolete starting in the Oracle Solaris 11.1 release.

rstchown

This parameter is obsolete starting in the Oracle Solaris 11 release.

Description	Indicates whether the POSIX semantics for the chown system call are in effect. POSIX semantics are as follows: <ul style="list-style-type: none"> ▪ A process cannot change the owner of a file, unless it is running with UID 0. ▪ A process cannot change the group ownership of a file to a group in which it is not currently a member, unless it is running as UID 0. For more information, see chown(2) .
Data Type	Signed integer
Default	1, indicating that POSIX semantics are used
Range	0 = POSIX semantics not in force or 1 = POSIX semantics used
Units	Toggle (on/off)
Dynamic?	Yes
Validation	None
When to Change	When POSIX semantics are not wanted. Note that turning off POSIX semantics opens the potential for various security holes. Doing so also opens the possibility of a user changing ownership of a file to another user and being unable to retrieve the file without intervention from the user or the system administrator.
Commitment Level	Obsolete

Obsolete TCP/IP Module Parameters (Oracle Solaris 11)

ip_multidata_outbound

This parameter is obsolete in the Oracle Solaris 11 release.

tcp_mdt_max_pbufs

This parameter is obsolete in the Oracle Solaris 11 release.

Revision History for This Manual

This section describes the revision history for this manual.

- [“Current Version: Oracle Solaris 11.1 Release”](#) on page 187
- [“New or Changed Parameters in the Oracle Solaris Release”](#) on page 187

Current Version: Oracle Solaris 11.1 Release

The current version of this manual applies to the Oracle Solaris 11.1 release.

New or Changed Parameters in the Oracle Solaris Release

The following sections describe new, changed, or obsolete kernel tunables.

- Oracle Solaris 11.1: Oracle Solaris ZFS tunable information is provided in [Chapter 3, “Oracle Solaris ZFS Tunable Parameters.”](#)
- Oracle Solaris 11.1: The `maxusers`, `max_nprocs`, `ngroups_max`, `pidmax`, and `segkpsize` parameters have been revised in the Oracle Solaris 11.1 release. For more information, see [Chapter 2, “Oracle Solaris Kernel Tunable Parameters.”](#)
- Oracle Solaris 11.1: The `dnlc_dircache_percent` parameter is new in the Oracle Solaris 11.1 release. For more information, see [“dnlc_dircache_percent”](#) on page 65.
- Oracle Solaris 11: The `rstchown` parameter is obsolete. For more information, see [“What's New in Oracle Solaris System Tuning?”](#) on page 17.
- Oracle Solaris 11: The `ipadm` command replaces the `ndd` command for setting TCP, IP, UDP, and SCTP properties. In addition, the names of the network parameters have changed to better correlate to the `ipadm` format. For more information, see [“Overview of Tuning IP Suite Parameters”](#) on page 129.
- Oracle Solaris 11: This release includes the `disp_rechoose_interval` parameter. For more information, see [“disp_rechoose_interval”](#) on page 73.

- Oracle Solaris 11: This release includes the `ngroups_max` parameter description. For more information, see [“ngroups_max” on page 38](#).
- Oracle Solaris 11: This release includes the `zfs_arc_min` and `zfs_arc_max` parameter descriptions. For more information, see [“zfs_arc_min” on page 82](#) and [“zfs_arc_max” on page 83](#).
- Oracle Solaris 11 : This release includes several `igb` and `ixgbe` network driver parameters. For more information, see [“igb Parameters” on page 56](#) and [“ixgbe Parameters” on page 57](#).
- Oracle Solaris 11: This release includes the `ddi_msix_alloc_limit` parameter that can be used to increase the number of MSI-X interrupts that a device instance can allocate. For more information, see [“ddi_msix_alloc_limit” on page 55](#).
- Oracle Solaris 11: This release includes the `kmem_stackinfo` parameter, which can be enabled to monitor kernel thread stack usage. For more information, see [“kmem_stackinfo” on page 53](#).
- Oracle Solaris 11: Memory locality group parameters are provided in this release. For more information about these parameters, see [“Locality Group Parameters” on page 78](#).
- Oracle Solaris 11: Parameter information was updated to include sun4v systems. For more information, see the following references:
 - [“maxphys” on page 61](#)
 - [“tmpfs : tmpfs_maxkmem” on page 66](#)
 - [“SPARC System Specific Parameters” on page 75](#)

Index

A

_addip_enabled, 166
_addrs_per_if, 133
autofs, 172
autoup, 32

C

_conn_req_max_q, 145
_conn_req_max_q0, 146
_conn_req_min, 146
_cookie_life, 164
cron, 172
_cwnd_max, 141,158

D

ddi_msix_alloc_limit parameter, 55
default_stksize, 26
default_tsb_size, 76
_deferred_ack_interval, 136,161
_deferred_acks_max, 137
desfree, 41
dhcpageant, 172
disp_rechoose_interval, 73,178
dnlc_dir_enable, 64
dnlc_dir_max_size, 65
dnlc_dir_min_size, 64
dnlc_dircache_percent, 65
doiflush, 33

dopageflush, 33

E

ecn, 144
enable_tsb_rss_sizing, 77

F

fastscan, 46
forwarding, 132
fs, 173
fsflush, 30
ftp, 173

H

handspreadpages, 48
_heartbeat_interval, 159
hires_tick, 74
hoplimit (ipv6), 132
hostmodel, 133

I

_icmp_err_burst, 130
_icmp_err_interval, 130
_icmp_return_data_bytes, 135
_ignore_path_mtu, 161

inetinit, 173
init, 174
_initial_mtu, 160
_initial_out_streams, 165
_initial_ssthresh, 161
intr_force, 56
intr_throttling, 58
_ip_abort_interval, 151
ip_squeue_fanout, 134
ip_squeue_worker_wait, 149
ipcl_conn_hash_size, 149
ipsec, 174
_ipv4_ttl, 159
_ipv6_hoplimit, 159

K

kbd, 174
_keepalive_interval, 150
keyserv, 174
kmem_flags, 52
kmem_stackinfo, 53

L

largest_anon_port, 148, 156, 167
lgrp_mem_pset_aware, 79
_local_dack_interval, 137
_local_dacks_max, 138
_local_slow_start_initial, 142
logevent_max_q_sz, 28
login, 175
lotsfree, 40
lpg_alloc_prefer, 78
lwp_default_stksize, 27

M

max_buf (SCTP), 163
max_buf (TCP), 140
max_buf (UDP), 155
_max_in_streams, 165

_max_init_retr, 157
max_nprocs, 37, 178
maxpgio, 49
maxphys, 61
maxpid, 36
maxuprc, 37
maxusers, 34, 178
min_percent_cpu, 47
minfree, 42
moddebug, 54
mpathd, 175
mr_enable, 56

N

ncsize, 63
ndd, 130
_new_secret_interval, 160
nfs_max_threads, 105
nfs:nacache, 118
nfs:nfs_allow_preepoch_time, 97
nfs:nfs_async_clusters, 115
nfs:nfs_async_timeout, 118
nfs:nfs_cots_timeo, 98
nfs:nfs_disable_rddir_cache, 113
nfs:nfs_do_symlink_cache, 100
nfs:nfs_dynamic, 101
nfs:nfs_lookup_neg_cache, 102
nfs:nfs_nra, 107
nfs:nfs_shrinkreaddir, 110
nfs:nfs_write_error_interval, 112
nfs:nfs_write_error_to_cons_only, 112
nfs:nfs3_async_clusters, 116
nfs:nfs3_bsize, 114
nfs:nfs3_cots_timeo, 98
nfs:nfs3_do_symlink_cache, 100
nfs:nfs3_dynamic, 102
nfs:nfs3_jukebox_delay, 119
nfs:nfs3_lookup_neg_cache, 103
nfs:nfs3_max_threads, 106
nfs:nfs3_max_transfer_size, 120
nfs:nfs3_max_transfer_size_clts, 122
nfs:nfs3_max_transfer_size_cots, 122
nfs:nfs3_nra, 108

nfs:nfs3_pathconf_disable_cache, 96
 nfs:nfs3_shrinkreaddir, 111
 nfs:nfs4_async_clusters, 117
 nfs:nfs4_bsize, 114
 nfs:nfs4_cots_timeo, 99
 nfs:nfs4_do_symlink_cache, 101
 nfs:nfs4_lookup_neg_cache, 104
 nfs:nfs4_max_threads, 107
 nfs:nfs4_max_transfer_size, 121
 nfs:nfs4_nra, 109
 nfs:nfs4_pathconf_disable_cache, 96
 nfs:nrnode, 109
 nfslogd, 175
 ngroups_max, 38
 noexec_user_stack, 30
 nss, 175
 nstrpush, 70

O

Oracle database tuning, ZFS file systems, 90

P

pageout_reserve, 44
 pages_before_pager, 48
 pages_pp_maximum, 45
 passwd, 175
 _pathmtu_interval, 135
 physmem, 26
 pidmax, 36, 178
 _policy_mask, 156
 _pp_max_retr, 158
 primarycache, ZFS file system property, 90
 _prsctp_enabled, 166
 pt_cnt, 68
 pt_max_pty, 69
 pt_pctofmem, 69

R

recordsize, ZFS file system property, 89

recv_buf (SCTP), 162
 recv_buf (TCP), 140
 recv_buf (UDP), 154
 _recv_hiwat_minmss, 153
 reserved_procs, 35
 _respond_to_echo_broadcast, 131
 _respond_to_echo_multicast, 131
 _rev_src_routes, 143
 _rexit_interval_extra, 153
 _rexit_interval_initial, 151
 _rexit_interval_max, 152
 _rexit_interval_min, 152
 rlim_fd_cur, 62
 rlim_fd_max, 61
 rpcmod:clnt_idle_timeout, 124
 rpcmod:clnt_max_conns, 123
 rpcmod:cotsmaxdupreqs, 126
 rpcmod:maxdupreqs, 125
 rpcmod:svc_default_stksize, 124
 rpcmod:svc_idle_timeout, 124
 _rst_sent_rate, 147
 _rst_sent_rate_enabled, 147
 rstchown, 185
 _rto_max, 164
 _rto_min, 163
 rx_copy_threshold, 60
 rx_limit_per_intr, 58
 rx_queue_number, 58
 rx_ring_size, 59

S

sack, 142
 sctp_maxburst, 166
 secondarycache, ZFS file system property, 90
 segkpsize, 178
 segspt_minfree, 72
 send_buf (SCTP), 162
 send_buf (TCP), 139
 send_buf (UDP), 154
 _send_redirects, 131
 _shutack_wait_bound, 165
 _slow_start_after_idle, 142
 _slow_start_initial, 141

slowscan, 47
smallest_anon_port, 148,155,167
strmsgsz, 70,71
su, 176
swapfs_minfree, 51,179
swapfs_reserve, 50
syslog, 176

T

tar, 176
throttlefree, 43
_time_wait_interval, 143
timer_max, 74
tmpfs_maxkmem, 66
tmpfs_minfree, 67
tsb_alloc_hiwater, 75
tsb_rss_factor, 77
_tstamp_always, 139
_tstamp_if_wscale, 153
ttl (ipv4), 132
tune_t_fsflushr, 31
tune_t_minarmem, 45
tx_copy_threshold, 60
tx_queue_number, 57
tx_ring_size, 59

U

utmpd, 176

W

_wscale_always, 138

X

_xmit_lowat, 162

Z

zfs_arc_max, 83,177
zfs_arc_min, 82,177
ZFS file system property
 primarycache, 90
 recordsize, 89
 secondarycache, 90
ZFS file systems, tuning for an Oracle database, 90
zfs_prefetch_disable, 84