



Xdd

User's Guide

Version 6.4

I/O Performance, Inc.

Copyright © 1992-2006, I/O Performance, Inc.

Author:	Thomas M. Ruwart
Date:	December 16, 2005
Phone:	612-850-2918
Email:	tmruwart@ioperformance.com

Change History:

The revision numbers have the following meaning. The “number” such as 5.7, 5.8, 5.9, ...etc. represent major release changes such as the addition of new options or sections to xdd. The letter following the number represents a maintenance release (such as 5.9z is maintenance release z). A number appended to the maintenance release letter (such as z7) is a site specific fix to a particular maintenance release.

Revision	Date	Author	Description of Changes
5.9b	2/1//2002	Thomas M. Ruwart	
5.9z4	4/1/2002	Thomas M. Ruwart	Updated minor parts
5.9z7	3/1/2003	Thomas M. Ruwart	Updated the sysinfo output
5.9z7e	4/24/2003	Thomas M. Ruwart	-queue-depth and -seek interleave work
5.9z7f	5/20/2003	Thomas M. Ruwart	Fixed problems with trace output on Windows
5.9z7g	6/19/2003	Thomas M. Ruwart	Fixed problem random I/O using -quedepth Fixed 4GB limit problem on Windows
6.0	6/25/2003	Thomas M. Ruwart	<p>Added the following options</p> <ul style="list-style-type: none"> -throttle <MB/sec> -runtime <seconds> -ts wrap -rwratio <percent> <p>The “-procoffset” option has been removed. The functionality has been replaced with the -startoffset target <target#> <offset></p> <p>Modified the syntax for the following options to take the “target <target#>” argument</p> <ul style="list-style-type: none"> -op target <target#> <operation> -reqsize target <target#> <reqsize> -timelimit target <target#> <seconds> -throttle target <target#> <MB/sec> <p>Fixed some bugs with saving seek lists</p>
6.1	10/1/2003	Thomas M. Ruwart	Added the -starttrigger, -stoptrigger, and -lockstep options

6.2	3/31/04	Thomas M. Ruwart	<p>Added –deskew option</p> <p>Added –sharedmemory option for AIX</p> <p>Added –verify location/content option</p> <p>Removed the old –verify option</p> <p>Fixed problem with semop error on exit</p> <p>Revised the way semaphore resources are allocated and removed.</p>
6.2c	1/16/05	Guess.	Added support for Mac OS X - Darwin
6.3	1/23/05	Guess again...	<p>Restructuring the code</p> <p>Added shared memory buffer support for linux, solaris, and AIX</p> <p>Added support for zone specification</p> <ul style="list-style-type: none"> -zone start -zone size -zone map -zone bins <p>Added the –profile option that will automatically perform request-size profiles</p> <p>Enabled Read-After-Write for all platforms</p> <p>Added the “Under the Hood” section to this manual</p> <p>Added more examples to this manual</p>
6.4	12/15/2005	Thomas M. Ruwart	<p>Added IOType and Request Size to the output line.</p> <p>Restructured the code as follows:</p> <ul style="list-style-type: none"> • Added parse.c file for all the commands • Added barrier.c file that contains the barrier management code • Added initialization.c to handle all the initialization stuff that happens when xdd is started • Added results.c that contains all the routines that manage and display results. • Added sg.c that contains all the routines to do I/O to the SCSI Generic devices in Linux • Added access_pattern.c that contains all the code that comes up with the seek list for xdd to run • Added time_stamp.c that contains all the time stamp management functions • Changed globtim.c to global_time.c to be more descriptive • Added hpss.c that contains all the hpss access methods – not yet actually implemented

			<ul style="list-style-type: none">• Added read_after_write.c that manages read_after_write operations whenever that option is specified. <p>Added the following options:</p> <ul style="list-style-type: none">-miall-nomemlock-noproclock-version-report_threshold-sgio

1	INTRODUCTION	7
1.1	ABOUT THIS DOCUMENT	7
1.2	ABOUT XDD	7
1.3	LICENSE AGREEMENT.....	7
1.4	DISTRIBUTION OF XDD.....	7
1.5	SYSTEM REQUIREMENTS	7
2	WHAT'S NEW	8
2.1	IN XDD 6.4.....	8
2.2	IN XDD 6.3.....	8
2.3	IN XDD6.2	9
2.4	IN XDD6.1	9
2.5	IN XDD6.0	9
3	COMPILING AND INSTALLING XDD.....	9
3.1	XDD SOURCE CODE OVERVIEW	9
3.1.1	<i>xdd.c and xdd.h</i>	10
3.1.2	<i>timeserver.c</i>	10
3.1.3	<i>gettime.c</i>	10
3.1.4	<i>access_pattern.c</i>	10
3.1.5	<i>barrier.c</i>	10
3.1.6	<i>global_time.c</i>	10
3.1.7	<i>initialization.c</i>	10
3.1.8	<i>parse.c</i>	10
3.1.9	<i>pclk.c and pclk.h</i>	11
3.1.10	<i>read_after_write.c</i>	11
3.1.11	<i>results.c</i>	11
3.1.12	<i>ticker.c and ticker.h</i>	11
3.1.13	<i>time_stamp.c</i>	11
3.1.14	<i>misc.h</i>	11
3.1.15	<i>nt_unix_compat.c</i>	11
3.1.16	<i>Compile Notes</i>	11
3.2	WINDOWS™ NT™, WINDOWS™ 2000™, AND WINDOWS™ XPTM	12
3.3	LINUX	12
3.4	SOLARIS™ FROM SUN (SPARC™ VERSION)	12
3.5	AIX™ FROM IBM.....	12
3.6	HPUX™ FROM HEWLETT-PACKARD	12
3.7	IRIX™ FROM SGI.....	13
3.8	MAC OS X™ FROM APPLE	13
4	THEORY OF OPERATIONS	13
4.1	BASIC OPERATION	13
4.2	COMMAND LINE OPTIONS AND THE SETUP FILE	13
4.3	OPERATION SPECIFICATION AND REQUEST SIZES	14
4.4	TARGET SPECIFICATION AND MULTIPLE TARGET SYNCHRONIZATION	14
4.5	TIME SYNCHRONIZATION ACROSS MULTIPLE COMPUTER SYSTEMS.....	16
4.6	XDD RUN TIME OR RUN LENGTH	17
4.7	I/O RANGE	17
4.8	ACCESS PATTERNS	17

4.9	READING AND WRITING DEVICES AND FILES	18
4.10	PROCESSOR ALLOCATION AND PRIORITY ASSIGNMENT	20
4.11	I/O TIME STAMPING.....	20
5	RUNNING XDD PROGRAMS	21
5.1	XDD COMMAND LINE ARGUMENTS SYNOPSIS	21
5.2	TARGET-SPECIFIC OPTIONS	29
5.3	LOCKSTEP OPERATIONS	30
5.4	TIMESERVER AND GETTIME	30
5.5	DESKEW	31
6	RUNTIME HINTS	31
6.1	WINDOWS	31
6.2	IRIX.....	31
6.3	AIX.....	32
6.4	SOLARIS	32
6.5	LINUX	32
6.6	HPUX	32
6.7	TRU64.....	32
7	OUTPUT AND REPORTS.....	34
7.1	REPORTING OPTIONS	34
7.2	WHAT THE NUMBERS REALLY MEAN.....	36
8	PERFORMANCE TUNING HINTS	36
8.1	CACHES AND WRITE PERFORMANCE.....	36
8.2	FIBRE CHANNEL FRAME SIZES	36
9	EXAMPLES.....	36
9.1	EXAMPLE DISPLAY OUTPUT	38
9.2	EXAMPLE TIME STAMP OUTPUT.....	39
10	UNDER THE HOOD	42
10.1	XDD GENERAL OPERATION	42
10.2	THE XDD BUFFER MEMORY LAYOUT	43
10.3	THE XDD THREAD STRUCTURE	43
10.4	XDD BARRIERS.....	44
11	THE GNU PUBLIC LICENSE	44
11.1	PREAMBLE	45
11.2	TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION	45
11.3	NO WARRANTY	46
12	ACKNOWLEDGEMENTS	46

1 Introduction

1.1 About this document

This document is a user's guide to compiling, installing, and running the xdd program. It also has a variety of examples and hints for understanding the use and results of xdd measurements.

1.2 About xdd

Xdd is a tool for measuring and characterizing disk subsystem I/O on single systems and clusters of systems. It is a command-line based tool that grew out of the UNIX world and has been ported to run in Windows environments as well. It is designed to provide consistent and reproducible performance measurements of disk I/O traffic. There are three basic components to xdd that include the xdd program itself, a timeserver program, and a gettime program. The timeserver and gettime programs are used to synchronize the clocks of xdd programs simultaneously running across multiple computer systems.

1.3 License Agreement

It is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program/document in a file named 'Copying' or 'gpl.txt' or in the section entitled *The GNU Public License* in this document; if not, write to the Free Software Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139 or visit their web site at <http://www.gnu.org/licenses/gpl.html>.

1.4 Distribution of xdd

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

1.5 System Requirements

Xdd can run on a relatively minimally configured system but it is recommended that adequate main memory capacity and processor speed be available when using some of the more advanced features of xdd. The xdd program itself is not CPU intensive but the faster the processor and the less loaded a system is, the more accurate and consistent the results will be. As a rule, a minimum of 128MB of main memory is recommended. All other system parameters are left to the discretion of the user.

2 What's new

2.1 In xdd 6.4

Added IOType and Request Size to the output line. This information is also included in the target information that is reported before the results but it was always a pain to have to look back through all that to figure out where that information was. So it is now in each output line.

Of course there cannot be a new release of xdd without new options. Thus, the following options have been added to this release:

- nomemlock – do not lock memory buffers
- noproclock – do not lock process into memory
- miall – essentially does a -nomemlock and a -noproclock to get rid of annoying messages that get generated if you are running xdd as a normal user and not super user.
- version – Display the xdd version number
- report_threshold – display I/O operations that exceed a certain time threshold – great for identifying those operations that take much longer than others
- sgio – My personal favorite – do I/O to the SCSI Generic devices in a Linux system

2.2 In xdd 6.3

The xdd source file was getting way too large so in an attempt to get some sanity back into the source tree the source code files were restructured. The restructured the code as follows:

- Added parse.c file for all the commands
- Added barrier.c file that contains the barrier management code
- Added initialization.c to handle all the initialization stuff that happens when xdd is started
- Added results.c that contains all the routines that manage and display results.
- Added sg.c that contains all the routines to do I/O to the SCSI Generic devices in Linux
- Added access_pattern.c that contains all the code that comes up with the seek list for xdd to run
- Added time_stamp.c that contains all the time stamp management functions
- Changed globtim.c to global_time.c to be more descriptive
- Added hpss.c that contains all the hpss access methods – not yet actually implemented
- Added read_after_write.c that manages read_after_write operations whenever that option is specified.

Added shared memory buffer support for linux, solaris, and AIX

Added support for zone specification

- zone start
- zone size
- zone map
- zone bins

Added the -profile option that will automatically perform request-size profiles

Enabled Read-After-Write for all platforms

Added the “Under the Hood” section to this manual

Added more examples to this manual

Added support for Mac OS X a.k.a. Darwin. The direct I/O and multi-processor support is not yet implemented but should be by release 6.3.

2.3 In xdd6.2

Xdd6.2 includes the `-deskew` option that is useful when testing a large number of targets. Deskew will measure the I/O performance when all targets are actually running rather than from the time the first one starts to the time when the last one ends. There is also a new option called `-sharedmemory` that currently only works with AIX. This option causes xdd to allocate the I/O buffer in a shared memory segment using `shmget()` and `shmat()` rather than the normal `valloc/malloc`. There are performance reasons for doing this in AIX and it is useful to try it if testing peak bandwidth on an AIX system. Finally, there were several annoying little bugs that have been squashed.

2.4 In xdd6.1

New stuff includes the `-starttrigger`, `-stoptrigger`, and `-lockstep` options that allow for targets to control the operation of other targets.

2.5 In xdd6.0

There are several new options in version 6.2 of xdd that allow for the specification of the parent directory of targets (**-targetdir**), specify the total transfer size in kilobytes (**-kbytes**), mixing read and write operations within a single run (**-rwratio**), to run the xdd program for a specific amount of time rather than a number of bytes or operations (**-runtime**), and the ability to throttle the transfer rate (**-throttle**). The most notable change to this release however is the ability to set many of the options for specific targets rather than having all targets use the same options (see section on Target Specific Options). This allows for tailoring the I/O behavior of each target to meet specific I/O performance/behavior requirements.

3 Compiling and Installing xdd

This section describes how to compile and install xdd on the following systems:

- Windows™ NT™ and Windows™ 2000™
- Linux
- Solaris™ from Sun on both SPARC™ and Intel platforms
- AIX™ from IBM
- HPUX™ from Hewlett-Packard
- IRIX™ from SGI
- Mac OS X - Darwin

3.1 Xdd Source Code Overview

The xdd distribution comes with all the source code necessary to install xdd and the companion programs for the timeserver and the gettime utility programs. For UNIX systems, a *make* file is included to build any of these programs. This release of xdd comes with the following source files:

<i>Program files</i>	<i>Header files</i>
xdd.c	xdd.h
timeserver.c	
gettime.c	
<i>Subroutine files</i>	<i>Subroutine header files</i>
access_pattern.c	
barrier.c	

global_time.c	
initialization.c	
parse.c	
pclk.c	pclk.h
read_after_write.c	
results.c	
ticker.c	ticker.h
time_stamp.c	
	misc.h
nt_unix_compat.c	

Table 2. The xdd source code tree.

3.1.1 xdd.c and xdd.h

xdd.c contains all the program routines that are specific to running xdd. The xdd.h header file is common to xdd.c and to all the other programs that use xdd subroutine files.

3.1.2 timeserver.c

timeserver is the master time server that runs on a single machine and is used as a reference clock by xdd programs running on other machines. This program does not have to be compiled in order to compile/run xdd on a single machine.

3.1.3 gettime.c

gettime is a program that is used in conjunction with the time server functions when running xdd on multiple machines (see section on Timeserver Functions on Multiple Machines). This program does not have to be compiled in order to compile/run xdd on a single machine.

3.1.4 access_pattern.c

These subroutines generate the seek list used by xdd to access a target.

3.1.5 barrier.c

This contains all the barrier subroutines that are used to control the execution and timing of I/O operations.

3.1.6 global_time.c

global_time.c contains all the functions necessary to contact the master time server machine and establish the Global Time.

3.1.7 initialization.c

This file contains most of the subroutines used during program and target thread initialization.

3.1.8 parse.c

This file contains the subroutines that parse the command-line options and the setup file if it is specified. The help and usage functions are also located in this file.

3.1.9 *pclk.c* and *pclk.h*

pclk.c contains all the subroutines provide pico-second clock values to the calling function. It calls the ticker functions that read the actual clock values specific to a particular architecture.

3.1.10 *read_after_write.c*

This file contains the subroutines that control a read-after-write operation.

3.1.11 *results.c*

results.c contains all the routines that process and display the results information.

3.1.12 *ticker.c* and *ticker.h*

ticker.c contains all the subroutines that are specific to a specific machine architecture that are required to access that's machine's high resolution clock.

3.1.13 *time_stamp.c*

This contains the time stamp subroutines.

3.1.14 *misc.h*

misc.h contains all the miscellaneous definitions that are common to all the programs.

3.1.15 *nt_unix_compat.c*

nt_unix_compat.c contains all the UNIX subroutines that are not supported in a standard Windows™ NT™ or Windows™ 2000™ environment. These subroutines provide a mapping from the UNIX system call (i.e. sleep, getpid, pthread_create, ...etc.) to the equivalent Windows™ system call (i.e. Sleep, GetCurrentThreadID, CreateThread, ...etc.). This file is only used when compiling in a Windows™ environment. It is not part of a UNIX compile.

3.1.16 *Compile Notes*

Before building xdd on any of the UNIX-based systems, it may be necessary to “clean” the files first. Since the xdd distribution is built on a Windows platform, certain Windows artifacts may contaminate the source, header, and make files. More specifically, any text file (such as *make* files, *.c* and *.h* files) may contain hex characters “0D” (ctl-M) at the end of each line. This is not interpreted correctly in most UNIX environments and makes it impossible to compile any of the xdd source code. Therefore, these hex “0D” characters must first be removed before any compilation can be done.

A simple program and script is provided to perform this removal operation automatically. The program is called *stripm.c* and the associated script is *stripm.csh*. Simply run the *stripm.csh* script and it will compile the *strip.c* program and run it on all affected files. At the UNIX command prompt, use the following command to execute *stripm.csh*:

```
root#    csh stripm.csh
Found 32 instances of 0x0d
Found  0 instances of 0x0d
...
Found 12 instances of 0x0d
root#
```

There will be a number of the “Found...” messages that can be ignored. After has been run, it is safe to make xdd, the timeserver, and the gettime programs.

3.2 Windows™ NT™ , Windows™ 2000™, and Windows™ XP™

The distribution directory hierarchy includes the source files and header files as well as a “debug” directory. The debug directory contains the executable xdd file (*xdd.exe*) that is sufficient to run on any Windows™ platform. The source directory contains the *.dsw* files necessary to compile xdd using Microsoft’s Visual C++™ compiler. This is the only compiler that has been used to build the xdd executable in the Windows™ environment. Use of any other compiler can produce unpredictable results.

3.3 Linux

To make xdd, timeserver, and gettimeofday in a uni-processor Linux environment, use the following command:

```
make -f linuxup.makefile
```

or for an SMP version use the following:

```
make -f linuxsmp.makefile
```

This uses the gcc compiler and associated libraries. Insure that the pthreads libraries are installed for correct compilation and linking.

3.4 Solaris™ from Sun (SPARC™ version)

To make xdd, timeserver, and gettimeofday in a Solaris™ environment, use the following command:

```
make -f solaris.makefile all    (for Solaris™ on a SPARC™ platform)
```

```
make -f isolaris.makefile all  (for Solaris™ on an Intel platform)
```

This uses the gcc compiler and associated libraries. Insure that the pthreads libraries are installed for correct compilation and linking.

There is an issue with running xdd on Solaris that involves the default number of semaphores that Solaris allows each process to use. It is normally too low for xdd to run and xdd generates errors that say things like “cannot allocate barrier for ..., not enough space”. The following parameters can be put in */etc/system* on a solaris system and a reboot will take care of this problem:

```
set semsys:seminfo_semmni=4096
set semsys:seminfo_semmns=8192
set semsys:seminfo_semmmap=4098
set shmsys:shminfo_shmmni=512
set shmsys:shminfo_shmseg=32
```

These settings may be a little high and can be adjusted to meet the local system constraints.

3.5 AIX™ from IBM

To make xdd, timeserver, and gettimeofday in a AIX™ environment, use the following command:

```
make -f aix.makefile all
```

This uses the xlc compiler and associated libraries. Insure that the pthreads libraries are installed for correct compilation and linking.

3.6 HPUX™ from Hewlett-Packard

To make xdd, timeserver, and gettimeofday in a HPUX™ environment, use the following command:

```
make -f hpux.makefile all
```

This uses the gcc compiler and associated libraries. Insure that the pthreads libraries are installed for correct compilation and linking.

3.7 IRIX™ from SGI

To make xdd, timeserver, and gettime in an IRIX™ environment, use the following command:

```
make -f irix.makefile all
```

This uses the IRIX™ compiler and associated libraries. Insure that the pthreads libraries are installed for correct compilation and linking.

3.8 Mac OS X™ from Apple

To make xdd, timeserver, and gettime in a Mac OS X™ environment, use the following command:

```
make -f macosx.makefile all
```

This uses the gcc compiler and associated libraries. Make sure these are installed on your Mac because even though they come with the Mac, they are not installed by default.

4 Theory of Operations

4.1 Basic Operation

Xdd is a program that performs data transfer operations between memory and a disk device or file (or multiple disk devices / files) and collects and displays performance information about these I/O operations. Xdd creates one or more threads for every device or file under test. Each xdd thread issues I/O operations to a "target" which is either a disk storage device or a file. Each I/O operation is either a read or a write operation of a fixed size known as the "request size". An xdd "run" consists of several "passes", the number of which is specified by the "--passes" option. Each pass will execute some number of I/O requests on the specified target(s) at the given request size. In general, each pass is identical to the previous passes in a run with respect to the request size, number of requests to issue, and the access pattern unless certain options are used to alter these parameters between passes. Passes are run one after another with no delays between passes unless a pass delay is specified with the "--delay" option.

Multiple passes within an xdd run are used to determine the reproducibility of the results. In theory, the results from each pass in an xdd run should be the same or at least very close to the same given the same set of run-time parameters.

4.2 Command Line Options and the Setup File

This version of xdd has a command-line interface that requires all the run-time parameters to be specified either on the xdd invocation command line or in a "setup" file. The format of the setup file is similar to the xdd command-line in that the options can be simply entered into the setup file the same as they would be seen on the command line. The following example shows an xdd invocation using just the command-line and the same invocation using the setup file along with the contents of the setup file.

Command line:

```
xdd -op read -targets 1 /dev/scsi/disk1 -reqsize 8 -numreqs 128  
-verbose
```

Using a Setup file:

```
xdd -setup xddrun.txt
```

Where the setup file `xddrun.txt` is an ASCII text file that contains the following:

```
-op read -targets 1 /dev/scsi/disk1  
-reqsize 8  
-numreqs 128  
-verbose
```

4.3 Operation Specification and Request Sizes

The operation to perform is specified by the “**-op**” option. This can be either “*read*” or “*write*”. This version of `xdd` *will* mix read and write operations within an `xdd` run according to the read/write ratio set using the **-rwratio** option. Each r/w operation will transfer a given amount of data known as the “request size”. The request size is specified by the “**-reqsize**” option in units of “blocks”. A block is, by default, 1024 bytes but can be specified to be any positive integer value by using the “**-blocksize**” option. The “block” is used as the basic unit for all other options requiring a data size unless otherwise noted. It is recommended that the block size be specified as numbers that are integer multiples of 512 bytes (i.e. 1024, 2048, 5120, ...etc.) since this tends to be the predominant sector size for most storage devices at the current time.

4.4 Target Specification and Multiple Target Synchronization

All requests are sent to a “target” which can be either a disk device or a file. A single `xdd` run can operate on a single target or multiple targets simultaneously. Target names are specified using the “**-targets**” option. It is always necessary to specify the number of targets followed by the individual target names. In order to simplify the list of target names, the “**-targetdir**” option can be used to specify the directory where the target devices or files reside.

The execution of the `xdd` threads on multiple targets is synchronized through the use of “barriers”. Each `xdd` I/O thread initializes itself and waits for all the other `xdd` I/O threads to “reach at starting point”. Once all the `xdd` threads reach this point, they are all released simultaneously. Each `xdd` thread will run independently until it has either completed all of its requested I/O operations or reached its time limit (as specified by the **-timelimit** input parameter). At this point the `xdd` threads resynchronize with one another and begin another pass.

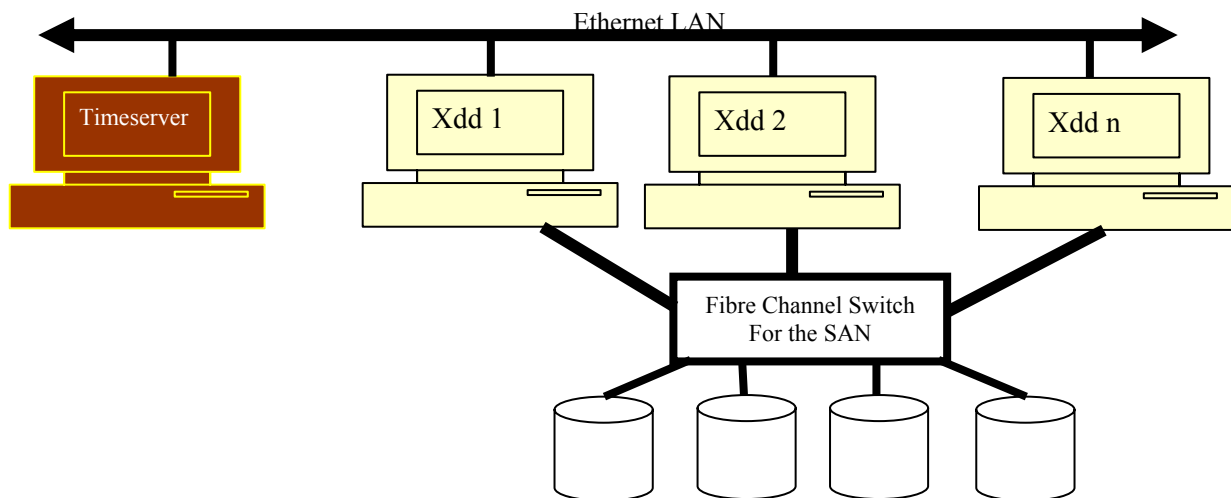
It is also possible to resynchronize the threads at specific points within a pass by using the “**-syncio**” and/or “**-syncwrite**” options. The **-syncio** option instructs each of the I/O threads to synchronize after some number of I/O operations specified as an argument to this option. The default is to synchronize only at the beginning of each pass. The **-syncwrite** option is used to synchronize “buffered” write operations at the end of each pass by flushing all the file system data buffers to the physical media. This option is not able to flush and cache buffers on the disk controllers or disk drives themselves.

Synchronizing the `xdd` I/O threads is done to insure that all the `xdd` I/O threads start at precisely the same time in order to avoid misleading results due to skewing the start times. It is possible to eliminate synchronization by specifying the **-nobARRIER** parameter. The result is that many of the

I/O start times can be significantly skewed from one another for all participating xdd I/O threads. However, that may in fact be the desired effect.

4.5 Time Synchronization Across Multiple Computer Systems

Synchronization can also be done across multiple machines each running xdd. This is accomplished by using the timeserver and gettime programs. The timeserver program provides a single global reference clock that is used by the time stamping in order to be able to more accurately correlate events in time from multiple computer systems. The timeserver program should be run on a single machine as a background task. This machine must be accessible via a LAN to all the machines that will be running xdd. This LAN should preferably be a lightly used LAN for optimum results.



The preferred way to make this work is to use “rsh” to start the xdd programs on each of the nodes using the “-starttime” and “-timeserver” options. The “-timeserver” option tells the xdd program the host name or IP address of the time server machine. The xdd program will contact this machine to determine the “global” time that all the other xdd programs will use as a frame of reference. The “-starttime” option specifies the time to start in “global time” units.

Before running the “rsh xdd” command line in a script, it is necessary to determine a global time sometime in the future at which all the xdd program will start. The way to do this is to use the “gettime” program to contact the time server, determine the global time, add a specified number of seconds, and display the result on standard out. This global start time can then be used as the argument to the “-starttime” option for each of the “rsh xdd” commands. An example script would look like so:

```

...
set g=`gettime -timeserver 192.10.11.12 -add 20`
# At this point ${g} will be the current global time plus
# 20 seconds.
foreach i ( 1 2 3 4 )
    rsh host${i} xdd -starttime ${g} -op read -targets 1
    /dev/dsk/c1d2s0 -mbytes 5 ...&
end
wait
...

```


This example scriptlet will set the local variable “\${g}” to the global time plus 20 seconds. This value is then passed to each of the xdd programs that are started on host1, host2, host3, and host4. This will result in each xdd program starting at exactly the same time. However, if any or all of the host machines running xdd are in the presence of a black hole, neutron star, or other extremely massive body (such as Oprah), the relativistic effects on the space-time continuum may produce unpredictable results.

4.6 Xdd Run Time or Run Length

It is necessary to specify a limit on how *long* xdd will run. There are several ways to do this. First, it is possible to explicitly specify the number of transfers to perform using the “**–numreqs**” option. This specifies the number of read or write calls to make for a single “pass”. It is also possible to simply specify the number of MegaBytes to transfer using the “**–mbytes**” option. It is important to note that a MegaByte in this context is 1048576 bytes. Finally, it is possible to specify a time limit for each xdd pass using the **–timelimit** option. This will cause each pass to terminate after the specified number of seconds has elapsed or after executing the specified number of requests or transferring the specified number of megabytes whichever occurs first.

4.7 I/O Range

For random I/O operations, each xdd I/O thread performs its I/O operations within a certain consecutive “range” of blocks on the target (see Figure 1). The range is either *implicitly* specified as the number of MegaBytes to transfer or *explicitly* specified using the “**–range**” option. For example, if the user specifies 2048 purely sequential data transfers at a request size of 128 blocks each, then the range will be implied as 262144 blocks (2048 * 128). However, if the user wants to transfer the same 2048 requests randomly over a 2 GigaByte area (or range) on the target then the range needs to be explicitly specified as 268435456 blocks or 2 GigaBytes.

The beginning of the I/O range defaults to the beginning of the target whether it is a device or a file. It is possible to specify a “starting offset” such that the I/O range begins at some distance into the target (see Figure 1). This can be done several ways. First, the “**–startoffset**” option can be used to start I/O operations at some distance into the target for an xdd run. The “**–passoffset**” can be used to incrementally move the starting offset further into the target device on subsequent passes within an xdd run by some specified number of blocks. Finally, the “**–targetoffset**” can be used to move the starting offset into the target device by a number of blocks that is determined by the target offset value times the target’s ordinal number. For example, given an xdd run on 4 targets with a target offset of 1024 and a starting offset of 0, I/O will start at block 0 on target 0, block 1024 on target 1, block 2048 on target 2, and block 3072 on target 3.

4.8 Access Patterns

The range can be specified to start anywhere within the target so long as care is taken to insure that the end of the range is still within the confines of the target. This is particularly true when randomly accessing blocks within a target that is a regular file. Within the I/O range, data access patterns can be either

- 1) purely sequential
- 2) staggered sequential
- 3) purely random

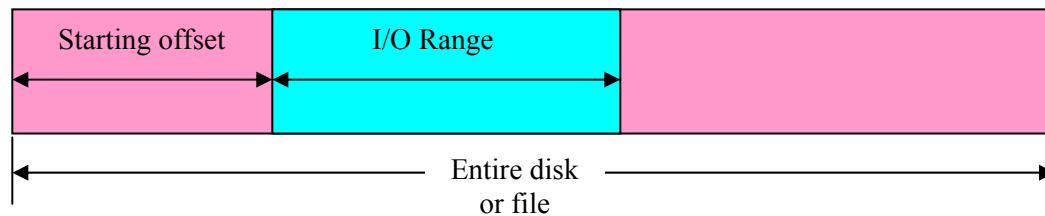


Figure 1. Example of the I/O range and the offset.

Purely sequential I/O is the default access pattern. This access pattern accesses consecutive data blocks starting at the beginning of the range to the end of the range.

A Staggered Sequential access pattern also starts at the beginning of the range and ends at the end of the range but only transfers every N^{th} data block. This access pattern is specified by the "-seek staggered" parameter. For example, if reading 256 MegaBytes within a 2 GigaByte range, the 256 MegaBytes is spread evenly over the entire 2GB range. Therefore, for a given request size, every 8th data block is read with gaps of unread data in between since 256 MegaBytes is 1/8th of 2 GigaBytes.

A purely Random access pattern is one that accesses data blocks at the specified request size randomly throughout the specified range. It is important to note that for a given range and given request size, the same random pattern is generated for each successive pass/run in order to yield *reproducible* results. The random access pattern used for each pass within a run can be changed by using the "-randomize" parameter. The random access pattern used for each run can be changed by using the "-seek seed" parameter.

It should be noted that when running *xdd* on a regular file especially in "random" mode, the Direct IO option (-dio) should be used to avoid using the file system buffer cache. The file system data caching mechanisms will produce misleading (and non-reproducible) results. See the section on Reading and Writing Devices and Files for more information.

Finally, it is possible to specify a "null" access pattern whereby the first block in the range is continually accessed. This is accomplished by specifying the "-seek none" option. This is useful for testing the effectiveness of caching algorithms and/or the speed of the cache on the target device or file.

4.9 Reading and Writing Devices and Files

Xdd read or writes devices or files. When running *xdd* on regular "files", the target files tend to be large and can "accidentally" be left behind after the testing. The -deletefile option will remove a target file at the conclusion of the *xdd* run. This option is not recommended when running *xdd* on device files for obvious reasons.

As previously implied, sequentially reading data is relatively straight forward: start at the beginning of the device or file plus the starting offset and read data until the end of the data range is reached or the time limit expires. When reading a file it is assumed that the file is at least as large

as the desired data range so that the read does not go past the end of the file. If the file is smaller than the desired or specified data range then a warning message is displayed regarding this condition.

Reading a file on most systems uses a file system buffer cache. In this mode, data is read into a file system buffer and then copied into the xdd I/O buffer by the CPU. This can cause two different problems with respect to the performance results. First, the maximum bandwidth for reading data into the machine for the first time is limited to the memory copy speed of the processor that can be much lower than the true bandwidth of the disk I/O subsystem under evaluation. Secondly, if all the data from the file fits into the file system buffer then subsequent reads to the same file will be satisfied by copying the data from the file system buffer cache and not from the actual disk subsystem. In this case, xdd is reporting the memory copy performance rather than the speed of the disk subsystem. To avoid this problem the Direct I/O, (“-dio”), option can be used to bypass the use of the file system buffer cache and forcing all read requests to access the disks. There are certain I/O request size and alignment restrictions that must be observed in order to use Direct I/O and these restrictions are Operating System dependent. These restrictions essentially state that the I/O requests must start on a “page-size” boundary and must be in units of the native block size of the underlying file system.

Writing devices is very similar to read files except for the data transfer direction. However, writing “files” has additional complexities that can affect the performance results. This is due primarily to the allocation mechanisms used by the file system manager to allocate disk space when writing a file for the first time. Generally speaking, each write operation will cause the file system manager to allocate space on the disk to accommodate the data being written. These allocation operations can require additional accesses to the disk that are “invisible” to xdd but do show up in the performance results as essentially slower write operations. In order to minimize the effects of this allocation process, the “-preallocate” option can be used to do the entire allocation operation before the file write operations take place. The argument to this option specifies the number of blocks to pre-allocate before the write operations start. This number should be greater than or equal to the number of blocks that will be in the file.

File write performance results can also be affected by the file system buffer cache. When writing a file the data is first copied into the file system buffer and later actually written to the physical disk subsystem. As in the case of the read operation, it is possible to use the “-dio” option to avoid this data copy operation. The same request size and alignment restrictions apply Direct I/O in this case as in the read case.

One last note about writing to a target device or file. The data pattern to be written to the target can be specified by the “-pattern” option. This will fill the data buffer with a single character data pattern or with a random data pattern depending on the argument to this option. No data verification is done with the data that is written out using this pattern. It is simply an option to use in case it is necessary to find out where xdd has written to.

Finally, there is an option to re-align the internal data buffer within memory if necessary. This is used more for testing computer system memory performance rather than I/O performance of a target and therefore may have limited usefulness. It is only mentioned here for completeness.

4.10 Processor Allocation and Priority Assignment

On multi-processor systems it is possible to assign xdd threads to specific processors. This is accomplished with the `-processor`, `-singleproc`, and `-roundrobin` options. The `-processor` option allows the explicit assignment of a processor to a specific xdd thread.

The `-singleproc` option will assign all xdd threads to a single processor specified as an argument to this option. The `-roundrobin` option will distribute the xdd threads across M processors where M is the number of processors. M should be less than or equal to the number of processors in the system. The processor-numbering scheme used is 0 to N-1 where N is the number of processors in the system. For example, if there are five xdd threads running on a computer with eight processors, then the round robin processor assignment will assign threads 0 thru 4 on processors 0 thru 4. However, if there were only two processors on the computer, then xdd threads 0, 2, and 4 will be assigned to processor 0 and threads 1 and 3 will be assigned to processor 1.

The priority of each thread defaults to the “normal” priority on the system. It is possible to increase the priority to a maximum level by using the “`-maxpri`” option. Maximizing the runtime priority of the xdd threads decreases the effects on the I/O performance of other processes running on the system. It is also possible to lock the xdd process and I/O buffers using the “`-plock`” option. This is done to prevent the xdd process or any of its I/O buffers from being paged or swapped out of the system. The “`-maxall`” option is a shortcut for specifying both “`-maxpri`” and “`-plock`.”

4.11 I/O Time Stamping

While running, each xdd thread has the option to enter time stamp information into a table that is later written to a file for further analysis. Each I/O operation is time stamped before the operation starts and just after the operation ends. The time stamp table contains all the information necessary to understand when I/O operations started, ended, the block location being accessed, and the amount of data transferred. The time stamps themselves are taken from the system's high-resolution timer and re-normalized in units of picoseconds so that this data can easily be correlated with time stamp data from the other associated xdd output files.

The format of the Time Stamp binary output file contains a header followed by the time stamp data.

In addition to a binary output file, the time stamp table information can be dumped in ASCII readable text. There are several options for ASCII output including a summarized and a detailed output specified by the “`-output summary`” and “`-output detailed`” options. Appendix A includes examples of the summarized and detailed outputs of the Time Stamp table. The time stamp ASCII output file names have a “.csv” file extension so that it can more easily be read by a spreadsheet program such as Excel®.

5 Running xdd Programs

Xdd programs include xdd itself, the timeserver program, and the gettime program. (New options are highlighted in blue). Example command lines are given in Appendix B and are also contained in a file called tests.txt in the distribution bin directory.

5.1 Xdd Command Line Arguments Synopsis

```

xdd
-op [target <target#>] read|write
-rwratio [target <target#>] %read
-setup setup_filename
-targetdir [target <target#>] directoryname
-targets N filename1 filename2 ... filenameN
-deletefile [target <target#>]
-processor processor_number target_number
-roundrobin #
-singleproc processor_number
-nobarrier
-blocksize [target <target#>] number_of_bytes_per_block
-reqsize [target <target#>] number_of_blocks
-passes number_of_passes
-delay seconds
-deskew
-randomize [target <target#>]
-kbytes [target <target#>] number_of_kilobytes_to_transfer
-mbytes [target <target#>] number_of_megabytes_to_transfer
-numreqs [target <target#>] number_of_requests_to_perform
-timelimit [target <target#>] seconds_per_pass
-runtime seconds
-startoffset [target <target#>] starting_block_number
-passoffset [target <target#>] offset_in_blocks
-targetoffset [target <target#>] offset_in_blocks
-queuedepth [target <target#>] number_of_commands_per_target
-datapattern [target <target#>] character_pattern , “random” , “sequenced”
-throttle [target <target#>]
    ops operations/sec
    bw megabytes/second
-sharedmemory [target <target#>]
-align [target <target#>] alignment_value_in_bytes
-lockstep <master_target> <slave_target>
    <when> <howlong>
    <what> <howmuch>
    <startup> <completion>
-syncio number
-syncwrite [target <target#>] number
-preallocate [target <target#>] number_of_blocks
-dio [target <target#>]
-seek [target <target#>]
    save filename
    load filename
    disthist #
    seekhist #

```

random
range #
stagger
interleave #
seed #
none

-id *commandline* - or - "*id_string*"
-verbose
-version
-report_threshold [*target #*] <#. #>
-timerinfo
-output *filename*
-errout *filename*
-csvout *filename*
-combined *filename*
-maxerrors *number_of_errors*
-maxpri
-proclock
-noproclock
-nomemlock
-maxall
-minall
-sgio
-timeserver *hostname*
-port #
-starttime *#seconds*
-ts [*target <target#>*]
 output *output_filename_prefix*
 summary
 detailed
 normalize
 summary
 wrap
 oneshot
 size #
 triggertime *#seconds*
 triggerop *operation#*
 append
 dump *dump_filename_prefix*
-heartbeat *#seconds*
-verify [*target <target#>*]
 location
 contents
-fullhelp

Where:

-op specifies the operation to perform: either **read** or **write** may be specified.

-rwratio (or **-rw**) specifies the percentage of operations that should be read operations. The remaining operations will be write operations. For example, specifying a value of 30.2 (i.e. **-rwratio 30.2**) will cause 30.2% of the total number of operations being performed on the target to be *read* operations and 69.8% of the operations will be *write* operations. Values less than 0 or greater than 100 are not allowed.

-setup specifies a file that contains commonly used xdd options. This file is read in and the options contained within the file will be inserted into the command line.

-targetdir specifies the name of the directory to be pre-pended to the target(s). For example, specifying a parent directory of */dev/rdisk/* (i.e. **-targetdir /dev/rdisk/**) and target names of *"dks1d2s0 dks7d3s0"* will cause I/O to be directed to */dev/rdisk/dks1d2s0* and */dev/rdisk/dks7d3s0* respectively. It is important to remember to put the trailing slash ("/") at the end of the parent directory name.

-targets must first specify the number of targets (*N*) followed by the target device names or file names of each of the *N* targets. For example, **"-targets 2 dks1d2s0 dks7d3s0"** will perform I/O to the target devices *dks1d2s0* and *dks7d3s0* respectively. In the output reports, these two targets will also be identified as targets 0 and 1 respectively.

-deletefile will cause the target file to be deleted and re-created between passes.

-processor *procesor_number target_number* This option allows an xdd thread for a specific target to run on a specific processor. The xdd thread for target *target_number* is assigned to processor *processor_number*.

-roundrobin *#* will assign successive xdd threads to processors in a "roundrobin" fashion across *#* processors.

-singleproc *processor_number* will assign all xdd threads to the specified processor.

-nobarrier will cause the passes to run asynchronously.

-blocksize specifies the number of bytes per block. This defaults to 1024 bytes per block. Block sizes must be powers of 2 or the results are unpredictable.

-reqsize specifies the number of *blocks* to transfer where the size of the block is specified by the **-blocksize** parameter.

-passes *#* where *#* the number of passes to perform.

-delay *#* where *#* is the number seconds to delay between passes.

-deskew will adjust the performance calculations to deskew the results. See section on Deskew for more information.

-randomize will cause the seek locations to be randomized between passes.

-kbytes specifies the integer number of kilobytes to transfer on each pass. In this case, one kilobyte is equal to 1024 bytes. If the **-numreqs** option is also specified, **-numreqs** takes precedence.

-mbytes specifies the integer number of megabytes to transfer on each pass. In this case, one megabyte is equal to 1024*1024 or 1048576 bytes. If the **-numreqs** option is also specified, **-numreqs** takes precedence.

-numreqs specifies the integer number of “reqsize” requests to perform on each pass. If the **-mbytes** or **-kbytes** option is also specified, **-numreqs** takes precedence.

-timelimit # will impose a time limit of # seconds on each pass. This value must be a positive integer.

-runtime #seconds will cause xdd to terminate completely after it has run for the specified number of seconds. It is important to note that if the *timestamp* option is also specified the timestamp buffer *wrap* option is automatically enabled so that the internal timestamp buffer is not overrun.

-startoffset specifies the starting block number. This defaults to block 0. The value must be a positive integer.

-passoffset # where # is the number of blocks to offset for each pass.

-targetoffset specifies the offset in blocks that is used by each xdd process to determine their respective starting locations. The purpose of this is to be able to run multiple xdd threads on a single device but to have each thread start at a different location. (*Note: This option was formerly called -procoffset*)

-queuedepth specifies the number of commands to send to each target at one time. This exercises the command queuing capabilities of a storage device or, if I/O is to a file, it will mimic parallel I/O – multiple readers/writers to a single file.

-datapattern specifies either a single byte data pattern character (or a random pattern if the word “**random**” is specified) that is used as the data in the data buffer that is written to a target device or file. If the word “**sequenced**” is used as the data pattern, then the data pattern will be sequential 64-bit integers starting with the current block address times the size of a 64-bit integer. It should be noted that writing this sequenced pattern to the device will result in additional CPU overhead that may affect overall performance. Similarly, for read operations, if the “**sequenced**” data pattern is specified, the data is checked to see if the data read is in fact what was expected. (*Note: This option was formerly called -patttern*)

-throttle specifies the I/O Operations per second (**ops**) or bandwidth (**bw**) limit for the target(s) depending on which of the two parameters are specified. Valid values are positive real numbers greater than 0.000. The parameters **ops** and **bw** are mutually exclusive and the last one specified for a targets takes precedence. Example usages:

- “-throttle ops 7.8” will limit all targets to running at 7.8 I/O operations per second.
- “-throttle bw 87.2” will limit all targets to running at 87.2 megabytes per second.
- “-throttle target 2 ops 7.8” will limit only target 2 to running at 7.8 I/O operations per second and all other targets will have no throttle limit unless specified with another – throttle option.

-sharedmemory tells xdd to use a shared memory segment (via shmget/shmat) for the I/O buffer rather than using the normal valloc()/malloc() system calls.

-align *memory_alignment_value_in_bytes* will cause the internal memory address alignment of the I/O buffer to be offset by the number of bytes specified as *memory_alignment_value_in_bytes*. The I/O buffer is normally page aligned.

-lockstep <master_target> <slave_target> <when> <howlong> <what> <howmuch> <startup> <completion>

Where

"master_target" is the target that tells the slave when to do something.

"slave_target" is the target that responds to requests from the master.

"when" specifies when the master should tell the slave to do something.

The word *"when"* should be replaced with the word:

"time"

"op"

"percent"

"mbytes"

"kbytes"

"howlong" is either the number of seconds, number of operations, ...etc.

- The interval time in seconds (a floating point number) between task requests from the master to the slave. i.e. if this number were 2.3 then at every 2.3-second interval the master would request the slave to perform its task.
- The operation number that defines the interval on which the master will request the slave to perform its task. i.e. if the operation number is set to 8 then upon completion of every 8 (master) operations, the master will request the slave to perform its task.
- The percentage of operations that must be completed by the master before requesting the slave to perform a task
- The number of megabytes (1024*1024 bytes) or the number of kilobytes (1024 bytes)

"what" is the type of task the slave should perform each time it is requested to perform a task by the master. The word *"what"* should be replaced by:

"time"

"op"

"mbytes"

"kbytes"

"howmuch" is either the number of seconds, number of operations, ...etc.

- The amount of time in seconds (a floating point number) the slave should run before pausing and waiting for further requests from the master.
- The number of operations the slave should perform before pausing and waiting for further requests from the master.
- The number of megabytes (1024*1024 bytes) or the number of kilobytes (1024 bytes) the slave should transfer before pausing and waiting for further requests from the master

"startup" is either **"wait"** or **"run"** depending on whether the slave should start running upon invocation and perform a single task or if it should simply wait for the master to request it to perform its first task.

"completion" - in the event that the master finishes before the slave, then the slave will have the option to complete all of its remaining operations or to just stop at this point. This should be specified as either **"complete"** or **"stop"**.

-syncio number will cause each of the xdd I/O threads to synchronize every n^{th} I/O operation where N is specified as *"number"*. .

-syncwrite will cause each of the xdd I/O threads to synchronize write operations at the end of each pass flushing all data to the physical media.

-preallocate # will preallocate # 1024-byte blocks. This is used when writing to a target that is a regular file. This option has no effect when reading or when the target is a real device.

-seek specifies a number of parameters that are specific to the access pattern used on each target. The default access pattern is purely sequential. These parameters are:

- **save filename** - will save the list of seek locations in an ASCII text file specified by *filename*. This file can later be used by the **-seek load** option. See Appendix B for the format of this file.
- **load filename** - will load the list of seek locations from an ASCII text file specified by *filename*.
- **range #blocks** - will specify the range in blocks over which to perform random seek operations.
- **random** will generate a random list of locations to access over the "range"
- **seed seed_value** specifies a seed value to use when generating random locations
- **stagger** will stagger the requests sequentially and evenly over the "range"
- **interleave factor** where "factor" is the interleave factor to used (see section on parallel I/O and seek interleave).
- **none** will cause xdd to continuously read the *starting* block on a target until for a total of -mbytes or -numreqs of data transfers completes.
- **disthist #categories** - will display an ASCII readable histogram of the seek "distances".
- **seekhist #categories** - will display an ASCII readable histogram of the seek "locations".

-id allows the user to specify an ASCII string to be displayed in the output in order to identify the run. If the word **commandline** is used as the character string then the input command line is used as the id. Multiple instances of the **-id** option will concatenate each specified id to the previous one.

-dio will turn on the DIRECT IO option when accessing a regular file. This option cannot be used when accessing special device files. Certain blocksize, request size, and alignment restrictions apply and will cause problems if the wrong combination of block size, request size, and offsets are chosen.

-verbose will display performance information for each pass.

-version will display the version number for this xdd program.

-report_threshold will report the byte location of the operation that exceeded the specified threshold time.

-timerinfo will display internal timer information.

-output *filename* will send all the results output to the file specified by *filename*. This does not include error messages. See the **-errout** option for more information on redirecting error output.

-errout *filename* will send all the error message output to the file specified by *filename*. This does not include normal results output. See the **-output** option for more information on redirecting results output.

-csvout *filename* will send all the results output to the file specified by *filename* similar to the **-output** option. The difference between **-csv** output and the normal output is that this is a Comma Separated Values file that is directly importable into MS Excel. This does not include error messages. See the **-errout** option for more information on redirecting error output.

-combined *filename* will append just the “Combined” results output to the file specified by *filename*. This allows for collecting the Combined performance data from multiple runs in a single file. This does not include error messages. See the **-errout** option for more information on redirecting error output.

-maxerrors *number_of_errors* will limit the number of errors to *number_of_errors* so as not to clutter up the output with endless lines of errors.

-maxpri will set the priority of all xdd threads to maximum. NOTE: Use of this option can result in system hangs due to deadlocks with other system functions.

-proclock will lock the xdd process in memory so that it cannot be paged or swapped out. This is useful on a crowded system.

-noproclock will prevent xdd process from being locked in memory so that it can be paged or swapped out.

-nomemlock will prevent the xdd memory buffers from being locked so that they can be paged or swapped out.

-maxall will set the **-maxpri** and **-plock** options.

-minall will set the **-noproclock** and **-nomemlock** options

-sgio will perform I/O operations to the specified target using the SCSI Generic protocol rather than the normal read/write system calls. This is only valid on Linux systems and is used to provide raw-like access to a device.

-timeserver *hostname* use the host computer system specified by *hostname* as the ‘master clock’ for all timing information. See the **-starttime** option for additional information.

-port *port_number* will use the specified port number to connect to the time server. See the **-timeserver** option for more information.

-starttime *global_time* will cause the target I/O threads to all start at the specified time. The global time is the time value returned by the time server and is consistent for all systems using the time server. See the **-timeserver** for more information.

-ts specifies a number of parameters that are specific to the time stamping capabilities. These are:

- **summary** will generate a summary of all the I/O operations in the time stamp trace (see figure 4 for details).
- **detailed** will generate a detailed report of each I/O operation in the time stamp trace and a summary report. It is recommended that the **output** filename be specified when using **detailed** reporting since the trace data can be exceedingly verbose (see figure 4 for details).
- **normalize** will cause all the time stamp values to be normalized to the global clock. This is useful when running xdd on multiple machines so that the events in the time stamp file can be correlated in time.
- **output** *output_filename* will cause the detailed and/or summary reports to be written to a file of “*output_filename*”. The output defaults to standard out.
- **append** will cause the detailed and/or summary reports to be appended to the specified output file.
- **dump** *dump_filename_prefix* will dump a binary file that contains all the time stamp data. The following structure contains the format of that file.
- **wrap** will cause the internal timestamp buffer to *wrap* around to the beginning of the buffer if/when the end is reached. This is used in conjunction with the “size” option described below. The reason for wrapping the timestamp buffer is to essentially capture the most recent I/O operations in a timestamp buffer that is smaller than required for the number of operations being processed by xdd.
- **oneshot** specifies that time stamping will stop once the internal timestamp buffer is full.
- **size** *#* specifies the size of the internal timestamp buffer in terms of the number of operations that will fit into the buffer. If the size specified is smaller than the number of operations to be

performed, the timestamp buffer will be “*wrapped*” after the last timestamp buffer entry is used.

- **triggertime** *#seconds* will cause timestamping to start at the specified time as measured in *global-time seconds*.
- **triggerop** *operation#* will cause timestamping to start when the specified operation number is reached.

-heartbeat *#seconds* will display the current operation and pass number for each target I/O thread every N seconds where N is specified as *#seconds*.

-verify specifies a number of parameters that are specific to the time stamping capabilities. These are:

- **immediate** will verify all data written at then end of each write operation rather than on a per-pass basis.
- **compare** will compare the data read with the expected data pattern for data integrity.
- **Report** *report_level* will report errors at various levels. Level 1 will simply report that there was a difference, level 2 will report the offset into the run and the length of the difference, and level 3 will display the data read and the expected data for the length of the miscompare.
- **Action** *action* – specifies maximum error count before the verify operation is terminated.

-fullhelp will display a list of these options *and* a one line explanation of each.

5.2 Target-specific options

Many of the options can be target-specific. These options are listed with the optional [**target** <*target#*>] arguments that immediately follow the option name. The word “**target**” indicates that the associated option is to be set for the target with a target number of <*target#*>. Target numbers are from 0 to N-1 where N is the number of targets being accessed in this run. For example, specifying “**-op target 3 read**” will cause target 3 to issue read operations regardless of what the other targets are doing. This capability is useful for tailoring the behavior of each target in a run to meet specific I/O requirements. For example, it is possible to have a single xdd run accessing several targets using different throttle values so that one target does not overwhelm the others.

It is important to note that the options are evaluated from left to right on the command line or from top to bottom in a setup file and that latter options (to the right) take precedence over prior options (to the left). Take the following command line for example:

```
xdd -op read -op target 1 write -targets 3 s1 s2 s3 -reqsize 1 -numreqs 7
```

The “**-op read**” option will cause all three targets (s1, s2, and s3) to perform *read* operations. However, the “**-op target 1 write**” option will override the *read* operation for target number 1 (target s2) causing it to perform *write* operations.

5.3 Lockstep Operations

Lockstep operations are used to simulate the I/O interaction between multiple applications running on a system. For example, one application may be creating files that a second application will use

just after their creation – aka the “read-after-write” scenario where a file is being ingested from a source, written to a file and another application is reading blocks just after they are written in order to process the data as quickly as possible. An example of this is as follows:

```
xdd -targets 2 /dev/disk1 /dev/disk1 \  
    -op target 0 write \  
    -op target 1 read \  
    -reqsize 1024 -mbytes 2048 \  
    -lockstep 0 1 op 1 op 1 wait complete
```

This will cause target 0 to write a block and then signal target 1 to start reading. Since they are the same target starting at the same locations, target 1 will be exactly 1 operation behind target 0 all the time. Essentially this tells target 0 to do 1 operation, signal target 1 which will do 1 operation and then wait for target 0 to signal it again and so on.

The current version of xdd only supports lockstep operations on a single computer system. The next version of xdd will enable lockstep operations across physically separate computer systems.

5.4 Timeserver and gettime

Command synopsis of the timeserver command is

```
timeserver  
    -port #
```

Where

-port # specifies the port number to use for the time serving function.

The gettime command is used to obtain the global clock value from the timeserver computer. The command synopsis is:

```
gettime  
    -timeserver hostname  
    -port #  
    -add seconds  
    -bounce times  
    -verbose
```

Where:

-timeserver *hostname* specifies the name of the computer running the timeserver. This may be either a host name or an IP address.

-port # specifies the port number to use when contacting the timeserver.

-add *seconds* specifies number of seconds to add to the global time that is displayed as the output of this program.

-bounce *times* specifies the number of times to ping the timeserver in order to get a minimum round trip time. The higher the bounce count, the more accurate the global time will be.

-verbose will cause gettime to display more information than simply the global clock value.

5.5 Deskew

De-skewing the performance results becomes particularly necessary when testing a large number of targets on a single system. The reason is that when all devices are started there can be a significant amount of time lag between the time the first targets starts and the last target starts its data transfer. Furthermore, there may be a long lag time between the time the first target finishes and the last target finishes particularly if there is one device that is unusually slow. This causes the overall results to be skewed and does not represent the true “cross sectional” bandwidth of the system as a whole.

The deskew option will report the bandwidth during the time in which all the targets are transferring data. This is effectively from the time the last target starts to the time the first target finishes. During that time period all targets are transferring data. This “deskewed” bandwidth is a more accurate representation of the bandwidth of the system.

6 Runtime Hints

6.1 Windows

The Windows physical drive is the equivalent of the raw volume in a Unix environment. The physical drive can be specified by using the following xdd command line:

```
xdd -op read -targets 1 \\.\PhysicalDrive0 -reqsize 1 -mbytes 1
```

xdd may complain about an incorrect function but that message can be ignored. It will be fixed in a future release. If a *cygwin* shell window is being used then the syntax must include extra “\” characters like so:

```
xdd -op read -targets 1 \\\\.\\PhysicalDrive0 -reqsize 1 -mbytes 1
```

6.2 IRIX

IRIX is reasonably well behaved and has no known idiosyncrasies.

6.3 AIX

AIX can use very large page sizes (on the order of 16Mbytes per page) but it is still necessary to “pin” all the pages in memory for every I/O operation. For most I/O performance testing this is not a problem unless the number of targets gets large and/or the bandwidth is very high. In this case the page pinning operation takes a significant percentage of the I/O time and can have a negative affect on the results.

To avoid the page pinning penalty, the “-sharedmemory” operation can be used to allocate the I/O buffer in a shared memory segment which causes the system to bypass the page pinning since shared memory pages are already pinned by default.

6.4 Solaris

There is an issue with running xdd on Solaris that involves the default number of semaphores that Solaris allows each process to use. It is normally too low for xdd to run and xdd generates errors that say things like

“cannot allocate barrier for ..., not enough space”. The following parameters can be put in /etc/system on a solaris system and a reboot will take care of this problem:

```
set semsys:seminfo_semmni=4096
set semsys:seminfo_semmns=8192
set semsys:seminfo_semmap=4098
set shmsys:shminfo_shmmni=512
set shmsys:shminfo_shmseg=32
```

Another change that is sometimes necessary is

```
set maxphys = 0x1000000
```

This will change the maximum request size to 16 MB that is useful when doing really large request sizes for raw device tests.

These settings may be a little high and can be adjusted to meet the local system constraints.

6.5 Linux

Raw I/O is only supported using the “raw” command to create the “raw” device associates and then doing I/O to those devices. The Linux “raw” devices are only an approximation of real raw devices on most other Unix systems. It should also be mentioned that request sizes larger than about 256Kbytes may not actually be issued to the device but rather be split into multiple requests by the “raw” device driver.

In this release of xdd there is no support for assigning targets to processors in Linux. Apparently this functionality is only supported in Linux kernels 2.5.8 and above. This will be supported in the next release of xdd.

6.6 HPUX

None.

6.7 Tru64

None.

7 Output and Reports

7.1 Reporting Options

The `-verbose` option will display performance information for each pass within an xdd run along with per-target totals and combined averages. There is also an “id” option (`-id`) that will display a specified string of identification information along with the normal output of the run. If the id string is specified to be “commandline” then the entire xdd command line plus all the options will be display. This is useful when running xdd from a shell script to get more qualitative information about the run into the output file.

Since xdd I/O operations can fail due to device failures or option specifications it is possible to generate tremendous numbers of error messages. To avoid filling up output files with too many error messages, the `-maxerrors` option can be used to limit the number of error messages to display to some small, finite number. Once this number of errors has been reached, I/O operations to that target are halted for that pass.

The output of xdd comes in several sections. The **first section describes the state of options selected for the run that apply to all the targets being tested.** The **second section of output describes the options that apply to each of the individual targets.** The **third section contains information about the system that xdd is being run on.** The **fourth section describes the rate information.**

```
IOIOIOIOIOIOIOIOIOIOIOIOI XDD IOIOIOIOIOIOIOIOIOIOIOIOI
xdd - Version, 6.4, I/O Performance Inc.
Run Time, Wed Dec 15 21:24:13 2005

ID for this run, 'Test_using_4096K_Ops'
Maximum Process Priority, enabled
Passes, 3
Pass Delay in seconds, 0
Maximum Error Threshold, 3
I/O Synchronization, 0
Target Offset, 0
Total run-time limit in seconds, 0
Output file name, ./results/txt/Test.4096.txt
CSV output file name, ./results/csv/Test.4096.csv
Error output file name, ./results/txt/errors.4096.txt
Combined output file name, ./results/csv/Test.combined.csv
Pass seek randomization, disabled
File write synchronization, disabled
Pass synchronization barriers, enabled
Timeserver hostname, mp-1
Timeserver port number, 2000
Global start time, 186719
Number of Targets, 2
Number of I/O Threads, 2

Computer Name, testsgi, User Name, george
OS release and version, IRIX64 6.5 04100803
Machine hardware type, IP35
Number of processors on this system, 16
Page size in bytes, 16384
Number of physical pages, 524288
```

Megabytes of physical memory, 8192
 Seconds before starting, 9

```

Target information,
  target[0] Q[0], /scratch/testfile1
    target directory, "/"
    Processor, 0
    Read/write ratio, 100.00, 0.00
    Throttle in MB/sec, 0.00
    Per-pass time limit in seconds, 30
    Blocksize in bytes, 1024
    Request size, 4096, blocks, 4194304, bytes
    Start offset, 0
    Number of MegaBytes, 8192
    Pass Offset in blocks, 0
    I/O memory buffer alignment in bytes, 16384
    Data pattern in buffer, 0x0
    Direct I/O, disabled
    Seek pattern, sequential
    Seek range, 1048576
    Preallocation, 0
    Queue Depth, 1
    Timestamping, enabled for DETAILED SUMMARY
    Timestamp ASCII output file name, ./results/ts/Test.4096.target.0.csv
    Delete file, disabled
  target[1] Q[0], /scratch/testfile2
    target directory, "/"
    Processor, 0
    Read/write ratio, 100.00, 0.00
    Throttle in MB/sec, 0.00
    Per-pass time limit in seconds, 30
    Blocksize in bytes, 1024
    Request size, 4096, blocks, 4194304, bytes
    Start offset, 0
    Number of MegaBytes, 8192
    Pass Offset in blocks, 0
    I/O memory buffer alignment in bytes, 16384
    Data pattern in buffer, 0x0
    Direct I/O, disabled
    Seek pattern, sequential
    Seek range, 1048576
    Preallocation, 0
    Queue Depth, 1
    Timestamping, enabled for DETAILED SUMMARY
    Timestamp ASCII output file name, ./results/ts/Test.4096.target.1.csv
    Delete file, disabled

```

	T	Q	Bytes	Ops	Time	Rate	IOPS	Latency	%CPU	I/Otype	ReqSize
T PASS0001	1	1	1333788672	318	30.01	44.44	10.60	0.0944	1.27	read	4096
T PASS0001	0	1	2864709632	683	30.19	94.89	22.62	0.0442	1.46	read	4096
T PASS0002	0	1	2944401408	702	30.03	98.06	23.38	0.0428	1.30	read	4096
T PASS0002	1	1	1283457024	306	30.03	42.74	10.19	0.0981	1.32	read	4096
T PASS0003	0	1	2583691264	616	30.02	86.06	20.52	0.0487	1.20	read	4096
T Average	0	1	8392802304	2001	90.24	93.01	22.17	0.0451	1.32	read	4096
T PASS0003	1	1	1296039936	309	30.14	43.01	10.25	0.0975	1.19	read	4096
T Average	1	1	3913285632	933	90.18	43.39	10.35	0.0967	1.26	read	4096
Combined	2	2	12306087936	2934	90.24	136.37	32.51	0.0308	2.58	read	4096

The rate information output of xdd is a line of text with the following format:

T Q Bytes Ops Time Rate IOPS Latency %CPU I/Otype Reqsize

Where these fields have the following meanings:

T This is the target number, relative to 0 (zero). A list of the targets and the associated numbers precedes this line of output.

Q This is the Queue Thread number, relative to 0 (zero) for this target. This is only used if the `-queuedepth` option is specified and `-qthreadinfo` is requested.

Bytes Is the total number of bytes that were transferred during the xdd run for this target.

Ops is the total number of read or write operations performed on this target.

Time is the number of seconds that elapsed for the current pass in a multi-pass run.

Rate is the average I/O rate in Mega Bytes per Second where one Mega Byte is 1,000,000 bytes.

IOPS is the Average number of I/O operations per second for this pass in a multi-pas run.

Latency is the Average time it takes to perform each operation.

%CPU is the percentage of the CPU that was used by this xdd thread.

IOType is either read, write, or mixed.

Reqsize is the *average* request size in block-size blocks that was used for that target

If the **-verbose** option is specified then each line has a pass number associated with it and the final output lines report overall averaged values for the Average Rate and Elapsed time. For a Multi-Target run, the sum total of all the targets is presented as the **Combined** average. For example, if *two* targets are being tested and *each* target performs at an average of 75MB/sec, the Combined average is 150 MB/sec.

7.2 What the numbers really mean

There are several performance values reported. These include the per-pass target results, the individual queue results, the target averages over all passes, and the combined average of all targets over all passes.

8 Performance Tuning Hints

This section describes various hints about performance tuning.

8.1 Caches and write performance

When writing to a storage device whether it is a single disk drive or a disk array it is important to know the status of the caches on each of the devices that have cache. For maximum performance for write operations, it is necessary to enable the write caches on a disk array controller as well as the write caches on the disk drives themselves.

8.2 Fibre Channel Frame Sizes

Fibre Channel host bus adapters (HBA), switches, and target devices all have frame sizes defined and negotiated when any two FC devices are connected together. For maximum bandwidth performance it is important to make certain that the FC frame size is set to 2048-bytes. For maximum transaction performance for small transaction sizes (i.e. around 512 bytes per transaction) a smaller frame size of 512 or 1024 bytes can be used.

9 Examples

The following is a list of examples on how to run xdd.

Example 1.

```
xdd -op read -targets 1 /dev/rdisk/dsk10d2s0 -reqsize 128
-mbytes 64 -passes 3 -verbose
```

This is a very basic test that will **read** sequentially from target device **/dev/rdisk/dks10d2s0** starting at block 0 using a fixed request size of **128** blocks until it has read **64** MegaBytes ($64 * 1024 * 1024$ bytes). It will do this **3** times and display performance information for each pass. The default block size is 1024 bytes per block so the request size in bytes is 128 Kbytes ($128 * 1024$ bytes). Please note that all these options need to be on a single command line unless they are in the setup file where they can be on separate lines.

Example 2.

```
xdd -op write -targets 2 /raid/BIGFILE1 /raid/BIGFILE2
    -blocksize 512 -reqsize 128 -mbytes 64 -verbose
    -pass count 3 -timelimit 10
```

This test will **write** sequentially from **2** target files **/raid/BIGFILE1** and **/raid/BIGFILE2** starting at the beginning of each file using a fixed request size of **128** blocks of **512** bytes per block until it has read **64** MegaBytes ($64 * 1024 * 1024$ bytes) – *or* – until it has reached a time limit of **10** seconds at which time it will end the current pass and proceed to the next pass. It will do this **3** times and display performance information for each pass. The *combined* performance of both devices is calculated and displayed at the end of the run.

Example 3.

```
xdd -op write -targets 2 /dev/rdisk/dks10d2s0
    /dev/rdisk/dks10d2s0
    -setup xdd.setup -ts detailed -ts output ts.write
```

This test that will **read** sequentially from **2** targets that are actually a single device: **/dev/rdisk/dks10d2s0**. The request size of **128** blocks at **2048** bytes per block, read limit of **4096** MegaBytes ($4096 * 1024 * 1024$ bytes), the time limit of **10** seconds for each pass, verbose output, and pass count of **3** are all specified in the **xdd.setup** file which looks like so:

```
-blocksize 2048
-reqsize 128
-mbytes 4096
-verbose
-passes 3 -timelimit 10
```

The time stamp option is also used in this example to dump an ASCII output file called **ts.write**. It should be noted that these time stamp file names are appended with a **t#** where **#** is the number of the target that belongs to the particular time stamp file. In this example, since there are two targets, the time stamp files will be **ts.write.t0** and **ts.write.t1**.

Example 4.

```
xdd -op read -targets 1 /dev/rdisk/dsk10d2s0 -reqsize 8
    -mbytes 16 -passes 3 -verbose -seek random -seek range
    4000000
```

This is a very basic *random I/O* test that will **read** from target device **/dev/rdisk/dks10d2s0** starting at some random location using a fixed request size of **8** blocks until it has read **16** MegaBytes ($16 * 1024 * 1024$ bytes). It will do this **3** times and display performance information for each pass. The default block size is 1024 bytes per block so the request size in bytes is 8 Kbytes ($8 * 1024$ bytes). The number of requests that need to be generated to read 16 MegaBytes is 8192

byte chunks is 2048. Since this is a purely random I/O pattern, these 2048 requests are distributed over a range of 4,000,000 blocks (again 1024 bytes per block). This is useful in constraining the area over which the random locations are chosen from. The same seek locations are used for each pass in order to generate reproducible results. In fact, upon each invocation of xdd using the same parameters, the same random locations are generated each time. This allows the user to change the disk or starting offset or some such thing and observe the effects. The random locations may be changed from pass to pass within an xdd run by using the **"-randomize"** option in which case a new set of locations is generated for each pass. Furthermore, the random locations may be changed from run to run using the **-seek seed** option to specify a different random number generation seed value for each invocation of xdd.

9.1 Example Display Output

The command line used to generate this output was

```
C:\Users\tmr\Program Source Files\xdd\xdd62c\Debug>xdd -op write -
targets 2 targetfile1 targetfile2 -reqsize 64 -numreqs 10 -verbose
-dio -maxpri -passes 3 -cvs csvoutput.csv -output output.txt -
errout erroroutput.txt -ts output timestamp.csv -ts detailed -
starttime 355390 -timeserver gigabits -syncwrite -deletetfile
```

The output file called **output.txt** is:

```
IOIOIOIOIOIOIOIOIOIOIOI XDD IOIOIOIOIOIOIOIOIOIOIOIOI
xdd - Version, 6.4.121505.1533MST, I/O Performance Inc.
Run Time, Wed Dec 15 21:24:13 2005

ID for this run, 'Test_using_4096K_Ops'
Maximum Process Priority, enabled
Passes, 3
Pass Delay in seconds, 0
Maximum Error Threshold, 3
I/O Synchronization, 0
Target Offset, 0
Total run-time limit in seconds, 0
Output file name, ./results/txt/Test.4096.txt
CSV output file name, ./results/csv/Test.4096.csv
Error output file name, ./results/txt/errors.4096.txt
Combined output file name, ./results/csv/Test.combined.csv
Pass seek randomization, disabled
File write synchronization, disabled
Pass synchronization barriers, enabled
Timeserver hostname, mp-1
Timeserver port number, 2000
Global start time, 186719
Number of Targets, 2
Number of I/O Threads, 2

Computer Name, testsgi, User Name, george
OS release and version, IRIX64 6.5 04100803
Machine hardware type, IP35
Number of processors on this system, 12
Page size in bytes, 16384
Number of physical pages, 524288
Megabytes of physical memory, 8192
Seconds before starting, 9

Target information,
  target[0] Q[0], /scratch/testfile1
    target directory, "/"
    Processor, 0
    Read/write ratio, 100.00, 0.00
    Throttle in MB/sec, 0.00
    Per-pass time limit in seconds, 30
    Blocksize in bytes, 1024
```

```

Request size, 4096, blocks, 4194304, bytes
Start offset, 0
Number of MegaBytes, 8192
Pass Offset in blocks, 0
I/O memory buffer alignment in bytes, 16384
Data pattern in buffer, 0x0
Direct I/O, disabled
Seek pattern, sequential
Seek range, 1048576
Preallocation, 0
Queue Depth, 1
Timestamping, enabled for DETAILED SUMMARY
Timestamp ASCII output file name, ./results/ts/Test.4096.target.0.csv
Delete file, disabled
target[1] Q[0], /scratch/testfile2
target directory, "/"
Processor, 0
Read/write ratio, 100.00, 0.00
Throttle in MB/sec, 0.00
Per-pass time limit in seconds, 30
Blocksize in bytes, 1024
Request size, 4096, blocks, 4194304, bytes
Start offset, 0
Number of MegaBytes, 8192
Pass Offset in blocks, 0
I/O memory buffer alignment in bytes, 16384
Data pattern in buffer, 0x0
Direct I/O, disabled
Seek pattern, sequential
Seek range, 1048576
Preallocation, 0
Queue Depth, 1
Timestamping, enabled for DETAILED SUMMARY
Timestamp ASCII output file name, ./results/ts/Test.4096.target.1.csv
Delete file, disabled

```

	T	Q	Bytes	Ops	Time	Rate	IOPS	Latency	%CPU	I/O type	ReqSize
T PASS0001	1	1	1333788672	318	30.01	44.44	10.60	0.0944	1.27	read	4096
T PASS0001	0	1	2864709632	683	30.19	94.89	22.62	0.0442	1.46	read	4096
T PASS0002	0	1	2944401408	702	30.03	98.06	23.38	0.0428	1.30	read	4096
T PASS0002	1	1	1283457024	306	30.03	42.74	10.19	0.0981	1.32	read	4096
T PASS0003	0	1	2583691264	616	30.02	86.06	20.52	0.0487	1.20	read	4096
T Average	0	1	8392802304	2001	90.24	93.01	22.17	0.0451	1.32	read	4096
T PASS0003	1	1	1296039936	309	30.14	43.01	10.25	0.0975	1.19	read	4096
T Average	1	1	3913285632	933	90.18	43.39	10.35	0.0967	1.26	read	4096
Combined	2	2	12306087936	2934	90.24	136.37	32.51	0.0308	2.58	read	4096

9.2 Example Time Stamp Output

An example of the Time Stamp output file. IOIOIOIOIOIOIOIOIOIOIOIOIOIOIO XDD

IOIOIOIOIOIOIOIOIOIOIOIOIOIOIO

xdd - Version, 6.4.121505.1533MST, I/O Performance Inc.

Run Time, Wed Dec 15 21:24:13 2005

ID for this run, 'Test_using_4096K_Ops'

Maximum Process Priority, enabled

Passes, 3

Pass Delay in seconds, 0

Maximum Error Threshold, 3

I/O Synchronization, 0

Target Offset, 0

Total run-time limit in seconds, 0

Output file name, ./results/txt/Test.4096.txt

CSV output file name, ./results/csv/Test.4096.csv

Error output file name, ./results/txt/errors.4096.txt

Combined output file name, ./results/csv/Test.combined.csv

Pass seek randomization, disabled

File write synchronization, disabled

Pass synchronization barriers, enabled

Timeserver hostname, mp-1

Timeserver port number, 2000

Global start time, 186719

Number of Targets, 2

Number of I/O Threads, 2

Computer Name, testsgi, User Name, george
OS release and version, IRIX64 6.5 04100803
Machine hardware type, IP35
Number of processors on this system, 12
Page size in bytes, 16384
Number of physical pages, 524288
Megabytes of physical memory, 8192
Seconds before starting, 9

Target information,

```
target[0] Q[0], /scratch/testfile1
target directory, "/"
Processor, 0
Read/write ratio, 100.00, 0.00
Throttle in MB/sec, 0.00
Per-pass time limit in seconds, 30
Blocksize in bytes, 1024
Request size, 4096, blocks, 4194304, bytes
Start offset, 0
Number of MegaBytes, 8192
Pass Offset in blocks, 0
I/O memory buffer alignment in bytes, 16384
Data pattern in buffer, 0x0
Direct I/O, disabled
Seek pattern, sequential
Seek range, 1048576
Preallocation, 0
Queue Depth, 1
Timestamping, enabled for DETAILED SUMMARY
Timestamp ASCII output file name, ./results/ts/Test.4096.target.0.csv
Delete file, disabled

target[1] Q[0], /scratch/testfile2
target directory, "/"
Processor, 0
Read/write ratio, 100.00, 0.00
Throttle in MB/sec, 0.00
Per-pass time limit in seconds, 30
Blocksize in bytes, 1024
Request size, 4096, blocks, 4194304, bytes
Start offset, 0
Number of MegaBytes, 8192
Pass Offset in blocks, 0
I/O memory buffer alignment in bytes, 16384
Data pattern in buffer, 0x0
Direct I/O, disabled
Seek pattern, sequential
Seek range, 1048576
Preallocation, 0
Queue Depth, 1
Timestamping, enabled for DETAILED SUMMARY
Timestamp ASCII output file name, ./results/ts/Test.4096.target.1.csv
Delete file, disabled

Start of DETAILED Time Stamp Report, Number of entries, 933, PicoSeconds per clock tick, 800000, delta, -69635263638800000
Target, RWV Op, Pass, OP Number, Location, Distance, StartTS, EndTS, IO Time_ms, Relative_ms, Delta_us, Loop_ms, Inst MB/sec
1, r, 1, 0, 0, 1867190463036000000, 1867190732716000000, 26.96800, 0.00000, 0.00000, 0.00000, 155.529
1, r, 1, 1, 4096, 0, 1867190732748000000, 1867191125204000000, 39.24560, 26.97120, 3.20000, 39.24880, 106.873
1, r, 1, 2, 8192, 0, 1867191125228000000, 1867191387500000000, 26.22720, 66.21920, 2.40000, 26.22960, 159.922
1, r, 1, 3, 12288, 0, 1867191387524000000, 1867191493268000000, 10.57440, 92.44880, 2.40000, 10.57680, 396.647
1, r, 1, 4, 16384, 0, 1867191493292000000, 1867191612788000000, 11.94960, 103.02560, 2.40000, 11.95200, 351.000
1, r, 1, 5, 20480, 0, 1867191612804000000, 1867191736284000000, 12.34800, 114.97680, 1.60000, 12.34960, 339.675
1, r, 1, 6, 24576, 0, 1867191736308000000, 1867191849956000000, 11.36480, 127.32720, 2.40000, 11.36720, 369.061
1, r, 1, 7, 28672, 0, 1867191849980000000, 1867192004852000000, 15.48720, 138.69440, 2.40000, 15.48960, 270.824
1, r, 1, 8, 32768, 0, 1867192004876000000, 1867192209204000000, 20.43280, 154.18400, 2.40000, 20.43520, 205.273
1, r, 1, 9, 36864, 0, 1867192209228000000, 1867192333716000000, 12.44880, 174.61920, 2.40000, 12.45120, 336.924
...
1, r, 2, 0, 0, 1302528, 1867492386676000000, 1867494531956000000, 214.52800, 30192.36400, 181588.80000, 396.11680, 19.551
1, r, 2, 1, 4096, 0, 1867494531980000000, 1867496128364000000, 159.63600, 30406.89440, 2.40000, 159.63840, 26.274
1, r, 2, 2, 8192, 0, 1867496128364000000, 1867498886244000000, 275.78800, 30566.53280, 2.40000, 275.79040, 15.208
1, r, 2, 3, 12288, 0, 1867498886268000000, 1867500006900000000, 112.06320, 30842.32320, 2.40000, 112.06560, 37.428
1, r, 2, 4, 16384, 0, 1867500006932000000, 1867504434156000000, 442.72240, 30954.38960, 3.20000, 442.72560, 9.474
1, r, 2, 5, 20480, 0, 1867504434180000000, 1867504578820000000, 14.46400, 31397.11440, 2.40000, 14.46640, 289.982
1, r, 2, 6, 24576, 0, 1867504578844000000, 1867504856036000000, 27.71920, 31411.58080, 2.40000, 27.72160, 151.314
1, r, 2, 7, 28672, 0, 1867504856052000000, 1867505319276000000, 46.32240, 31439.30160, 1.60000, 46.32400, 90.546
1, r, 2, 8, 32768, 0, 1867505319292000000, 1867505823708000000, 50.44160, 31485.62560, 1.60000, 50.44320, 83.152
1, r, 2, 9, 36864, 0, 1867505823732000000, 1867506126044000000, 30.23120, 31536.06960, 2.40000, 30.23360, 138.741
...
1, r, 3, 0, 0, 12533376, 1867792707924000000, 1867795001996000000, 229.40720, 60224.48880, 260.00000, 229.66720, 18.283
1, r, 3, 1, 4096, 0, 1867795002020000000, 1867796257572000000, 125.55520, 60453.89840, 2.40000, 125.55760, 33.406
1, r, 3, 2, 8192, 0, 1867796257596000000, 1867797598540000000, 134.09440, 60579.45600, 2.40000, 134.09680, 31.279
1, r, 3, 3, 12288, 0, 1867797598564000000, 1867798074300000000, 47.57360, 60713.55280, 2.40000, 47.57600, 88.165
1, r, 3, 4, 16384, 0, 1867798074324000000, 1867798156300000000, 8.19760, 60761.12880, 2.40000, 8.20000, 511.650
1, r, 3, 5, 20480, 0, 1867798156324000000, 1867798286060000000, 12.97360, 60769.32880, 2.40000, 12.97600, 323.295
1, r, 3, 6, 24576, 0, 1867798286084000000, 1867798403012000000, 11.69280, 60782.30480, 2.40000, 11.69520, 358.708
1, r, 3, 7, 28672, 0, 1867798403028000000, 1867800226764000000, 182.37360, 60793.99920, 1.60000, 182.37520, 22.998
End of DETAILED Time Stamp Report
Start of SUMMARY Time Stamp Report
Average seek distance in 1024 byte blocks, 0, request size in blocks, 4096
Range: Longest seek distance in blocks, 0, shortest seek distance in blocks, 0
Average I/O time in milliseconds, 96.14740, average dead time in microseconds, 2.35193
```

I/O Time Range: Longest I/O time in milliseconds,	515.83040, shortest I/O time in milliseconds,	8.19760
Dead Time Range: Longest dead time in microseconds,	5.60000, shortest dead time in microseconds,	1.60000
End of SUMMARY Time Stamp Report		

10 Under the Hood

This section is a look at xdd program organization and data structures. This section is primarily here for the author's benefit because his brain is getting old and forgetful.

Starting with xdd6.3 the source code files have been separated out into several categories of source files. The process of moving functions around within the files is mostly complete but further changes may be made in future releases. The rationale behind this change is that some of the source files were getting too large. The categories currently defined are:

Initialization: initialization.c, parse.c

Reporting: results.c and timestamp.c

Core functions: access_patterns.c, barrier.c, global_time.c, pclk.c, read_after_write.c, ticker.c, and of course, xdd.c

Special functions: nt_compat.c and hpss.c

Other programs: timeserver: timeserver.c, gettime: gettime.c, strip: strip.c

The restructuring of the source code files did not affect the structure of the xdd program itself.

10.1 Xdd general operation

Xdd is a command-line driven program. The first thing it does upon invoking it is to parse the command line. If the command line contains the “-setup” option, it will parse the command line options from left to right up to the -setup option. It will then parse the setup file options followed by parsing the remaining command line options.

After the all the requested options have been set, the main xdd program will start all the appropriate threads one at a time. Each thread will go through it initialization phase which includes the following:

- Open the target and prepare it for access
- Generate the list of locations to access and the associated access pattern and operations
- Allocate I/O buffer
- Perform timestamp setup if necessary
- Display target-specific information as requested

After the thread has completed its setup process, it will enter a serialization barrier that will release the main xdd parent thread which will in turn start the next thread. The target thread that has just completed initialization will enter the main barrier waiting for all the other target threads to complete initialization. Once all the threads have been started, the last thread to start will enter the starting barrier and cause all the threads to be released and the fun begins.

All the threads will do their respective I/O operations for a single pass and then enter the pass barrier. The pass barrier causes threads to wait until all threads have completed a pass before starting the next pass. If this is the last pass, then upon being released from this barrier, all the threads will perform any clean-up operations and exit.

Another function of the pass barrier is to hold all the threads dormant whilst thread 0 gathers all the results information from each of the threads and generates the appropriate intermediate results and displays them if requested. This is also the case when all threads have completed all passes and thread 0 needs to process summary information as well.

10.2 The xdd buffer memory layout

Xdd uses several buffers during a normal run. All buffers are page aligned. These buffers include:

- The Read/Write I/O buffer. This buffer is simply the size of a single I/O request and there is one of these buffers per target being tested. This buffer is normally allocated from user space memory but if the “-sharedmemory” option is specified the buffer will be allocated from a shared memory segment. The use of a shared memory segment can have a significant impact on bandwidth performance in some cases.
- Access location / operation buffer (aka seeklist). This buffer is variable in size and depends on the number of I/O operations that need to be performed. There is one access location/operation buffer for each target being tested. This buffer contains the following information:
 - The block location to access
 - The operation to be performed (read or write)
 - The amount of data to transfer for that operation
 - The time to issue the I/O request
- Time stamp buffer if time stamping is enabled. This buffer can get very large because there is one time stamp entry for every I/O operation to be performed for the entire xdd run (which includes all passes). For example, if there are three targets and each target will perform 8192 operations per pass and there are 4 passes, then there will be three time stamp buffers (one for each target) with 32768 entries in each buffer (4 passes times 8192 entries per pass). Each time stamp entry contains the following information:
 - Start time stamp (64-bit)
 - End time stamp (64-bit)
 - Operation (read or write)
 - Amount of data transferred
 - Pass number
 - Operation number
- Per Thread Data Structure (ptds). This data structure contains all the information related to a single target during an xdd run. This is explained in more detail in the section on the *xdd thread structure*.

10.3 The xdd thread structure

Xdd uses POSIX threads for all processes that it controls. There are one or more threads per target. Each target will have a primary thread and potentially some number of Q threads. The Q threads are used to perform parallel, asynchronous I/O on a particular target. The concept of Q threads is essentially “aio” but since “aio” is different from system to system, it was easier to simply implement one within xdd.

Each thread (primary or Q thread) has a PTDS (per thread data structure) data structure associated with it. The PTDS contains all the information related to a thread running at any given time. This structure is passed to the various routines within xdd. The PTDS's are allocated on an as-needed basis. The layout of the PTDS structures is shown in Figure 1.

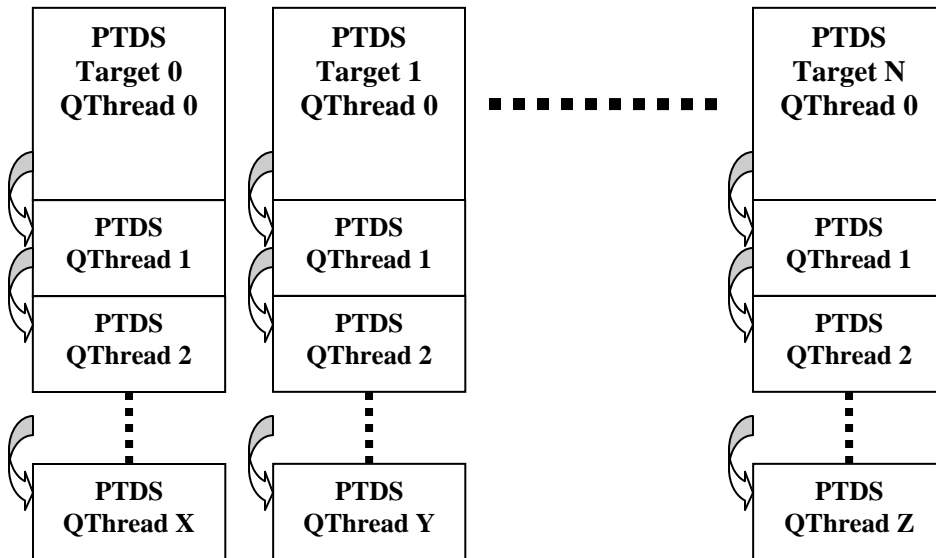


Figure 1 The Per Target Data Structure (PTDS) layout including the QThreads.

In the above diagram there are N targets each with some number of QThreads. The number of QThreads is equal to the Queue Depth specified for each target (using the `-queuedepth` option). The number of QThreads can be different for each target as is shown in this diagram but in practice the number of QThreads is the same for all targets during a run. The primary PTDS for each target is, in fact, QThread 0. If there is more than one QThread for a target, the subsequent QThreads are chained together as shown in the diagram. The last QThread in the chain is identified by the fact that its “next QThread” point is null.

10.4 Xdd barriers

Xdd uses a number of synchronization barriers in order to provide precise control and certain functionality such as lock-stepping. The primary barriers cause all the threads to begin execution at precisely the same time for each pass of a run. The barriers are implemented using mutexes and semaphores. The amount of time required to process a semaphore request is significantly less than the time scales of an I/O operation and has no measurable effect on the results.

11 The GNU Public License

Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.

59 Temple Place - Suite 330, Boston, MA 02111-1307, USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

11.1 Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

11.2 TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.
- b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.
- c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

- a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
- b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

- c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

11.3 NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

End of TERMS AND CONDITIONS

12 Acknowledgements

HPUX is a registered trademark of Hewlett-Packard, Inc.

AIX is a registered trademark of IBM Corporation.

Tru64 is a registered trademark of Compaq Computer Corporation.

IRIX is a registered trademark of SGI.

Solaris and SPARC are registered trademarks of Sun Microsystems, Inc.

Windows, Windows NT, and Windows 2000 are registered trademarks of Microsoft, Inc.

The author would like to acknowledge and thank all those people who have helped with adding various functions and bug fixes to xdd over the years.