



Contents

Foreword	xvii
Introduction	xix
Chapter 1 UNIX Evolution and Standardization	1
A Brief Walk through Time	1
How Many Versions of UNIX Are There?	3
Why Is UNIX So Successful?	3
The Early Days of UNIX	3
The Early History of the C Language	4
Research Editions of UNIX	5
AT&T's Commercial Side of UNIX	5
The Evolution of BSD UNIX	7
BSD Networking Releases	8
UNIX Goes to Court	8
The NetBSD Operating System	8
The FreeBSD Operating System	9
The OpenBSD Operating System	9
Sun Microsystems and SunOS	9
System V Release 4 and Variants	10
Novell's Entry into the UNIX Market	10
Linux and the Open Source Movement	11
UNIX Standardization	11
IEEE and POSIX	11
The X/Open Group	12
The System V Interface Definition	12
Spec 11/70 and the Single UNIX Specification	13
UNIX International and OSF	13
The Data Management Interfaces Group	14
The Large File Summit	14
Summary	15

Chapter 2	File-Based Concepts	17
	UNIX File Types	18
	File Descriptors	19
	Basic File Properties	20
	The File Mode Creation Mask	23
	Changing File Permissions	24
	Changing File Ownership	26
	Changing File Times	28
	Truncating and Removing Files	29
	Directories	30
	Special Files	31
	Symbolic Links and Hard Links	32
	Named Pipes	33
	Summary	34
Chapter 3	User File I/O	35
	Library Functions versus System Calls	35
	Which Header Files to Use?	36
	The Six Basic File Operations	37
	Duplicate File Descriptors	40
	Seeking and I/O Combined	41
	Data and Attribute Caching	42
	VxFS Caching Advisories	43
	Miscellaneous Open Options	46
	File and Record Locking	46
	Advisory Locking	47
	Mandatory Locking	51
	File Control Operations	51
	Vectored Reads and Writes	52
	Asynchronous I/O	54
	Memory Mapped Files	59
	64-Bit File Access (LFS)	65
	Sparse Files	66
	Summary	71
Chapter 4	The Standard I/O Library	73
	The FILE Structure	74
	Standard Input, Output, and Error	74
	Opening and Closing a Stream	75
	Standard I/O Library Buffering	77
	Reading and Writing to/from a Stream	79
	Seeking through the Stream	82
	Summary	84

Chapter 5	Filesystem-Based Concepts	85
	What's in a Filesystem?	85
	The Filesystem Hierarchy	86
	Disks, Slices, Partitions, and Volumes	88
	Raw and Block Devices	90
	Filesystem Switchout Commands	90
	Creating New Filesystems	92
	Mounting and Unmounting Filesystems	94
	Mount and Umount System Call Handling	98
	Mounting Filesystems Automatically	98
	Mounting Filesystems During Bootstrap	99
	Repairing Damaged Filesystems	100
	The Filesystem Debugger	101
	Per Filesystem Statistics	101
	User and Group Quotas	103
	Summary	104
 Chapter 6	 UNIX Kernel Concepts	 105
	5th to 7th Edition Internals	105
	The UNIX Filesystem	106
	Filesystem-Related Kernel Structures	107
	User Mode and Kernel Mode	107
	UNIX Process-Related Structures	109
	File Descriptors and the File Table	110
	The Inode Cache	112
	The Buffer Cache	112
	Mounting Filesystems	115
	System Call Handling	115
	Pathname Resolution	116
	Putting It All Together	117
	Opening a File	118
	Reading the File	119
	Closing the File	120
	Summary	120
 Chapter 7	 Development of the SVR4 VFS/Vnode Architecture	 121
	The Need for Change	121
	Pre-SVR3 Kernels	122
	The File System Switch	122
	Mounting Filesystems	123
	The Sun VFS/Vnode Architecture	126
	The uio Structure	129
	The VFS Layer	129
	The Vnode Operations Layer	130

Pathname Traversal	131
The Veneer Layer	132
Where to Go from Here?	133
The SVR4 VFS/Vnode Architecture.....	133
Changes to File Descriptor Management.....	133
The Virtual Filesystem Switch Table	134
Changes to the Vnode Structure and VOP Layer	135
Pathname Traversal	139
The Directory Name Lookup Cache	140
Filesystem and Virtual Memory Interactions.....	142
An Overview of the SVR4 VM Subsystem	143
Anonymous Memory	146
File I/O through the SVR4 VFS Layer.....	146
Memory-Mapped File Support in SVR4	149
Flushing Dirty Pages to Disk	152
Page-Based I/O.....	153
Adoption of the SVR4 Vnode Interface.....	153
Summary	154
 Chapter 8 Non-SVR4-Based Filesystem Architectures	155
The BSD Filesystem Architecture	155
File I/O in 4.3BSD	156
Filename Caching in 4.3BSD	157
The Introduction of Vnodes in BSD UNIX	157
VFS and Vnode Structure Differences.....	159
Digital UNIX / True64 UNIX	159
The AIX Filesystem Architecture.....	161
The Filesystem-Independent Layer of AIX.....	161
File Access in AIX.....	162
The HP-UX VFS Architecture.....	163
The HP-UX Filesystem-Independent Layer	164
The HP-UX VFS/Vnode Layer.....	164
File I/O in HP-UX	164
Filesystem Support in Minix	165
Minix Filesystem-Related Structures.....	166
File I/O in Minix	167
Pre-2.4 Linux Filesystem Support.....	168
Per-Process Linux Filesystem Structures	168
The Linux File Table.....	169
The Linux Inode Cache.....	170
Pathname Resolution	172
The Linux Directory Cache	172
The Linux Buffer Cache and File I/O	173
Linux from the 2.4 Kernel Series	174
Main Structures Used in the 2.4.x Kernel Series	175

The Linux 2.4 Directory Cache.....	175
Opening Files in Linux.....	177
Closing Files in Linux.....	178
The 2.4 Linux Buffer Cache	178
File I/O in the 2.4 Linux Kernel.....	179
Reading through the Linux Page Cache	179
Writing through the Linux Page Cache	180
Microkernel Support for UNIX Filesystems	180
High-Level Microkernel Concepts	181
The Chorus Microkernel	182
Handling Read Operations in Chorus	183
Handling Write Operations in Chorus.....	184
The Mach Microkernel	185
Handling Read Operations in Mach.....	185
Handling Write Operations in Mach.....	186
What Happened to Microkernel Technology?	186
Summary	187
 Chapter 9 Disk-Based Filesystem Case Studies	 189
The VERITAS Filesystem.....	189
VxFS Feature Overview	190
Extent-Based Allocation	190
VxFS Extent Attributes	191
Caching Advisories.....	193
User and Group Quotas	194
Filesystem Snapshots / Checkpoints	194
Panic Free and I/O Error Handling Policies	194
VxFS Clustered Filesystem	195
The VxFS Disk Layouts	195
VxFS Disk Layout Version 1	196
VxFS Disk Layout Version 5.....	197
Creating VxFS Filesystems	200
Forced Unmount	201
VxFS Journaling	201
Replaying the Intent Log.....	204
Extended Operations	204
Online Administration	204
Extent Reorg and Directory Defragmentation.....	206
VxFS Performance-Related Features.....	206
VxFS Mount Options	206
VxFS Tunable I/O Parameters	209
Quick I/O for Databases	209
External Intent Logs through QuickLog.....	211
VxFS DMAPI Support.....	212
The UFS Filesystem	212

Early UFS History.....	213
Block Sizes and Fragments.....	214
FFS Allocation Policies	215
Performance Analysis of the FFS	216
Additional Filesystem Features.....	216
What's Changed Since the Early UFS Implementation?	217
Solaris UFS History and Enhancements	217
Making UFS Filesystems	217
Solaris UFS Mount Options.....	219
Database I/O Support.....	220
UFS Snapshots.....	220
UFS Logging	224
The ext2 and ext3 Filesystems	224
Features of the ext2 Filesystem.....	225
Per-File Attributes	225
The ext2 Disk Layout	226
ext2 On-Disk Inodes.....	231
Repairing Damaged ext2 Filesystems.....	232
Tuning a ext2 Filesystem	233
Resizing ext2 Filesystems	234
The ext3 Filesystem	234
How to Use an ext3 Filesystem.....	234
Data Integrity Models in ext3	235
How Does ext3 Work?	235
Summary	236
Chapter 10 Mapping Filesystems to Multiprocessor Systems	237
The Evolution of Multiprocessor UNIX.....	237
Traditional UNIX Locking Primitives.....	238
Hardware and Software Priority Levels	239
UP Locking and Pre-SVR4 Filesystems.....	241
UP Locking and SVR4-Based Filesystems	241
Symmetric Multiprocessing UNIX	242
SMP Lock Types	243
Mapping VxFS to SMP Primitives	245
The VxFS Inode Reader/Writer Lock.....	246
The VxFS Getpage and Putpage Locks.....	246
The VxFS Inode Lock and Inode Spin Lock.....	246
The VxFS Inode List Lock.....	246
Summary	247
Chapter 11 Pseudo Filesystems	249
The /proc Filesystem.....	249
The Solaris /proc Implementation	250
Accessing Files in the Solaris /proc Filesystem	253

Tracing and Debugging with /proc	253
The Specfs Filesystem	255
The BSD Memory-Based Filesystem (MFS)	258
The BSD MFS Architecture	259
Performance and Observations	259
The Sun tmpfs Filesystem	260
Architecture of the tmpfs Filesystem	260
File Access through tmpfs	261
Performance and Other Observations	261
Other Pseudo Filesystems	262
The UnixWare Processor Filesystem	262
The Translucent Filesystem	262
Named STREAMS	263
The FIFO Filesystem	263
The File Descriptor Filesystem	263
Summary	264
Chapter 12 Filesystem Backup	265
Traditional UNIX Tools	265
The tar, cpio, and pax Commands	266
The tar Archive Format	266
The USTAR tar Archive Format	266
Standardization and the pax Command	268
Backup Using Dump and Restore	268
Frozen-Image Technology	270
Nonpersistent Snapshots	270
VxFS Snapshots	270
Accessing VxFS Snapshots	272
Performing a Backup Using VxFS Snapshots	273
How VxFS Snapshots Are Implemented	274
Persistent Snapshot Filesystems	274
Differences between VxFS Storage Checkpoints and Snapshots	275
How Storage Checkpoints Are Implemented	276
Using Storage Checkpoints	277
Writable Storage Checkpoints	279
Block-Level Incremental Backups	279
Hierarchical Storage Management	280
Summary	283
Chapter 13 Clustered and Distributed Filesystems	285
Distributed Filesystems	286
The Network File System (NFS)	286
NFS Background and History	286
The Version 1 and 2 NFS Protocols	287

NFS Client/Server Communications.....	288
Exporting, Mounting, and Accessing NFS Filesystems	290
Using NFS	292
The Version 3 NFS Protocol	292
The NFS Lock Manager Protocol.....	294
The Version 4 NFS Protocol and the Future of NFS	295
The NFS Automounter	298
Automounter Problems and the Autofs Filesystem	300
The Remote File Sharing Service (RFS)	300
The RFS Architecture	301
Differences between RFS and NFS.....	302
The Andrew File System (AFS)	303
The AFS Architecture.....	303
Client-Side Caching of AFS File Data	304
Where Is AFS Now?	305
The DCE Distributed File Service (DFS)	305
DCE / DFS Architecture.....	306
DFS Local Filesystems.....	306
DFS Cache Management	306
The Future of DCE / DFS	307
Clustered Filesystems.....	307
What Is a Clustered Filesystem?	308
Clustered Filesystem Components	309
Hardware Solutions for Clustering.....	309
Cluster Management.....	309
Cluster Volume Management.....	310
Cluster Filesystem Management	311
Cluster Lock Management	313
The VERITAS SANPoint Foundation Suite.....	313
CFS Hardware Configuration.....	313
CFS Software Components	314
VERITAS Cluster Server (VCS) and Agents	315
Low Latency Transport (LLT).....	316
Group Membership and Atomic Broadcast (GAB)	317
The VERITAS Global Lock Manager (GLM)	317
The VERITAS Clustered Volume Manager (CVM)	317
The Clustered Filesystem (CFS)	318
Mounting CFS Filesystems.....	319
Handling Vnode Operations in CFS	319
The CFS Buffer Cache	320
The CFS DNLC and Inode Cache.....	321
CFS Reconfiguration	321
CFS Cache Coherency	321
VxFS Command Coordination	322
Application Environments for CFS.....	322

Other Clustered Filesystems	323
The SGI Clustered Filesystem (CXFS)	323
The Linux/Sistina Global Filesystem	323
Sun Cluster	323
Compaq/HP True64 Cluster	324
Summary	324
 Chapter 14 Developing a Filesystem for the Linux Kernel	325
Designing the New Filesystem	326
Obtaining the Linux Kernel Source	328
What's in the Kernel Source Tree	329
Configuring the Kernel	330
Installing and Booting the New Kernel	332
Using GRUB to Handle Bootstrap	333
Booting the New Kernel	333
Installing Debugging Support	334
The printk Approach to Debugging	334
Using the SGI kdb Debugger	335
Source Level Debugging with gdb	337
Connecting the Host and Target Machines	337
Downloading the kgdb Patch	338
Installing the kgdb-Modified Kernel	339
gdb and Module Interactions	340
Building the uxfS Filesystem	341
Creating a uxfS Filesystem	342
Module Initialization and Deinitialization	344
Testing the New Filesystem	345
Mounting and Unmounting the Filesystem	346
Scanning for a UxfS Filesystem	348
Reading the Root Inode	349
Writing the Superblock to Disk	350
Unmounting the Filesystem	352
Directory Lookups and Pathname Resolution	353
Reading Directory Entries	353
Filename Lookup	354
Filesystem/Kernel Interactions for Listing Directories	356
Inode Manipulation	359
Reading an Inode from Disk	359
Allocating a New Inode	361
Writing an Inode to Disk	362
Deleting Inodes	363
File Creation and Link Management	365
Creating and Removing Directories	368
File I/O in uxfS	370
Reading from a Regular File	371

Writing to a Regular File	373
Memory-Mapped Files	374
The Filesystem Stat Interface	376
The Filesystem Source Code	378
Suggested Exercises	403
Beginning to Intermediate Exercises	403
Advanced Exercises	404
Summary	405
 Glossary	 407
 References	 425
 Index	 429