

# Table of Contents

|                                                          |            |
|----------------------------------------------------------|------------|
| <b>Preface</b>                                           | <b>1</b>   |
| The Audience for This Book                               | 1          |
| Organization of the Material                             | 1          |
| Overview of the Book                                     | 3          |
| Background Information                                   | 4          |
| Conventions in This Book                                 | 4          |
| How to Contact Us                                        | 4          |
| Acknowledgments                                          | 5          |
| <b>1. Introduction</b>                                   | <b>6</b>   |
| 1.1 Linux Versus Other Unix-Like Kernels                 | 6          |
| 1.2 Hardware Dependency                                  | 10         |
| 1.3 Linux Versions                                       | 11         |
| 1.4 Basic Operating System Concepts                      | 12         |
| 1.5 An Overview of the Unix Filesystem                   | 16         |
| 1.6 An Overview of Unix Kernels                          | 22         |
| <b>2. Memory Addressing</b>                              | <b>36</b>  |
| 2.1 Memory Addresses                                     | 36         |
| 2.2 Segmentation in Hardware                             | 37         |
| 2.3 Segmentation in Linux                                | 41         |
| 2.4 Paging in Hardware                                   | 44         |
| 2.5 Paging in Linux                                      | 52         |
| 2.6 Anticipating Linux 2.4                               | 63         |
| <b>3. Processes</b>                                      | <b>64</b>  |
| 3.1 Process Descriptor                                   | 64         |
| 3.2 Process Switching                                    | 78         |
| 3.3 Creating Processes                                   | 86         |
| 3.4 Destroying Processes                                 | 93         |
| 3.5 Anticipating Linux 2.4                               | 94         |
| <b>4. Interrupts and Exceptions</b>                      | <b>96</b>  |
| 4.1 The Role of Interrupt Signals                        | 96         |
| 4.2 Interrupts and Exceptions                            | 97         |
| 4.3 Nested Execution of Exception and Interrupt Handlers | 106        |
| 4.4 Initializing the Interrupt Descriptor Table          | 107        |
| 4.5 Exception Handling                                   | 109        |
| 4.6 Interrupt Handling                                   | 112        |
| 4.7 Returning from Interrupts and Exceptions             | 126        |
| 4.8 Anticipating Linux 2.4                               | 129        |
| <b>5. Timing Measurements</b>                            | <b>131</b> |
| 5.1 Hardware Clocks                                      | 131        |
| 5.2 The Timer Interrupt Handler                          | 133        |
| 5.3 PIT's Interrupt Service Routine                      | 134        |
| 5.4 The TIMER_BH Bottom Half Functions                   | 136        |
| 5.5 System Calls Related to Timing Measurements          | 145        |
| 5.6 Anticipating Linux 2.4                               | 148        |

|                                                   |            |
|---------------------------------------------------|------------|
| <b>6. Memory Management</b>                       | <b>149</b> |
| 6.1 Page Frame Management                         | 149        |
| 6.2 Memory Area Management                        | 160        |
| 6.3 Noncontiguous Memory Area Management          | 176        |
| 6.4 Anticipating Linux 2.4                        | 181        |
| <b>7. Process Address Space</b>                   | <b>183</b> |
| 7.1 The Process's Address Space                   | 183        |
| 7.2 The Memory Descriptor                         | 185        |
| 7.3 Memory Regions                                | 186        |
| 7.4 Page Fault Exception Handler                  | 201        |
| 7.5 Creating and Deleting a Process Address Space | 212        |
| 7.6 Managing the Heap                             | 214        |
| 7.7 Anticipating Linux 2.4                        | 216        |
| <b>8. System Calls</b>                            | <b>217</b> |
| 8.1 POSIX APIs and System Calls                   | 217        |
| 8.2 System Call Handler and Service Routines      | 218        |
| 8.3 Wrapper Routines                              | 229        |
| 8.4 Anticipating Linux 2.4                        | 230        |
| <b>9. Signals</b>                                 | <b>231</b> |
| 9.1 The Role of Signals                           | 231        |
| 9.2 Sending a Signal                              | 239        |
| 9.3 Receiving a Signal                            | 242        |
| 9.4 Real-Time Signals                             | 251        |
| 9.5 System Calls Related to Signal Handling       | 252        |
| 9.6 Anticipating Linux 2.4                        | 257        |
| <b>10. Process Scheduling</b>                     | <b>258</b> |
| 10.1 Scheduling Policy                            | 258        |
| 10.2 The Scheduling Algorithm                     | 261        |
| 10.3 System Calls Related to Scheduling           | 272        |
| 10.4 Anticipating Linux 2.4                       | 276        |
| <b>11. Kernel Synchronization</b>                 | <b>277</b> |
| 11.1 Kernel Control Paths                         | 277        |
| 11.2 Synchronization Techniques                   | 278        |
| 11.3 The SMP Architecture                         | 286        |
| 11.4 The Linux/SMP Kernel                         | 290        |
| 11.5 Anticipating Linux 2.4                       | 302        |
| <b>12. The Virtual Filesystem</b>                 | <b>303</b> |
| 12.1 The Role of the VFS                          | 303        |
| 12.2 VFS Data Structures                          | 308        |
| 12.3 Filesystem Mounting                          | 324        |
| 12.4 Pathname Lookup                              | 329        |
| 12.5 Implementations of VFS System Calls          | 333        |
| 12.6 File Locking                                 | 337        |
| 12.7 Anticipating Linux 2.4                       | 342        |

|                                                 |            |
|-------------------------------------------------|------------|
| <b>13. Managing I/O Devices</b>                 | <b>343</b> |
| 13.1 I/O Architecture                           | 343        |
| 13.2 Associating Files with I/O Devices         | 348        |
| 13.3 Device Drivers                             | 353        |
| 13.4 Character Device Handling                  | 360        |
| 13.5 Block Device Handling                      | 361        |
| 13.6 Page I/O Operations                        | 377        |
| 13.7 Anticipating Linux 2.4                     | 380        |
| <b>14. Disk Caches</b>                          | <b>382</b> |
| 14.1 The Buffer Cache                           | 383        |
| 14.2 The Page Cache                             | 396        |
| 14.3 Anticipating Linux 2.4                     | 398        |
| <b>15. Accessing Regular Files</b>              | <b>400</b> |
| 15.1 Reading and Writing a Regular File         | 400        |
| 15.2 Memory Mapping                             | 408        |
| 15.3 Anticipating Linux 2.4                     | 416        |
| <b>16. Swapping: Methods for Freeing Memory</b> | <b>417</b> |
| 16.1 What Is Swapping?                          | 417        |
| 16.2 Swap Area                                  | 420        |
| 16.3 The Swap Cache                             | 429        |
| 16.4 Transferring Swap Pages                    | 433        |
| 16.5 Page Swap-Out                              | 437        |
| 16.6 Page Swap-In                               | 442        |
| 16.7 Freeing Page Frames                        | 444        |
| 16.8 Anticipating Linux 2.4                     | 450        |
| <b>17. The Ext2 Filesystem</b>                  | <b>451</b> |
| 17.1 General Characteristics                    | 451        |
| 17.2 Disk Data Structures                       | 453        |
| 17.3 Memory Data Structures                     | 459        |
| 17.4 Creating the Filesystem                    | 463        |
| 17.5 Ext2 Methods                               | 464        |
| 17.6 Managing Disk Space                        | 466        |
| 17.7 Reading and Writing an Ext2 Regular File   | 473        |
| 17.8 Anticipating Linux 2.4                     | 475        |
| <b>18. Process Communication</b>                | <b>476</b> |
| 18.1 Pipes                                      | 477        |
| 18.2 FIFOs                                      | 483        |
| 18.3 System V IPC                               | 486        |
| 18.4 Anticipating Linux 2.4                     | 499        |
| <b>19. Program Execution</b>                    | <b>500</b> |
| 19.1 Executable Files                           | 500        |
| 19.2 Executable Formats                         | 512        |
| 19.3 Execution Domains                          | 514        |
| 19.4 The exec-like Functions                    | 515        |
| 19.5 Anticipating Linux 2.4                     | 519        |

|                                                    |                |
|----------------------------------------------------|----------------|
| <b>A. System Startup</b> .....                     | <b>520</b>     |
| A.1 Prehistoric Age: The BIOS .....                | <b>520</b>     |
| A.2 Ancient Age: The Boot Loader .....             | <b>521</b>     |
| A.3 Middle Ages: The setup( ) Function .....       | <b>523</b>     |
| A.4 Renaissance: The startup_32( ) Functions ..... | <b>523</b>     |
| A.5 Modern Age: The start_kernel( ) Function ..... | <b>524</b>     |
| <br><b>B. Modules</b> .....                        | <br><b>526</b> |
| B.1 To Be (a Module) or Not to Be? .....           | <b>526</b>     |
| B.2 Module Implementation .....                    | <b>527</b>     |
| B.3 Linking and Unlinking Modules .....            | <b>529</b>     |
| B.4 Linking Modules on Demand .....                | <b>531</b>     |
| <br><b>C. Source Code Structure</b> .....          | <br><b>533</b> |
| <br><b>Colophon</b> .....                          | <br><b>536</b> |