



"The author does an outstanding job explaining Linux clusters."

—JOE BROZEC, INFORMATION TECHNOLOGY
Southwest Power Pool

Building Clustered Linux Systems

Presents a practical approach to ~~meeting~~ ~~meeting~~ ~~meeting~~ needs and designing clustered Linux systems

Covers information on physical construction, cabling, planning for power, and dealing with heat dissipation

Provides concrete examples of configuring the Linux operating system, management techniques, and selecting management software packages

Robert W. Lucke



CONTENTS

List of Figures	xxiii
List of Tables	xxix
Preface	xxxii
Acknowledgments	xxxv
Introduction	xxxvii

Part I Introduction to Cluster Concepts	1
--	----------

Chapter 1 Parallel Power: Defining the Clustered System Approach	3
---	----------

1.1 Avoiding Difficulties with the Word <i>Cluster</i>	3
1.2 Defining a Cluster	4
1.3 The Evolution of a Clustered Solution	4
1.3.1 Uniprocessor Systems (UPs)	5
1.3.2 SMP Systems	6
1.3.3 Networks of Independent Systems	7
1.4 Collapsed Network Computing for Engineering	15
1.5 Scientific Cluster Computing	16
1.5.1 An Example Parallel Problem	17
1.5.2 Refining the Parallel Example	19
1.5.3 Software Communication Facilities	21
1.5.4 High-Speed Interconnect (HSI)	22
1.6 Revisiting the Definition of <i>Cluster</i>	22
1.7 Commercial Cluster Computing	23
1.8 High Performance, High Throughput, and High Availability	23
1.9 A Formal Definition of <i>Cluster</i>	24
1.10 The Why and Wherefore of Clusters	24
1.11 Summary	25

Chapter 2 One Step at a Time: A Process for Building Clusters	27
--	-----------

2.1 Building Clusters as a Complex Endeavor	27
---	----

2.2	Talking about the "P Word"	29
2.3	Presenting a Formal Cluster Creation Process	30
2.3.1	Phase 1: Cluster Solution Design	30
2.3.2	Phase 2: Cluster Installation	34
2.3.3	Phase 3: Cluster Testing	39
2.4	Formal Cluster Process Summary	41

Part II Cluster Architecture and Hardware Components **43**

Chapter 3 Underneath the Hood: Cluster Hardware Components and Architecture **45**

3.1	Hardware Categories in a Cluster	45
3.1.1	Passive Hardware Elements in a Cluster	46
3.1.2	Active Hardware Elements in a Cluster	47
3.1.3	Cluster Resources and the "Outside" World	49
3.2	A Survey of Cluster Hardware Configurations	51
3.3	High-Throughput Cluster Configurations	51
3.3.1	A "Carpet" Cluster	52
3.3.2	Compute "Farms and Ranches"	53
3.4	High-Availability Cluster Configurations	57
3.4.1	An Example "Virtual" Web Server	57
3.4.2	A Parallel Database Server	59
3.5	High-Performance Cluster Configurations	61
3.5.1	A Visualization Cluster	62
3.5.2	High-Performance Parallel Application Configurations	65
3.6	Common Cluster Hardware Architecture	67
3.7	Cluster Hardware Architecture Summary	68

Chapter 4 Any Way You Slice It: Work and Master Nodes in a Cluster **69**

4.1	Criteria for Selecting Compute Slices	69
4.2	An Example Compute Slice from Hewlett-Packard	70
4.2.1	Analysis of the Example Compute Slice	72
4.2.2	Comparing the Example Compute Slice with Similar Systems	77
4.2.3	Example Clusters Using Our Compute Slices	80
4.3	Thirty-two Bit and 64-Bit Compute Slices	82
4.3.1	Physical RAM Addressing	82
4.3.2	Process Virtual Address Space	83
4.3.3	Software Implications of 64-Bit Hardware	85

4.4	Memory Bandwidth	86
4.5	Memory and Cache Latency	87
4.6	Number of Processors in a Compute Slice	89
4.7	I/O Interface Capacity and Performance	89
4.7.1	PCI Implementation	90
4.7.2	Accelerated Graphics Port	91
4.8	Compute Slice Operating System Support	91
4.9	Master Node Characteristics	91
4.10	Compute Slice and Master Node Summary	92

Chapter 5 Packet In: Cluster Networking Basics and Example Devices 95

5.1	A Short View of Ethernet Networking History	95
5.2	The Open System Interconnect (OSI) Communication Model	96
5.3	Ethernet Network Topologies	96
5.3.1	Ethernet Frames	98
5.3.2	Ethernet Hubs	98
5.3.3	Network Routers	99
5.4	Internet Protocol and Addressing	101
5.4.1	IP and TCP/UDP	102
5.4.2	IP Addressing	102
5.4.3	IP Subnetting	104
5.4.4	IP Supernetting	106
5.4.5	Ethernet Unicast, Multicast, and Broadcast Frames	106
5.4.6	Address Resolution Protocol (ARP)	107
5.4.7	IPv4 and IPv6	107
5.4.8	Private, Nonroutable Network Addresses	108
5.5	Ethernet Switching Technology	109
5.5.1	Half and Full Duplex Operation	109
5.5.2	Store and Forward versus Cut-through Switching	110
5.5.3	Collision Domains and Switching	110
5.5.4	Link Aggregation	111
5.5.5	Virtual LANs	113
5.5.6	Jumbo Frames	113
5.5.7	Managed versus Unmanaged Switches	114
5.6	Example Switches	115
5.6.1	A GbE Edge Switch	115
5.6.2	Ethernet Core Switches	116

5.7 Ethernet Networking Summary	118
---------------------------------	-----

Chapter 6 Tying It Together: Cluster Data, Management, and Control Networks 119

6.1 Networked System Management and Serial Port Access	119
6.1.1 Remote System Management Access	120
6.1.2 Keyboard, Video, and Mouse Switches	121
6.1.3 Serial Port Concentrators or Switches	121
6.2 Cluster Ethernet Network Design	122
6.2.1 Choosing a Clusterwide IP Address Scheme	123
6.2.2 IP Addressing Conventions	123
6.2.3 Using Nonroutable Network Addresses	124
6.3 An Example Cluster Ethernet Network Design	125
6.3.1 Choosing the Type of Network and Address Ranges	125
6.3.2 Device Addressing Schemes	126
6.3.3 The Management and Control Networks	127
6.3.4 The Data Network	128
6.3.5 Example IP Address Assignments	130
6.4 Cluster Network Design Summary	131

Chapter 7 Life in the Fast LAN: HSIs and Your Cluster 133

7.1 HSIs	133
7.2 HSI Latency and Bandwidth	134
7.3 Examining HSI Topologies	135
7.3.1 Some Common Topologies	136
7.3.2 Cross-Sectional Bandwidth	136
7.3.3 Clos Networks	138
7.3.4 Fat Tree Networks	138
7.4 Ethernet for HSI	139
7.4.1 An Example Ethernet HSI Network	141
7.4.2 Direct Attach Example Bandwidth	142
7.4.3 Multilevel Attach Example Bandwidth	143
7.4.4 A Larger Ethernet HSI Example	144
7.4.5 Other Ethernet HSI Configurations	145
7.5 Myricom's Myrinet HSI	146
7.6 Infiniband	148
7.7 Dolphin	151
7.8 Quadrics QsNet	152
7.9 HSI Technology Summary and Comparison	154

Part III Cluster Software Architecture	155
Chapter 8 The Right Stuff: Linux as the Basis for Clusters	157
8.1 Choosing a Cluster Operating System	157
8.1.1 Hardware Support	158
8.1.2 Operating System Stability	158
8.1.3 Software License Costs	158
8.1.4 Manageability	159
8.1.5 Software Flexibility	159
8.1.6 Openness	159
8.1.7 Scalability	160
8.1.8 Software Availability and Cost	160
8.1.9 Multiple Support Options	160
8.2 Introducing the Linux Operating System and Licensing	161
8.3 Linux Distributions	163
8.4 Managing Open-Source Software “Churn”	164
8.5 Commercial Linux Distributions	165
8.5.1 Red Hat Linux	166
8.5.2 SUSE Linux	168
8.5.3 Conclusions about Commercial Linux Distributions	170
8.6 Free Linux Distributions	171
8.6.1 The Fedora Project	171
8.6.2 Debian Linux	172
8.6.3 Conclusions about Free Distributions	172
8.7 Conclusions about Linux for Clusters	173
Chapter 9 Round and Round It Goes: Booting, Disks, Partitioning, and Local File Systems	175
9.1 Disk Partitioning, Booting, and the BIOS	175
9.1.1 Default Disk Partitioning	176
9.1.2 A Brief Note on IA-64 Disk Partitioning	181
9.1.3 Red Hat Linux Boot Loaders	182
9.2 Booting the Linux Kernel	183
9.3 The Linux Initial RAM Disk Image	185
9.4 Linux Local Disk Storage	188
9.4.1 Using the Software RAID 5 Facility	188
9.4.2 Using Software RAID 1 for System Disks	191
9.4.3 RAID Multipath	195

9.4.4	Recovering from Software RAID Failures	196
9.5	Linux File System Types	204
9.6	The Linux <code>/proc</code> and <code>devfs</code> Pseudo File Systems	205
9.7	The Linux <code>ext2</code> and <code>ext3</code> Physical File Systems	207
9.7.1	File System Volume Labels	209
9.7.2	Creating the Example <code>ext3</code> File System	209
9.7.3	Linux <code>ext3</code> Journal Behavior and Options	210
9.7.4	The <code>ext</code> File System Stride Option for RAID	211
9.8	Standard Mount Options for All File Systems	212
9.9	The Temporary File System	213
9.10	Other Available File System Types	214
9.11	Advanced Performance Tuning	214
9.12	A Word about SMART Monitoring for Disks	215
9.13	Local Disks and File Systems Summary	217
Chapter 10	Supporting Role: Infrastructure Services and Administration	219
10.1	The Big Infrastructure Picture	219
10.2	Initializing Your Cluster's Software Infrastructure	220
10.3	Infrastructure Implementation Recommendations	221
10.3.1	Avoiding Service Interference	222
10.3.2	Redundant Copies of Essential Services	223
10.3.3	Services with Fall-Back Capabilities	224
10.3.4	Single-Point Administration	224
10.3.5	Choosing Efficient Services	225
10.3.6	Management of Configuration Information	226
10.4	Protecting Active Configuration Information	227
10.5	Preparation for Infrastructure Installation	227
10.5.1	Order of Installation	228
10.5.2	Steps for Installing Infrastructure Services	228
10.5.3	Loading the Linux Operating System Distribution	232
10.6	Networking	233
10.6.1	Configuring Ethernet Switching Equipment	233
10.6.2	Network Aliases	234
10.6.3	Channel Bonding	237
10.6.4	Setting the Ethernet Link MTU Size	241
10.6.5	The Media-Independent Interface (MII) Tool	242
10.7	Enabling and Starting Linux Services	243

10.8	Time Synchronization	244
10.9	Name Services	246
10.9.1	Host Naming Conventions	247
10.9.2	The Name Service Switch File	248
10.9.3	The Hosts File	249
10.9.4	The DNS	250
10.9.5	The NIS	256
10.9.6	Name Resolution Recommendations	263
10.10	Infrastructure Services Summary	265

Chapter 11 Reach Out and Access Something: Remote Access Services, DHCP, and System Logging 267

11.1	Continuing Infrastructure Installation	267
11.2	"Traditional" User Login and Authentication	268
11.2.1	Using Groups and Directory Permissions	270
11.2.2	Distributing Password Information with NIS	272
11.2.3	Introducing Kerberos	272
11.2.4	Configuring a Kerberos KDC on Linux	274
11.2.5	Creating a Kerberos Slave KDC	278
11.2.6	Kerberos Summary	279
11.3	Remote Access Services	280
11.4	Using BSD Remote Access Services	281
11.5	Kerberized Versions of BSD/ARPA Remote Services	282
11.6	The Secure Shell	286
11.6.1	SSH and Public Key Encryption	287
11.6.2	Configuring the SSH Client and Server	289
11.6.3	Configuring User Identity for SSH	290
11.6.4	SSH Host Keys, and Known and Authorized Hosts	292
11.6.5	Using the Authorized Keys File	293
11.6.6	Fine-Tuning SSH Access	295
11.6.7	SSH <code>scp</code> and <code>sftp</code> Commands	296
11.6.8	SSH Forwarding	296
11.6.9	SSH Summary	299
11.7	The Parallel Distributed Shell	300
11.7.1	Getting and Installing PDSH	300
11.7.2	Compiling PDSH to Use SSH	303
11.7.3	Using PDSH in Your Cluster	304
11.7.4	PDSH Summary	306

11.8	Configuring DHCP	307
11.8.1	Client-side DHCP Information	307
11.8.2	Configuring the DHCP Server	310
11.9	Logging System Activity	313
11.9.1	Operation of the System Logging Daemon	314
11.9.2	Kernel Message Logging	316
11.9.3	Enabling Remote Logging	317
11.9.4	Using logrotate to Archive Log Files	319
11.9.5	Using logwatch Reporting	321
11.9.6	An Example Subsystem Logging Design	323
11.9.7	Linux System Logging Summary	325
11.10	Access and Logging Services Summary	325
Chapter 12	Installment Plan: Introduction to Compute Slice Configuration and Installation	327
12.1	Compute Slice Configuration Considerations	327
12.2	One Thousand Pieces Flying in Close Formation	328
12.3	The Single-System View	329
12.3.1	Shared System Structure, Individual System Personality	330
12.3.2	Accomplishing Shared System Structure	332
12.3.3	Compute Slice Software Requirements	333
12.4	A Generalized Network Boot Facility: pxelinux	333
12.4.1	Configuring TFTP for Booting	334
12.4.2	Configuring the pxelinux Software	336
12.4.3	The pxelinux Configuration Files	337
12.5	Configuring Network kickstart	340
12.5.1	The kickstart File Format	341
12.5.2	Making the Install Media Available for kickstart	343
12.5.3	The Network kickstart Directory	343
12.6	NFS Diskless Configuration	345
12.6.1	The Linux Terminal Server Project (LTSP)	346
12.6.2	Cluster NFS	348
12.7	Introduction to Compute Slice Installation Summary	349
Chapter 13	Improving Your Images: System Installation with SystemImager	351
13.1	Using the SystemImager Software	351
13.1.1	Downloading and Installing SI	352
13.1.2	Configuring the SI Server	356

13.1.3	The SI Cold Installation Boot Process	357
13.1.4	SI Server Commands	359
13.1.5	Installing and Configuring the SI Client Software	359
13.1.6	Capturing a Client Image	360
13.1.7	Forcing Hardware-to-Driver Mapping with SystemConfigurator	366
13.1.8	Installing a Client Image	366
13.1.9	Updating Client Software without Reinstalling	367
13.1.10	Image Management and Naming	368
13.1.11	Avoiding the Big MAC-Gathering Syndrome	369
13.1.12	Summary	369
13.2	Multicast Installation	371
13.2.1	Multicast Basics	371
13.2.2	An Open-Source Multicast Facility: udpcast	373
13.2.3	A Simple Multicast Example	374
13.2.4	A More Complex Example	375
13.2.5	Command-line Prototyping with Multicast	376
13.2.6	Prototyping a Network Multicast Installation	377
13.2.7	Making More Modifications	379
13.2.8	Generalizing the Multicast Installation Prototype	386
13.2.9	Triggering a Multicast Installation	389
13.3	The SI flamethrower Facility	391
13.3.1	Installing flamethrower	391
13.3.2	Activating flamethrower	392
13.3.3	Additional SI Functionality in Version 3.2.0	394
13.4	System Installation with SI Summary	394

Chapter 14 To Protect and Serve: Providing Data to Your Cluster **397**

14.1	Introduction to Cluster File Systems	397
14.1.1	Cluster File System Requirements	398
14.1.2	Networked File System Access	399
14.1.3	Parallel File System Access	401
14.2	The NFS	404
14.2.1	Enabling NFS on the Server	405
14.2.2	Adjusting NFS Mount Daemon Protocol Behavior	406
14.2.3	Tuning the NFS Server Network Parameters	408
14.2.4	NFS and TCP Wrappers	410
14.2.5	Exporting File Systems on the NFS Server	411

14.2.6	Starting the NFS Server Subsystem	412
14.2.7	NFS Client Mount Parameters	412
14.2.8	Using autofs on NFS Clients	414
14.2.9	NFS Summary	415
14.3	A Survey of Some Open-Source Parallel File Systems	415
14.3.1	The Parallel Virtual File System (PVFS)	416
14.3.2	The Open Global File System (OpenGFS)	419
14.3.3	The Lustre File System	421
14.4	Commercially Available Cluster File Systems	428
14.4.1	Red Hat Global File System (GFS)	428
14.4.2	The PolyServe Matrix File System	430
14.4.3	Oracle Cluster File System (OCFS)	431
14.5	Cluster File System Summary	431
Chapter 15	Stuck in the Middle: Cluster Middleware	433
15.1	Introduction to Cluster Middleware	433
15.1.1	Describing the Parallel Application Execution Environment	434
15.1.2	The HSI Message-Passing Facility	435
15.1.3	Load Balancing or Job Scheduling	435
15.1.4	Cluster Resource Management	437
15.1.5	Custom Scheduling	437
15.1.6	Monitoring, Measuring, and Managing Your Cluster	439
15.2	The MPICH Library	439
15.2.1	Introduction to MPICH	439
15.2.2	Downloading and Installing MPICH	440
15.2.3	Using <code>mpirun</code>	441
15.2.4	Special Versions of MPICH	444
15.2.5	MPICH Summary	444
15.3	The Simple Linux Utility for Resource Management	445
15.4	The Maui Scheduler	446
15.4.1	Maui Scheduler Software Architecture	446
15.4.2	Job Scheduling in Maui	448
15.4.3	Maui Scheduler Summary	448
15.5	The Ganglia Distributed Monitoring and Execution System	449
15.5.1	The Ganglia Software Architecture	450
15.5.2	Introducing RRD Software: <code>rrdtool</code>	451
15.5.3	Downloading and Installing Ganglia Software	455
15.5.4	Ganglia's <code>gmond</code> and <code>gmetad</code> Daemons	455

15.5.5	Adding Your Own Ganglia Metrics	457
15.5.6	Parallel Authentication with <code>authd</code> and <code>gexec</code>	460
15.5.7	Starting Parallel Programs with <code>gexec</code>	461
15.5.8	Ganglia Summary	462
15.6	Monitoring with Nagios	462
15.6.1	Explaining Nagios	463
15.6.2	Downloading and Installing Nagios	463
15.6.3	Configuring the Web Server for Nagios	465
15.6.4	Configuring and Using Nagios	466
15.6.5	Nagios Summary	473
15.7	Cluster Middleware Summary	474
15.8	An Afterword on Linux High-Availability and Open-Source	475

Chapter 16 Put Tab A in Slot C: OSCAR, Rocks, OpenMOSIX, and the Globus Toolkit **477**

16.1	Introducing Cluster-Building Toolkits	477
16.2	General Cluster Toolkit Installation Process	479
16.3	Installing a Cluster with OSCAR	480
16.3.1	OSCAR Initial Software Installation and Configuration	480
16.3.2	The OSCAR Installation Wizard	481
16.3.3	OSCAR Package Configuration	483
16.3.4	Building an OSCAR Compute Slice Image	485
16.3.5	Defining and Installing OSCAR Clients	487
16.3.6	Completing the OSCAR Installation	488
16.3.7	Adding and Deleting OSCAR Clients	489
16.3.8	OSCAR Summary	491
16.4	Installing a Cluster with NPACI Rocks	492
16.4.1	Getting the Rocks Software	493
16.4.2	Installing a Cluster Front-End Node Using Rocks	495
16.4.3	Completing the Installation	499
16.4.4	Rocks System Administration	501
16.4.5	Rocks Summary	502
16.5	The OpenMOSIX Project	502
16.5.1	Getting and Installing OpenMOSIX	503
16.5.2	Configuration of OpenMOSIX Clusters	503
16.5.3	OpenMOSIX Summary	504
16.6	Introduction to the Grid Concept	504
16.7	The Globus Toolkit	505

16.8 Cluster-Building Toolkit Summary

507

Part IV Building and Deploying Your Cluster	509
Chapter 17 Dollars and Sense: Cluster Economics	511
17.1 Initial Perceptions	511
17.2 Setting the Ground Rules	512
17.3 Cluster Cabling and Complexity	513
17.4 Eight-Compute Slice Cluster Hardware Costs	514
17.5 Sixteen-Compute Slice Cluster Hardware Costs	515
17.6 Thirty-two-Compute Slice Hardware Costs	516
17.7 Sixty-four-Compute Slice Hardware Costs	519
17.8 One Hundred Twenty-eight-Compute Slice Hardware Costs	519
17.9 The Land beyond 128 Compute Slices	521
17.10 Hardware Cost Trends and Analysis	522
17.11 Cluster Economics Summary	524
Chapter 18 Racking Your Brains: Example Cluster Rack Assembly Steps	529
18.1 Examining the Cluster Assembly Process	529
18.2 Assembly Assumptions	530
18.3 Some "Rules of Thumb" for Physical Cluster Assembly	530
18.4 Detailed Cluster Assembly Steps	531
18.4.1 Physical Rack Assembly	532
18.4.2 Physical Management Rack Assembly	533
18.4.3 Physical Compute Rack Assembly	534
18.4.4 Physical Compute Rack System Installation	534
18.4.5 Physical Rack Final Assembly and Checkout	536
18.4.6 Individual System Checkout	536
18.4.7 Physical Rack Cleanup	536
18.4.8 Physical Rack Positioning	537
18.4.9 Interrack Configuration	537
18.4.10 Interrack Cabling	537
18.4.11 Final Cluster Hardware Assembly and Checkout	538
18.5 Learning from the Example Steps	539
18.5.1 Finding Efficiencies in Cluster Construction	539
18.5.2 Parallelism in Rack Verification and Checkout	542
18.5.3 Parallelism in Interrack Cabling	542
18.5.4 Types of Teams and Specific Skills	542

18.6	Physical Assembly Conclusions	543
Chapter 19	Getting Your Cluster Wired: An Example Cable-Labeling Scheme	545
19.1	Defining the Cable Problem	546
19.2	Different Classes of Cabling	546
19.2.1	Intrrack Cables	547
19.2.2	Interrack Cables	547
19.3	A First Pass at a Cable-Labeling Scheme	548
19.4	Refining the Cable Documentation Scheme	549
19.4.1	Labeling Cable Ends	550
19.4.2	Tracking and Documenting the Connections	551
19.5	Calculating the Work in Cable Installation	552
19.6	Minimizing Interrack Cabling	553
19.7	Cable Labeling System Summary	555
Chapter 20	Physical Constraints: Heat, Space, and Power	557
20.1	Identifying Physical Constraints for Your Cluster	557
20.2	Space, the Initial Frontier	558
20.3	Power-Up Requirements	560
20.4	System Power Utilization	560
20.5	Taking the Heat	562
20.6	Physical Constraints Summary	564
Appendix A	Acronym List	567
Appendix B	List of URLs and Software Sources	573
Glossary		581
Bibliography		587
Index		589