



The

LINUX ENTERPRISE CLUSTER

BUILD A *HIGHLY AVAILABLE* CLUSTER WITH
COMMODITY HARDWARE AND FREE SOFTWARE



Karl Kopper



NO STARCH
PRESS

CONTENTS IN DETAIL

INTRODUCTION	1
Properties of a Linux Enterprise Cluster	2
Architecture of the Linux Enterprise Cluster	4
The Load Balancer	4
The Shared Storage Device	5
The Print Server	5
The Cluster Node Manager	5
No Single Point of Failure	6
In Conclusion	7

PRIMER	9
High Availability Terminology	10
Linux Enterprise Cluster Terminology	10

PART I: CLUSTER RESOURCES

1	
STARTING SERVICES	15
How Do Cluster Services Get Started?	15
Starting Services with init	16
The /etc/inittab File	16
Respawning Services with init	18
Managing the init Script Symbolic Links with chkconfig	18
Managing the init Script Symbolic Links with ntsysv	20
Removing Services You Do Not Need	21
Using the Red Hat init Scripts on Cluster Nodes	21
In Conclusion	31

2	
HANDLING PACKETS	33
Netfilter	34
A Brief History of Netfilter	35
Setting the Default Chain Policy	38
Using iptables and ipchains	39
Clear Existing Rules	39
Set the Default INPUT Chain Policy	39
FTP	40
Passive FTP	41
DNS	41

Telnet	42
SSH	42
Email	42
HTTP	43
ICMP	43
Review and Save Your Rules	43
Routing Packets with the Linux Kernel	44
Matching Packets for a Particular Destination Network	45
Matching Packets for a Particular Destination Host	45
Matching Packets for Any Destination or Host	46
To View Your Routing Rules	47
Making Your Routing Rules Permanent	47
iptables and Routing Resource Script	47
The ip Command	50
In Conclusion	51

3 COMPILING THE KERNEL 53

What You Will Need	54
Step 1: Get the Source Code	54
Using the Stock Kernel	54
Using the Kernel Supplied with Your Distribution	55
Decide Which Kernel Version to Use	55
Step 2: Set the Options You Want	56
Installing a New Kernel	56
Upgrading or Patching the Kernel	56
Upgrading a Kernel from a Distribution Vendor	57
Set Your Kernel Options	57
Step 3: Compile the Code	59
Step 4: Install the Object Code and Configuration File	60
Install the System.map	61
Save the Kernel Configuration File	62
Step 5: Configure Your Boot Loader	62
In Conclusion	66

PART II: HIGH AVAILABILITY

4 SYNCHRONIZING SERVERS WITH RSYNC AND SSH 69

rsync	70
Open SSH 2 and rsync	71
SSH: A Secure Data Transport	71
SSH Encryption Keys	72
Establishing the Two-Way Trust Relationship: Part 1	73
Establishing the Two-Way Trust Relationship: Part 2	76
Two-Node SSH Client-Server Recipe	77

Create a User Account That Will Own the Data Files	78
Configure the Open SSH2 Host Key on the SSH Server	79
Create a User Encryption Key for This New Account on the SSH Client	79
Copy the Public Half of the User Encryption Key from the SSH Client to the SSH Server	81
Test the Connection from the SSH Client to the SSH Server	82
Using the Secure Data Transport	86
Improved Security	86
rsync over SSH	87
Copying a Single File with rsync	89
rsync over Slow WAN Connections	89
Scheduled rsync Snapshots	90
ipchains/iptables Firewall Rules for rsync and SSH	92
In Conclusion	93

5 CLONING SYSTEMS WITH SYSTEMIMAGER 95

SystemImager	95
Cloning the Golden Client with SystemImager	96
SystemImager Recipe	96
Install the SystemImager Server Software on the SystemImager Server	97
Using the Installer Program from SystemImager	98
Install the SystemImager Client Software on the Golden Client	98
Create a System Image of the Golden Client on the SystemImager Server	99
Make the Primary Data Server into a DHCP Server	101
Create a Boot Floppy for the Golden Client	103
Start rsync as a Daemon on the Primary Data Server	106
Install the Golden Client System Image on the New Clone	106
Post-installation Notes	107
Performing Maintenance: Updating Clients	108
SystemInstaller	108
System Configurator	109
In Conclusion	109

6 HEARTBEAT INTRODUCTION AND THEORY 111

The Physical Paths of the Heartbeats	112
Serial Cable Connection	113
Ethernet Cable Connection	113
Partitioned Clusters and STONITH	113
Heartbeat Control Messages	114
Heartbeats	114
Cluster Transition Messages	114
Retransmission Requests	115
Ethernet Heartbeat Control Messages	115
Security and Heartbeat Control Messages	115
How Client Computers Access Resources	115
Failover Using IP Address Takeover (IPAT)	116

Secondary IP Addresses and IP Aliases	116
Ethernet NIC Device Names	119
Secondary IP Address Names	119
IP Aliases	120
Offering Services	121
Gratuitous ARP (GARP) Broadcasts	122
Resource Scripts	125
Status of the Resource	125
Resource Ownership	126
Using init Scripts as Heartbeat Resource Scripts	128
Heartbeat Configuration Files	130
In Conclusion	130

7

A SAMPLE HEARTBEAT CONFIGURATION

131

Recipe	132
Preparations	132
Step 1: Install Heartbeat	132
Step 2: Configure /etc/ha.d/ha.cf	134
Step 3: Configure /etc/ha.d/haresources	136
Configure the haresources File	137
Step 4: Configure /etc/ha.d/authkeys	139
Step 5: Install Heartbeat on the Backup Server	139
Step 6: Set the System Time	140
Step 7: Launch Heartbeat	140
Launch Heartbeat on the Primary Server	140
Launch Heartbeat on the Backup Server	142
Examining the Log Files on the Primary Server	143
Stopping and Starting Heartbeat	144
Monitoring Resources	145
In Conclusion	145

8

HEARTBEAT RESOURCES AND MAINTENANCE

147

The Haresources File Syntax	147
Haresources File Syntax: Primary-Server Name	148
Haresources File Syntax: IP Alias	148
Heartbeat's Automated Network Interface Card Selection Process	149
Specifying a Network Interface Card	151
Customizing IP Address Takeover with the iptakeover Script	151
The Haresources File Syntax: Resources	152
Load Sharing with Heartbeat	154
Load Sharing with Heartbeat: Round-Robin DNS	155
Problems with Round Robin DNS Load Balancing	156
Wide-Area Load Balancing	157
Operator Alerts: Audible Alarm	157
Operator Alerts: Email Alerts	158
Heartbeat Maintenance	158

Changing Heartbeat Configuration Files	158
Server Maintenance and the Heartbeat auto_failback Option	159
Forcing the Primary Server into Standby Mode	160
Tuning Heartbeat's Deadtime Value	161
Informational Messages in Heartbeat's Log	161
Failover and Respawn (Automatically Restarting Failed Resources)	161
License Manager Failover	162
In Conclusion	162

9 **STONITH AND IPFAIL** **163**

Stonith	164
An Unconventional Approach: Using a Single Stonith Device	165
Sample Heartbeat with Stonith Configuration	166
Stonith Sequence of Events	167
Stonith Devices	168
Viewing the Current List of Supported Stonith Devices	169
The Stonith Meatware "Device"	169
Using the Stonith Meatware Device with Heartbeat	170
Using a "Real" Stonith Device	173
Avoiding Multiple Stonith Events	174
Network Failures	175
ipfail	175
Watchdog and Softdog	176
Enable Watchdog in the Kernel	177
Kernel Panic—Hang or Reboot?	178
Configure Heartbeat to Support Watchdog	178
Testing Your Heartbeat Configuration	179
In Conclusion	181

PART III: CLUSTER THEORY AND PRACTICE

10 **HOW TO BUILD A LINUX ENTERPRISE CLUSTER** **185**

Steps for Building a Linux Enterprise Cluster	186
NAS Server	186
Kernel Netfilter and Kernel Packet Routing	187
Cloning a Linux Machine	187
Cluster Naming Scheme	187
Applying System Configuration Changes to All Nodes	187
Building an LVS-NAT Cluster	187
Building an LVS-DR Cluster	188
Installing Software to Remove Failed Cluster Nodes	188
Installing Software to Monitor the Cluster Nodes	189
Monitoring the Performance of Cluster Nodes	189
Updating Software on Cluster Nodes and Servers	189

Centralizing User Account Administration	189
Installing a Printing System	190
Installing a Highly Available Batch Job-Scheduling System	190
Purchasing the Cluster Nodes	190
In Conclusion	192

11

THE LINUX VIRTUAL SERVER: INTRODUCTION AND THEORY

193

LVS IP Address Name Conventions	194
The Virtual IP (VIP)	194
The Real IP (RIP)	195
The Director's IP (DIP)	195
The Client Computer's IP (CIP)	195
IP Addresses in an LVS Cluster	195
Types of LVS Clusters	196
Network Address Translation (LVS-NAT)	196
Direct Routing (LVS-DR)	198
IP Tunneling (LVS-TUN)	200
LVS Scheduling Methods	202
Fixed (or Non-dynamic) Scheduling Methods	202
Dynamic Scheduling Methods	203
In Conclusion	208

12

THE LVS-NAT CLUSTER

209

How Client Computers Access LVS-NAT Cluster Resources	209
Virtual IP Addresses on LVS-NAT Real Servers	214
Building an LVS-NAT Web Cluster	216
Recipe for LVS-NAT	217
Step 1: Install the Operating System	217
Step 2: Configure and Start Apache on the Real Server	218
Step 3: Set the Default Route on the Real Server	219
Step 4: Install the LVS Software on the Director	220
Step 5: Configure LVS on the Director	221
Step 6: Test the Cluster Configuration	224
LocalNode: Using the Director as a Real Server	227
In Conclusion	229

13

THE LVS-DR CLUSTER

231

How Client Computers Access LVS-DR Cluster Services	232
ARP Broadcasts and the LVS-DR Cluster	236
Client Computers and ARP Broadcasts	236
In Conclusion	238

14

THE LOAD BALANCER

239

LVS and Netfilter	239
The Director's Connection Tracking Table	242
Hash Table Structure	242
Controlling the Hash Buckets	242
Viewing the Connection Tracking Table	243
Timeout Values for Connection Tracking Records	243
Return Packets and the Netfilter Hooks	244
LVS Without Persistence	246
LVS Persistence	247
Persistent Connection Template	247
Types of Persistent Connections	248
Persistent Client Connection (PCC)	249
Persistent Port Connection (PPC)	249
Port Affinity	250
Netfilter Marked Packets	250
In Conclusion	253

15

THE HIGH-AVAILABILITY CLUSTER

255

Redundant LVS Directors	255
High-Availability Cluster Design Goals	256
The High-Availability LVS-DR Cluster	257
Introduction to ldirectord	258
How ldirectord Monitors Cluster Nodes [LVS Real Servers]	258
LVS, Heartbeat, and ldirectord Recipe	261
Hide the Loopback Interface	261
Install the Heartbeat on a Primary and a Backup Director	263
Install ldirectord and Its Required Software Components	263
Install ldirectord	266
Test Your ldirectord Installation	266
Create the ldirectord Configuration File	267
Create the Health Check Web Page	272
Start ldirectord Manually and Test Your Configuration	273
Add ldirectord to the Heartbeat Configuration	274
Stateful Failover of the IPVS Table	275
Modifications to Allow Failover to a Real Server Inside the Cluster	277
In Conclusion	278

16

THE NETWORK FILE SYSTEM

279

Lock Arbitration	280
The Lock Arbitrator	280
The Existing Kernel Lock Arbitration Methods	281
The Network Lock Manager (NLM)	283
NLM and Kernel Lock Arbitration	284
NLM and Kernel BSD flock	284

NLM and Kernel System V lockf	285
NLM and Kernel Posixfcntl	285
NFS and File Lock (dotlock) Arbitration	286
Finding the Locks Held by the Linux Kernel	286
Performance Issues with NFS—Bottlenecks and Perceptions	288
Single Transactions and User Perception of NFS Performance	289
Multiple Transactions and User Perception of NFS Performance	290
Managing Lock and GETATTR Operations in a Cluster Environment	291
Managing Attribute Caching	291
Managing Interactive User Applications and Batch Jobs in a Cluster Environment	292
Run Batch Jobs Outside the Cluster	292
Use Multiple NAS Servers	292
Measuring NFS Latency	293
Measuring Total I/O Operations	294
Achieving the Best NAS Performance Possible	295
NFS Client Configuration Options	296
Putting It All Together	296
Developing NFS	297
Additional Starting Points for Information on Linux and NFS	297
In Conclusion	298

PART IV: MAINTENANCE AND MONITORING

17

THE SIMPLE NETWORK MANAGEMENT PROTOCOL AND MON

301

Mon	301
Mon Alerts	302
Mon Monitoring Scripts	302
Where to Run Mon	302
Basic Mon Recipe	303
Step 1: Compile and Install the fping Package	304
Step 2: Install the SNMP Package	304
Step 3: Install the Required CPAN Modules for Mon	305
Step 4: Install the Mon Software	306
Step 5: Create the /etc/mon/mon.cf Configuration File	307
Step 6: Test by Running the fping.monitor and mail.alert Scripts Manually	308
Step 7: Create the Mon Log Directory and Mon Log File	309
Step 8: Start the Mon Program in Debugging Mode and Test	309
Mon and SNMP "Proof of Concept" Recipe	310
Step 1: Install Net-SNMP Client Software on Each Cluster Node	310
Step 2: Create the snmp.conf Configuration File	311
Step 3: Start the SNMP Agent	311
Step 4: View the SNMP MIB Locally	312
Step 5: Install the SNMP Monitoring Agents	313
Examine the Output of netsnmp-freespace.monitor	314
Install Mon on All Cluster Nodes	314

Mon and SNMP "Real-World" Recipe	315
Step 1: Create the snmpd.conf on All Cluster Nodes	315
Step 2: Install netsnmp-proc.monitor and the New mon.cf File on the Cluster Node Manager	317
Step 3: Install the Mon init Script	318
Step 4: Run SNMP and Mon	319
Email Alerts from Mon	319
Creating Your Own SNMP Script	320
A Sample Custom SNMP Script	320
Monitoring Your SNMP Script with Mon	321
Things to Monitor with SNMP Monitoring Scripts	322
Forcing a Stonith Event with Mon	323
Forcing a Heartbeat Failover with Mon	324
In Conclusion	325

18

GANGLIA

327

Introduction to Ganglia	328
gmond	328
gmetad	328
Installing Ganglia's Prerequisite Packages	329
Install PHP	329
Install RRDtool	329
Install Apache on the Cluster Node Manager	330
Installing Ganglia on the Cluster Node Manager	330
Installing Ganglia on the Cluster Nodes	330
Configuring gmetad and gmond on the Cluster Node Manager	331
Modify /etc/gmetad.conf	331
Modify /etc/gmond.conf	331
Start gmond and gmetad	332
Add the Ganglia Page to Your Apache Configuration	332
The Ganglia Web Package	333
The Title Section	333
The Node Snapshot Section	334
Examining the Cluster Node from the Ganglia Web Package	336
Examining the Cluster Node from the Shell Prompt	337
gstat	338
Running a Command on the Least-Loaded Cluster Node Using gstat	338
Creating Custom Metrics with gmetric	338
In Conclusion	340

19

CASE STUDIES IN CLUSTER ADMINISTRATION

341

Administering Accounts Without Active Directory	342
Legacy Unix Account Administration Methods: The Problem	342
The Best of Both Worlds	343
Building a Reliable Cluster Account Authentication Mechanism	343
Using the Local passwd and Group File on Each Cluster Node	343

A Simple Script	344
Building a Fault-Tolerant Print Spooler	345
Cluster Nodes and Job Ordering	345
LPRng: A Linux Enterprise Cluster Printing System	346
Cluster Nodes and Print Jobs	346
Building the Cluster Printing System Based on LPRng	347
Install LPRng on the Central Print Server	347
Install LPRng on the Cluster Nodes	348
Modify the /etc/printcap.local File on the Cluster Nodes	348
Modify the /etc/printcap File on the Central Print Server	349
Managing Print Jobs on the Central Print Server	350
Managing Print Jobs from the Cluster Nodes	351
Rebooting Nodes for Preventative Maintenance	352
Using ipvsadm Commands to Remove a Cluster Node	352
Changing the Weight of a Cluster Node to 0	352
Disabling Telnet Access to One of the Cluster Nodes	353
Sending and Receiving Email in a Cluster Environment	353
Creating a Batch Job-Scheduling System with No Single Point of Failure	354
Run ssh-keygen	355
Modify the sshd_config File on Each Cluster Node	355
Create the (RSA) known_hosts Entries on the Cluster Node Manager	355
The Batch Job Scheduler	356
In Conclusion	357

20

THE LINUX CLUSTER ENVIRONMENT

359

The Linux Enterprise Cluster	360
Applications Running on the Cluster	361
The Cluster Node Manager	361
The Clients	362
The NAS Server	363
High-Availability NAS Server	364
Highly Available Serial Devices	364
Serial-to-IP Communication Device	365
High-Availability Modems	365
Highly Available Database Server	365
High-Availability SQL Server	366
Putting It All Together	367
The Cluster Environment	367
In Conclusion	368

A

DOWNLOADING SOFTWARE FROM THE INTERNET (FROM A TEXT TERMINAL)

369

Using Lynx	371
Using Wget	372
What to Do with the File You Downloaded	372
tar Files	373

rpm Files	373
Creating Your Own Tape Archive (tar) File	374

B

TROUBLESHOOTING WITH THE TCPDUMP UTILITY **375**

C

ADDING NETWORK INTERFACE CARDS TO YOUR SYSTEM **379**

Monolithic Versus Modular Kernels	380
View Your Existing Configuration	380
Install the Card and Reboot	381
Run linuxconf	382
Testing the New NIC	383
Changes Made by linuxconf	384
Using the New NIC	384
Testing and Troubleshooting	384

D

STRATEGIES FOR DEPENDENCY FAILURES **387**

What Are Dependencies?	387
Dynamic Executables and Shared Objects	388
rpm Packages and Shared Library Dependencies	390
Fixing a Dependency Failure Manually	391
Fixing a Dependency Failure Automatically	392
Using Yum to Install rpm Packages	392
Installing a New rpm Package with Yum	393
In Conclusion	393

E

OTHER POTENTIAL CLUSTER FILESYSTEMS AND LOCK ARBITRATION METHODS **395**

F

LVS CLUSTERS AND THE APACHE CONFIGURATION FILE **397**

ServerName	398
DocumentRoot	398
BindAddress	399
Port	399
Listen	399
Apache Virtual Host Configuration on Cluster Nodes	400
Apache IP-Based Virtual Hosts	400

Name-Based Virtual Hosts	401
Self-Referential (Redirection) URLs	403
IP-Based Virtual Hosts and Self-Referential URLs	404
Name-Based Virtual Hosts and Self-Referential URLs	404
Verify Your Virtual Host Configuration	404

INDEX

407