

SysAdmin Notes for RHCE

Somenath Sinha

January 2018

Contents

| | | |
|----------|---|-----------|
| I | Advanced System Management | 5 |
| 1 | Configuring Authentication | 6 |
| 1.1 | Understanding RedHat Identity Management | 6 |
| 1.1.1 | IdM Server Components and Requirements | 6 |
| 1.1.2 | Preparing IdM Installation | 7 |
| 1.1.3 | Installing IdM | 7 |
| 1.1.4 | Understanding Kerberos Tickets | 8 |
| 1.1.5 | Managing the IdM Server | 8 |
| 1.1.6 | Creating User Accounts | 8 |
| 1.2 | Using authconfig to Setup External Authentication | 9 |
| 1.3 | Configuring a System to Authenticate using Kerberos | 9 |
| 1.3.1 | Troubleshooting Authentication | 10 |
| 1.4 | Understanding authconfig Configuration Files | 11 |
| 1.4.1 | Authconfig Configuration | 11 |
| 1.4.2 | SSSD Configuration | 11 |
| 1.4.3 | Kerberos Configuration File | 12 |
| 1.4.4 | NSSwitch Configuration | 13 |
| 1.4.5 | NSLCD Configuration | 13 |
| 2 | Configuring iSCSI Target and Initiator | 14 |
| 2.1 | Understanding iSCSI Target and Initiator | 14 |
| 2.1.1 | iSCSI Operation | 14 |
| 2.1.2 | iSCSI Components | 15 |
| 2.1.3 | Basic iSCSI Terminology | 15 |
| 2.1.4 | After connecting an initiator to an iSCSI Target | 15 |

| | | |
|----------|--|-----------|
| 2.2 | Setting up an iSCSI Target | 16 |
| 2.2.1 | Creating the LVM | 16 |
| 2.2.2 | Creating the iSCSI configuration using targetcli | 17 |
| 2.2.3 | Target Creation | 18 |
| 2.2.4 | TPG Configuration | 19 |
| 2.2.5 | Adding a rule to the firewall | 21 |
| 2.2.6 | Starting target.service | 21 |
| 2.3 | Connecting the iSCSI Initiator to an iSCSI SAN | 22 |
| 2.3.1 | Setting up an initiator name | 22 |
| 2.3.2 | iscsiadm Command | 22 |
| 2.3.3 | Discovery | 22 |
| 2.3.4 | Login | 23 |
| 2.3.5 | Logout | 24 |
| 2.3.6 | Deleting node information | 24 |
| 2.4 | Verifying the iSCSI Connection | 24 |
| 2.4.1 | Verification on the iSCSI Initiator | 24 |
| 2.4.2 | Verification on the iSCSI Target | 26 |
| 3 | System Performance Reporting | 28 |
| 3.1 | Understanding System Performance Parameters | 28 |
| 3.1.1 | Typical Performance Focus Areas | 28 |
| 3.1.2 | Common Performance Monitoring Tools | 29 |
| 3.2 | Understanding top | 29 |
| 3.3 | Understanding iostat | 30 |
| 3.3.1 | Usage scenario | 30 |
| 3.3.2 | iotop | 31 |
| 3.4 | Understanding vmstat | 31 |
| 3.4.1 | Virtual Memory | 31 |
| 3.4.2 | Resident Memory | 32 |
| 3.4.3 | vmstat | 32 |
| 3.5 | Understanding sar components | 33 |
| 3.5.1 | /etc/cron.d/sysstat | 34 |

| | | |
|-----------|--|-----------|
| 3.5.2 | /var/log/sa/sa[dd] | 34 |
| 3.6 | Setting up sar | 34 |
| 3.7 | Analyzing sar data | 35 |
| 3.7.1 | I/O operations | 35 |
| 3.7.2 | Processor information | 35 |
| 3.7.3 | Network Statistics | 35 |
| 4 | System Optimization Basics | 37 |
| 4.1 | Understanding /proc contents | 37 |
| 4.2 | Analysing the /proc filesystem | 37 |
| 4.2.1 | Process Directories | 37 |
| 4.2.2 | Status files | 38 |
| 4.2.3 | /proc/sys | 39 |
| 4.3 | Optimizing through /proc | 40 |
| 4.3.1 | Sysctl | 41 |
| 4.4 | Introducing sysctl | 41 |
| 4.5 | Using sysctl | 42 |
| 4.5.1 | sysctl -w | 42 |
| 4.5.2 | sysctl -p | 43 |
| 4.6 | Modifying Network Behaviour through /proc and sysctl | 43 |
| 5 | Configuring Logging | 44 |
| 5.1 | Understanding Logging In RHEL 7 | 44 |
| 5.2 | Connecting Journald to Rsyslogd | 44 |
| 5.2.1 | Modules | 45 |
| 5.2.2 | Importing text files to log : httpd error log | 45 |
| 5.2.3 | Exporting data to an output module : exporting to a database | 46 |
| II | Networking and Apache | 47 |
| 6 | Configuring Advanced Networking | 48 |
| 7 | Managing Linux Based Firewalls | 49 |
| 8 | Configuring Apache Virtual Hosts | 50 |

| | | |
|------------|--|-----------|
| 9 | Managing Advanced Apache Features | 51 |
| III | DNS and File Sharing | 52 |
| 10 | Configuring a Cache-only DNS Server | 53 |
| 11 | Configuring NFS File Sharing | 54 |
| 12 | Managing SMB File Sharing | 55 |
| IV | Essential Back-end Services | 56 |
| 13 | Setting up an SMTP Server | 57 |
| 14 | Managing SSH | 58 |
| 15 | Managing MariaDB | 59 |
| 16 | Managing Time Services | 60 |
| 17 | Shell Scripting | 61 |

Part I

Advanced System Management

Chapter 1

Configuring Authentication

1.1 Understanding RedHat Identity Management

RedHat Identity Management is based on the FreeIPA (Identity, Policy, Audit) Project. The project bundles together several services in one solution. Some of the services are:

| Options | Description |
|--------------------------------------|---|
| 389 Directory Server | This is an LDAPv3 Directory Server – a replacement for <i>OpenLDAP</i> . |
| Single Sign-on | Provided by MIT Kerberos KDC. |
| Integrated Certificate System | Based on the <i>Dogtag</i> project. |
| Integrated NTP Server | <i>Chrony</i> must be disabled to use this! |
| Integrated DNS Server | Based on <i>ISC Bind</i> Service. |

Thus, the Identity Management provided by IPA bundles up some pretty complicated projects together and provides an easy interface to manage them all. However, IPA conflicts with other products, such as other *LDAP*, *Kerberos*, *Certificate System*, *NTP* or *DNS* servers shouldn't be running on the same system. Thus, ideally Identity Management should be set up on a dedicated server.

Kerberos is a Network Authentication Protocol that makes clients prove their identity to the server, and vice versa. Other than the authentication tools, it also supports strong cryptography over the network to keep the data safe in-transit.

1.1.1 IdM Server Components and Requirements

An IdM server needs some from of *Host Name Resolution*, which can be either through a DNS server or via the `/etc/hosts` file. Note that the hostname of the Identity Management server itself must also be specified.

Next we need both the **ipa-server** package, which installs the server components, and the **ipa-client** package which installs the client components. While the client package isn't required to be installed on the server, while configuring a client that talks to an IPA server, then this is one of the solutions available. Another method would be to use **authconfig**.

After the required RPM packages have been installed, we will run **ipa-server-install** which provides an easy, scripted way to install an IPA server, and all we have to do is answer a few questions, at the end of which we get a fully-functional IPA server.

The **ipa** tool is a generic client interface, that's also the administration interface. Thus, it can perform several tasks such as adding users (`ipa user-add <username>`), set the password for an user (`ipa passwd <username>`), see the IPA properties for a user account (`ipa user-find <username>`), etc. *ipa-xxx* can be used instead as well, where *xxx* represents the different tasks. Authentication can also be configured using **authconfig**.

1.1.2 Preparing IdM Installation

First and foremost, the *host name resolution* must be set up, since the installation will fail if the host can't find its own name. Additionally, the DNS name must also be known since the Kerberos domain that we'll configure will be based on the DNS name.

Next, the **nscd** service must be disabled, along with any existing LDAP and Kerberos services. If NTP and ISC Bind must also be disabled if installed (due to possible conflicts). The LDAP, Kerberos, NTP, DNS and certificate system ports must be opened in the firewall.

1.1.3 Installing IdM

The **ipa-server**, **bind** and **nds-ldap** packages must be installed using, following which, we have to run the command **ipa-server-install**, which will perform a wizard-like scripted installation.

```
1 # yum -y install ipa-server bind nds-ldap
2 # ipa-server-install
```

If we don't want to enter the information interactively, we can also provide them as options. The hostname, the domain name, a realm name (domain name in upper-case).

```
1 # ipa-server-install --hostname=vmPrime.somuVMnet.local -n somuVMnet.local -r
   ↪ SOMUVMNET.COM -p password -a password -U --no-ntp
```

The appropriate flags needed are:

| Options | Description |
|------------------|---|
| -hostname | The hostname of the server |
| -n | The Domain name of the server |
| -r | Realm Name (Domain name in All-Caps) |
| -p | Password for Directory Manager |
| -a | Password for admin user |
| -U | Unattended Install; Doesn't prompt for anything |
| -no-dns | Do not install the DNS Server |

Now, the SSH Daemon must be restarted to ensure that SSH obtains Kerberos credentials:

```
1 # systemctl restart sshd
```

Then, we generate a new Kerberos ticket and then verify Kerberos authentication for the default admin user by using:

```
1 # kinit admin
2 Password for admin@SOMUVMNET.LOCAL:
```

```
3 # klist
4 Ticket cache: KEYRING:persistent:0:0
5 Default principal: admin@SOMUVMNET.LOCAL
6
7 Valid starting      Expires            Service principal
8 2017-12-26T15:44:09  2017-12-27T15:44:05  krbtgt/SOMUVMNET.LOCAL@SOMUVMNET.LOCAL
```

This will show us if we have a valid Kerberos ticket. For any administrative tasks on the IPA server, having a valid Kerberos ticket is mandatory. Finally, we need to verify IPA access using:

```
1 # ipa user-find admin
```

This will show us the details of the admin user as created in the LDAP directory, along with all of its properties.

1.1.4 Understanding Kerberos Tickets

Kerberos tickets are the keys to the proper functioning of Identity Management. To be able to manage the IdM server, we need to log in to the IdM Domain and generate a Kerberos ticket for the admin user, using the command:

```
1 # kinit admin
```

We can check the validity of the ticket at any time using:

```
1 # klist
```

1.1.5 Managing the IdM Server

After generating a Kerberos ticket with `kinit admin`, we use the **ipa** command to manage the IdM server. `ipa help commands` shows us a short overview of all the available commands and their usage. For any specific command, we have `ipa help <command>` (such as `ipa help user-add`).

Another method to manage the IdM server is to navigate, using our web browser, to <https://vmPrime.somuVMnet.local> (if our server is named *vmPrime.somuVMnet.local*). This will load the IPA management web interface. Through this interface, after we've authenticated as admin, we will be guided through the various aspects of setting up the IdM environment.

1.1.6 Creating User Accounts

The required commands to create an user called *lisa* and verify the account creation are:

```
1 # kinit admin
2 # ipa user-add lisa
3 # ipa passwd lisa
4 # ipa user-find lisa
```

1.2 Using authconfig to Setup External Authentication

There are the **authconfig** utilities to setup external authentication (via LDAP), which consist of: *authconfig*, *authconfig-tui* and *authconfig-gtk*. The GUI utility can be installed using `yum -y install authconfig-gtk`. The utility is started with `authconfig-gtk`.

In the **authconfig-gtk** utility, we have to choose LDAP as the User Account Database in the Identity and Authentication tab. This might prompt for the installation of two packages: *nss-pam-ldapd* (the package that integrates the three) and *pam_krb5* (the package that integrates PAM with Kerberos). Now, we can enter the details for the LDAP server to setup authentication.

In cases of servers which don't have a GUI (or there is some inconvenience with the GUI, such as the apply button hidden by the status bar, etc.), the **authconfig-tui** is a very good alternative. In case of automated scripts, however, the **authconfig** command line utility is the best option.

1.3 Configuring a System to Authenticate using Kerberos

To connect a system for authentication to an LDAP server using Kerberos credentials, a part of the configuration has to be done with `authconfig`. But even before that, certain things must be ensured. *First*, we need to make sure that the IP address of the server we're trying to connect to can be resolved from the hostname, using `/etc/hosts`:

```
1 127.0.0.1    localhost localhost.localdomain localhost4 localhost4.localdomain4
2 ::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
3 90.0.16.100  vmDeux.somuVMnet.com      vmDeux
```

This is important so that we can use the FQDN of the server later while using the `authconfig-tui` utility. Next, the system must be configured to use the DNS component hosted within the IPA server. For this, all we need to do is add the IP address of the IPA server as the first nameserver entry in `/etc/resolv.conf`:

```
1 # Generated by NetworkManager
2 search somuvmnet.local
3 nameserver 90.0.16.100 # IP Address of DNS Server @ vmDeux.somuVMnet.local
4 nameserver 8.8.8.8
5 nameserver 202.38.180.7
```

It is important to place the IP address of the DNS server for a nameserver as the first entry because that's the only one configured to *know* the custom FQDNs of the machines on our network. So, this connectivity is essential since the Kerberos client needs to be connected to the Kerberos server.

Finally, we can start the `authconfig-tui` utility, and enter the following details:

```
1 # In Authentication Configuration:
2 [*] Use LDAP
3 [*] User Kerberos Authentication
4
5 # LDAP Settings
6 [*] Use TLS
7 Server:      ldap://vmdeux.somuvmnet.local
8 Base DN:     dc=somuvmnet, dc=local
```

```
9
10 # Kerberos Settings
11 Realm:          SOMUVMN.NET.LOCAL
12 [*] Use DNS to resolve hosts to realms
13 [*] Use DNS to resolve KDCs for realms
```

First, we've just setup the system to use LDAP using Kerberos authentication. Next, we've made it necessary to use a TLS certificate to ensure the security of the connection. Then, the details of the LDAP server have to be entered.

In the Kerberos authentication step, the *Realm* refers to the Kerberos realm that the server is a part of. If we've setup the DNS component of the IPA server properly, then the system is able to detect the KDCs properly for each realm, as well as assign hosts to their realm appropriately. Now, the TLS Certificate for the IPA Server have to be downloaded and put in the `/etc/openldap/cacerts` directory (from whichever location the IPA Server stored them in, typically `/root/cacert.p12` for the root user):

```
1 # cd /etc/openldap/cacerts/
2 # scp vmdeux.somuvmn.net:/root/cacert.p12 .
```

At this point, we should be good to go. We can verify the LDAP connectivity by trying to login as an LDAP user. For this we use (for an LDAP user lisa):

```
1 # su - lisa
2 Last login: Tue Dec 26 18:52:05 IST 2017 on pts/0
3 su: warning: cannot change directory to /home/lisa: No such file or directory
4 -sh-4.2$
```

The warning is natural if no home directory has been configured yet.

1.3.1 Troubleshooting Authentication

When authentication doesn't work, for some reason related to the certificates, then there is an easy fix as well. Depending on whether our LDAP and Kerberos credentials are being cached by **nsld** or **sssd**, we can edit their configuration file to ignore the validity of the certificate. This is because the *self-signed cacert* may not meet the standards dictated and required by the program. For this, we can add to `/etc/nsld.conf`:

```
1 tls_reqcert never
```

If SSSD is used instead, then we can edit `/etc/sss/sss.conf` and add the following line:

```
1 ldap_tls_reqcert = never
```

When using Certificates that are well signed from an External Certificate Authority, this of course becomes unnecessary.

1.4 Understanding authconfig Configuration Files

1.4.1 Authconfig Configuration

The primary configuration of the authconfig utility is located at `/etc/sysconfig/authconfig`. The contents of this file is used by other config files, such as `USELDAP=yes`.

```
1  CACHECREDENTIALS=yes
2  FAILLOCKARGS="deny=4 unlock_time=1200"
3  FORCELEGACY=no
4  FORCESMARTCARD=no
5  IPADOMAINJOINED=no
6  IPAV2NONTF=no
7  PASSWDALGORITHM=sha512
8  USEDB=no
9  USEECRYPTFS=no
10 USEFAILLOCK=no
11 USEFPRINTD=no
12 USEHESIOD=no
13 USEIPAV2=no
14 USEKERBEROS=yes
15 USELDAP=yes
16 USELDAPAUTH=no
17 USELOCAUTHORIZE=yes
18 USEMKHOMEDIR=no
19 USENIS=no
20 USEPAMACCESS=no
21 USEPASSWDQC=no
22 USEPWQUALITY=yes
23 USESHADOW=yes
24 USESMARTCARD=no
25 USESSSD=yes
26 USESSSDAUTH=no
27 USESYSNETAUTH=no
28 USEWINBIND=no
29 USEWINBINDAUTH=no
30 WINBINDKRB5=no
```

These are the settings we provided to the **authconfig** utility.

1.4.2 SSSD Configuration

Things like the Kerberos password, the LDAP search base, etc. and other IPA specific settings are stored in the `/etc/sss/sss.conf` file, to ensure that the connection to the IPA Server is successfully initiated and it's possible to login and use the services provided by it. Typical contents of this file look like:

```
1  [sss]
2  config_file_version = 2
3  services = nss, pam
4  # SSSD will not start if you do not configure any domains.
5  # Add new domain configurations as [domain/<NAME>] sections, and
6  # then add the list of domains (in the order you want them to be
7  # queried) to the "domains" attribute below and uncomment it.
8  ; domains = LDAP
9
```

```

10  [nss]
11
12  [pam]
13
14  # Example LDAP domain
15  ; [domain/LDAP]
16  ; id_provider = ldap
17  ; auth_provider = ldap
18  # ldap_schema can be set to "rfc2307", which stores group member names in the
19  # "memberuid" attribute, or to "rfc2307bis", which stores group member DNs in
20  # the "member" attribute. If you do not know this value, ask your LDAP
21  # administrator.
22  ; ldap_schema = rfc2307
23  ; ldap_uri = ldap://ldap.mydomain.org
24  ; ldap_search_base = dc=mydomain,dc=org
25  # Note that enabling enumeration will have a moderate performance impact.
26  # Consequently, the default value for enumeration is FALSE.
27  # Refer to the sssd.conf man page for full details.
28  ; enumerate = false
29  # Allow offline logins by locally storing password hashes (default: false).
30  ; cache_credentials = true
31
32  # An example Active Directory domain. Please note that this configuration
33  # works for AD 2003R2 and AD 2008, because they use pretty much RFC2307bis
34  # compliant attribute names. To support UNIX clients with AD 2003 or older,
35  # you must install Microsoft Services For Unix and map LDAP attributes onto
36  # msSFU30* attribute names.
37  ; [domain/AD]
38  ; id_provider = ldap
39  ; auth_provider = krb5
40  ; chpass_provider = krb5
41  ;
42  ; ldap_uri = ldap://your.ad.example.com
43  ; ldap_search_base = dc=example,dc=com
44  ; ldap_schema = rfc2307bis
45  ; ldap_sasl_mech = GSSAPI
46  ; ldap_user_object_class = user
47  ; ldap_group_object_class = group
48  ; ldap_user_home_directory = unixHomeDirectory
49  ; ldap_user_principal = userPrincipalName
50  ; ldap_account_expire_policy = ad
51  ; ldap_force_upper_case_realm = true
52  ;
53  ; krb5_server = your.ad.example.com
54  ; krb5_realm = EXAMPLE.COM

```

This is probably one of the most important configuration files when **SSSD** is being used. If **nsld** is being used instead, then the config file of interest is `/etc/nsld.conf`.

1.4.3 Kerberos Configuration File

The Kerberos configuration file (for connecting to a Kerberos Server) is stored in `/etc/krb5.conf` and typically has contents like:

```

1  # Configuration snippets may be placed in this directory as well
2  includedir /etc/krb5.conf.d/
3
4  includedir /var/lib/sss/pubconf/krb5.include.d/

```

```

5  [logging]
6  default = FILE:/var/log/krb5libs.log
7  kdc = FILE:/var/log/krb5kdc.log
8  admin_server = FILE:/var/log/kadmind.log
9
10 [libdefaults]
11 dns_lookup_realm = true
12 ticket_lifetime = 24h
13 renew_lifetime = 7d
14 forwardable = true
15 rdns = false
16 # default_realm = EXAMPLE.COM
17 default_ccache_name = KEYRING:persistent:%{uid}
18
19 dns_lookup_kdc = true
20 default_realm = SOMUVMNET.LOCAL
21 [realms]
22 # EXAMPLE.COM = {
23 #   kdc = kerberos.example.com
24 #   admin_server = kerberos.example.com
25 # }
26
27 SOMUVMNET.LOCAL = {
28 }
29
30 [domain_realm]
31 # .example.com = EXAMPLE.COM
32 # example.com = EXAMPLE.COM
33 somuvmnnet.local = SOMUVMNET.LOCAL
34 .somuvmnnet.local = SOMUVMNET.LOCAL

```

Here, the DNS domain to realm mapping is specified, to tell us which domain on the DNS belongs to which Kerberos realm.

1.4.4 NSSwitch Configuration

This file specifies the locations and the order in which passwords are searched for authentication. This includes the order in which passwords, shadow and groups are searched. The order is typically like:

```

1 passwd:      files sss ldap
2 shadow:      files sss ldap
3 group:       files sss ldap

```

This instructs the system to look for passwd files in the local file system first, then SSS and finally LDAP. The same is true for the two following categories of shadow and group.

1.4.5 NSLCD Configuration

While this file may be missing from newer versions of RHEL, this is an older version of LDAP configuration file. This file is supposed to be replaced by the `/etc/sss/sss.conf` file, and thus, all relevant settings should be provided in that file.

Chapter 2

Configuring iSCSI Target and Initiator

2.1 Understanding iSCSI Target and Initiator

SCSI (Small Computer System Interface) [read as *scuzzy*] is an alternative to ATA (a.k.a. IDE) Hard drives, which most consumer computers stick to. While SCSI drives provide significantly more throughput for certain scenarios, IDE suffices for most home computer usage. However, in case of servers, SCSI proves to be a much better alternative, since they provide more reliability and data transfer speed (much higher than ATA), owing to the fact that data transfer occurs in full-duplex mode (i.e., data can be read and written at the same time at full speeds). They also boast higher speeds (such as 15,000 RPM) as compared to ATA speeds (7200 RPM). Another reason servers tend to use SCSI (or related technologies, such as Serially Attached SCSI or SAS) is that the protocol makes it easy to *daisy-chain* several SCSI devices to the same controller, several times that of IDE devices. In fact, in the pre-USB era, SCSI was the go-to common interface for connecting peripherals or even devices such as printers.

Traditional SCSI devices use a long cable and a SCSI **Command Descriptor Block (CDB)** command to interact with the SCSI devices. In case of iSCSI, the same CDBs are used, but they're transmitted over IP packets over a network, instead of the cable. Thus, the SCSI devices are emulated by using a storage backend and presenting them on the network using iSCSI targets. SCSI targets are typically storage devices, while the hosts they're connected to are the initiators. Thus, this technology enables us to share PVs or LVs on the network, represented by iSCSI targets.

A **Storage Area Network (SAN)** is a network that provides access to a consolidated, block level data storage. *Block devices* provide a buffered data storage method, where data is transferred from the kernel buffer to the physical device. Also, data can be read and written in entire blocks. SANs thus present devices such as disk arrays as locally attached storage to servers. **Fiber Channel** or **FC** is a high speed network technology developed to enable fast data transfers between servers and SANs. Ethernet structures utilizing iSCSI technology can be as fast as their FC structure counterparts, thus making the technology enterprise ready for SAN creation.

2.1.1 iSCSI Operation

In the case of iSCSI storage, we have the SAN, on which runs a *iSCSI target* which can provide access to the storage backend. For any server that needs to access the files hosted

by the SAN, it needs to run an **iSCSI initiator**, which performs a discovery operation first. During this, the SAN tells it about the iSCSI devices it has to offer. Once this is complete, the iSCSI initiator can login to the devices.

2.1.2 iSCSI Components

Both the iSCSI targets and the storage backends need to be set up for the SAN to operate. The storage backend can be an entire disk, a dedicated partition, a logical volume or even a file! The servers, running the iSCSI initiators, will see the iSCSI targets as new storage devices after successfully logging in to them. This can be verified by viewing the output of the `/proc/partitions` file. A tool called `/sscsi` can also alternatively used, although it is not installed by default.

2.1.3 Basic iSCSI Terminology

| Terms | Description |
|------------------|---|
| IQN | iSCSI Qualified Name - an unique name assigned to each iSCSI target and initiator, used to identify them. |
| Initiator | The iSCSI client that is identified by its IQN. |
| Target | The service on the iSCSI server that gives access to the storage backend. |
| ACLs | Access Control Lists that are based on the node's IQNs. |
| Portal | Also known as nodes , this is the combination of the IP address and the port that are used by both targets and initiators to establish connections. |
| discovery | The process through which an indicator finds the available targets that are configured for a given portal. |
| LUNs | The <i>Logical Unit Number</i> is a number used to identify the logical unit (i.e., block devices shared through the target) being addressed by the iSCSI Controller. |
| login | The act which gives an initiator the relevant LUNs. |
| TPG | The <i>Target Portal Group</i> is a collection of IP Addresses and TCP ports to which a particular iSCSI Target will listen. |

So, there can be more than one portals per server, and more than one targets per portal.

2.1.4 After connecting an initiator to an iSCSI Target

The new block devices thus accessed will appear as local devices (`/dev/sdb`, `/dev/sdc` etc.) Note that if a LUN is available and used by multiple servers, multiple devices can access the LUN post connection, i.e., multiple servers can use the disk at the same time. This is a bit dangerous, since it requires using clustering, for providing multiple servers to use the storage. Otherwise for a file system like XFS or Ext4, two servers writing to the same file can cause data loss.

To avoid this, shared file systems such as GFS2 can be used. In GFS2, the file system cache is shared among all the nodes. Thus, all nodes writing to the file system know what all the other nodes are doing.

2.2 Setting up an iSCSI Target

The iSCSI target works with several storage backend devices on the SAN. These storage devices can be anything that can be used for PVs when using traditional LVM. All these devices are put together in a volume group, which is then subdivided into several LVs. These LVs are each assigned a LUN.

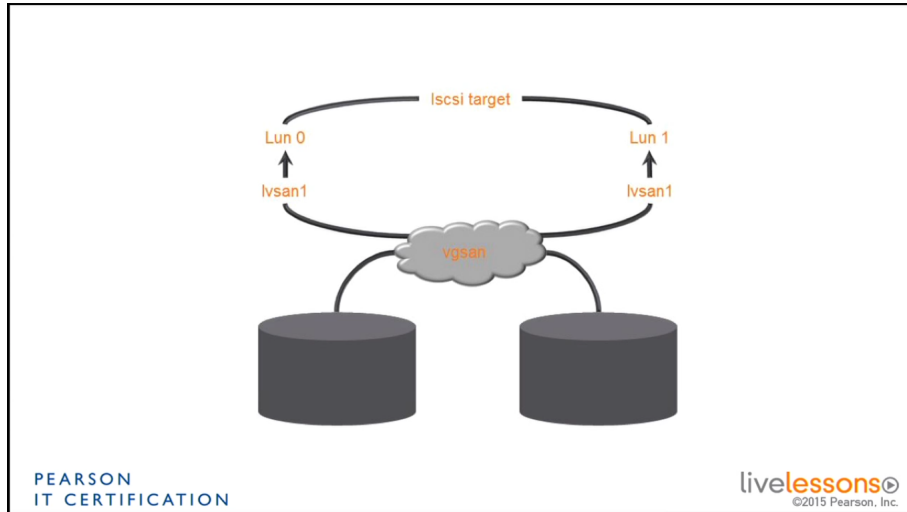


Figure 2.1: iSCSI Target Setup

These LUNs are presented using the iSCSI targets. Thus, the iSCSI configuration is created on top of a traditional LVM configuration.

2.2.1 Creating the LVM

Let us consider we have an empty disk of 1GB on which we want to build the iSCSI configuration. This can be verified using:

```
1 # lsblk
2 NAME        MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
3 sda          8:0    0   20G  0 disk
4 └─sda1       8:1    0    1G  0 part /boot
5 └─sda2       8:2    0   19G  0 part
6 └─centos-root 253:0    0   17G  0 lvm  /
7 └─centos-swap 253:1    0    2G  0 lvm  [SWAP]
8 sdb          8:16    0    1G  0 disk
9 sr0         11:0    1   8.1G  0 rom
```

We can directly create the VG `vgsan` on it, using:

```
1 # vgcreate vgSAN /dev/sdb
2 Physical volume "/dev/sdb" successfully created.
3 Volume group "vgSAN" successfully created
4 # pvs
5 PV          VG      Fmt  Attr PSize   PFree
6 /dev/sda2   centos lvm2 a--  <19.00g    0
7 /dev/sdb    vgSAN  lvm2 a--  1020.00m 1020.00m
8 # vgs
```

```

 9  VG      #PV #LV #SN Attr   VSize   VFree
10  centos    1  2  0 wz--n-   <19.00g     0
11  vgSAN     1  0  0 wz--n-  1020.00m 1020.00m

```

The output of the `pvs` and `vgs` commands show that the PV `/dev/sdb` is now a part of `vgSAN`, which has a free space of 1020MB. Now we create two LVs `lvSAN1` and `lvSAN2` on the VG, using:

```

 1  # lvcreate -L 500M -n lvSAN1 vgSAN
 2  Logical volume "lvSAN1" created.
 3  # lvcreate -l 100%FREE -n lvSAN2 vgSAN
 4  Logical volume "lvSAN2" created.
 5  # lvs
 6  LV      VG      Attr      LSize   Pool Origin ... Convert
 7  root    centos  -wi-ao---- <17.00g
 8  swap    centos  -wi-ao---- 2.00g
 9  lvSAN1   vgSAN   -wi-a----- 500.00m
10  lvSAN2   vgSAN   -wi-a----- 520.00m

```

Now, our LVM setup is complete, and we can proceed with the iSCSI setup. For this, first of all we need to install the iSCSI software, called **targetcli**. The `targetcli` utility is a relatively new one capable of managing multiple types of storage devices.

2.2.2 Creating the iSCSI configuration using targetcli

We start the utility using:

```

 1  # targetcli
 2  Warning: Could not load preferences file /root/.targetcli/prefs.bin.
 3  targetcli shell version 2.1.fb46
 4  Copyright 2011-2013 by Datera, Inc and others.
 5  For help on commands, type 'help'.
 6
 7  />

```

This interface can be navigated using the same commands as the bash shell. Using the `cd` command produces the output:

```

 1  /> ls
 2  o- / ..... [..]
 3    o- backstores ..... [..]
 4      | o- block ..... [Storage Objects: 0]
 5      | o- fileio ..... [Storage Objects: 0]
 6      | o- pscsi ..... [Storage Objects: 0]
 7      | o- ramdisk ..... [Storage Objects: 0]
 8    o- iscsi ..... [Targets: 0]
 9    o- loopback ..... [Targets: 0]
10  />

```

The *backstores* part allow us to work with the different storage devices. To enter backstores, we simply enter `cd` command, and select it from the menu. This will change the prompt to `/backstores>`. Here, we can see it contains the *block*, the *fileio*, the *pSCSI* and the *ramdisk* devices. Their significance is explained below:

| Types | Description |
|----------------|--|
| Block | Refers to any block device that we want to share using iSCSI. This includes all traditional disks, partitions and even LVMs. |
| fileio | Refers to a file that can be used as a storage source. This refers to a big file created using a tool such as dd. |
| pscsi | Physical SCSI - a SCSI pass-through backstore is created for such devices. |
| ramdisk | RAM storage, wiped with every reboot, and is thus a very bad idea. |

Now, since all our LVs are block devices (by their very nature), we have to create our LUNs inside the block category. This we can do using:

```

1 /backstores> block/ create block1 /dev/vgSAN/lvSAN1
2 Created block storage object block1 using /dev/vgSAN/lvSAN1.
```

The command instructs the targetcli utility to enter the block category, and create a block device called *block1* from the */dev/vgSAN/lvSAN1* device. We can create another block device for the partition and a 1G custom file device using:

```

1 /backstores> block/ create block2 /dev/vgSAN/lvSAN2
2 Created block storage object block2 using /dev/vgSAN/lvSAN2.
3 /backstores> fileio/ create file1 /root/diskFile1 1G
4 Created fileio file1 with size 1073741824
```

When creating a file, we can merely specify the size (1GB) and the name & location (*/root/diskFile1*) to have the *targetcli* utility create the file for us, instead of copying from */dev/zero* to a file using *dd*. All the different devices thus added can be seen with:

```

1 /backstores> ls
2 o- backstores ..... [....]
3   o- block ..... [Storage Objects: 2]
4     | o- block1 ..... [/dev/vgSAN/lvSAN1 (500.0MiB) write-thru deactivated]
5     | | o- alua ..... [ALUA Groups: 1]
6     | |   o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
7     | o- block2 ..... [/dev/vgSAN/lvSAN2 (520.0MiB) write-thru deactivated]
8     |   o- alua ..... [ALUA Groups: 1]
9     |     o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
10    o- fileio ..... [Storage Objects: 1]
11      | o- file1 ..... [/root/diskFile1 (1.0GiB) write-back deactivated]
12      |   o- alua ..... [ALUA Groups: 1]
13      |     o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
14    o- pscsi ..... [Storage Objects: 0]
15    o- ramdisk ..... [Storage Objects: 0]
```

Now that the block devices are ready, we can go to the */iscsi* environment and prepare the iSCSI targets. Initially, there will be no targets:

```

1 /backstores> cd /iscsi
2 /iscsi> ls
3 o- iscsi ..... [Targets: 0]
```

2.2.3 Target Creation

Now, we create a target that provides access to the backing storage devices called *block1*, *block2* and *file1*. This can be done using the *create* command, followed by an IQN. IQNs

are typically created using a naming format:

`iqn.<yearOfCreation>-<monthOfCreation>.<reverseDomainName>:<targetName>`

Thus, ours will be named: `iqn.2018-01.local.somuvmmnet:target1`. This can be done with:

```
1 /iscsi> create iqn.2018-01.local.somuvmmnet:target1
2 Created target iqn.2018-01.local.somuvmmnet:target1.
3 Created TPG 1.
4 Global pref auto_add_default_portal=true
5 Created default portal listening on all IPs (0.0.0.0), port 3260.
```

Thus, both a target and a TPG are created at the same time. The target thus created can be viewed with:

```
1 /iscsi> ls
2 o- iscsi ..... [Targets: 1]
3   o- iqn.2018-01.local.somuvmmnet:target1 ..... [TPGs: 1]
4     o- tpg1 ..... [no-gen-acls, no-auth]
5       o- acls ..... [ACLs: 0]
6       o- luns ..... [LUNs: 0]
7       o- portals ..... [Portals: 1]
8         o- 0.0.0.0:3260 ..... [OK]
```

2.2.4 TPG Configuration

Within the target is a TPG (Target portal group), which represents the entire configuration of the target. This includes all the ACLs, the LUNs and the portals related to the target.

ACLs

Next, we need to create the ACLs for our target. For this, we need to `cd` into the ACL environment of our target using (Note that tab-autocompletion works for this tool):

```
1 /iscsi> cd iqn.2018-01.local.somuvmmnet:target1/tpg1/acls
2 /iscsi/iqn.20...et1/tpg1/acls>
```

We create the ACL node using:

```
1 /iscsi/iqn.20...et1/tpg1/acls> create iqn.2018-01.local.somuvmmnet:vmdeux
2 Created Node ACL for iqn.2018-01.local.somuvmmnet:vmdeux
```

Note that the identifier provided to create the node ACL is the IQN that has been set on the second server. The structure now looks like:

```
1 /iscsi/iqn.20...et1/tpg1/acls> cd /iscsi
2 /iscsi> ls
3   o- iscsi ..... [Targets: 1]
4     o- iqn.2018-01.local.somuvmmnet:target1 ..... [TPGs: 1]
5       o- tpg1 ..... [no-gen-acls, no-auth]
6         o- acls ..... [ACLs: 1]
7           | o- iqn.2018-01.local.somuvmmnet:vmdeux ..... [Mapped LUNs: 0]
8         o- luns ..... [LUNs: 0]
9         o- portals ..... [Portals: 1]
10        o- 0.0.0.0:3260 ..... [OK]
```

LUNs

Now, inside the *tpg1* node, we create a LUN by using:

```
1 /iscsi/iqn.20...:target1/tpg1> luns/ create /backstores/block/block1
2 Created LUN 0.
3 Created LUN 0->0 mapping in node ACL iqn.2018-01.local.somuvmmnet:vmdeux
```

Now, we can repeat the command a couple of times to create the LUNs for *block2* and *file1* as well:

```
1 /iscsi/iqn.20...:target1/tpg1> luns/ create /backstores/block/block2
2 Created LUN 1.
3 Created LUN 1->1 mapping in node ACL iqn.2018-01.local.somuvmmnet:vmdeux
4 /iscsi/iqn.20...:target1/tpg1> luns/ create /backstores/fileio/file1
5 Created LUN 2.
6 Created LUN 2->2 mapping in node ACL iqn.2018-01.local.somuvmmnet:vmdeux
```

The contents of *tpg1* should now look like:

```
1 /iscsi/iqn.20...:target1/tpg1> ls
2 o- tpg1 ..... [no-gen-acls, no-auth]
3   o- acls ..... [ACLs: 1]
4     | o- iqn.2018-01.local.somuvmmnet:vmdeux ..... [Mapped LUNs: 3]
5       |   o- mapped_lun0 ..... [lun0 block/block1 (rw)]
6         |   o- mapped_lun1 ..... [lun1 block/block2 (rw)]
7           |   o- mapped_lun2 ..... [lun2 fileio/file1 (rw)]
8     o- luns ..... [LUNs: 3]
9       | o- lun0 ..... [block/block1 (/dev/vgSAN/lvSAN1) (default_tg_pt_gp)]
10      | o- lun1 ..... [block/block2 (/dev/vgSAN/lvSAN2) (default_tg_pt_gp)]
11      | o- lun2 ..... [fileio/file1 (/root/diskFile1) (default_tg_pt_gp)]
12     o- portals ..... [Portals: 1]
13       o- 0.0.0.0:3260 ..... [OK]
```

We can see that not only have the LUNs been created, but they've been assigned to the ACL as well! Thus, it becomes critical to create ACLs before the LUNs because the default behaviour of *targetcli* is to automatically assign any LUN that's been created to the ACLs in the TPG. Now, we have to create a portal.

Portals

We can create portal which will bear the IP address of the server on which our SAN will advertise the LUNs for this particular target. We do this by:

```
1 iscsi/iqn.20...:target1/tpg1> portals/ create 90.0.16.27
2 Using default IP port 3260
```

The complete configuration of the iSCSI setup can be viewed with:

```
1 /iscsi/iqn.20...:target1/tpg1> cd
2 /> ls
3 o- / ..... [...]
4   o- backstores ..... [...]
5     | o- block ..... [Storage Objects: 2]
6       | | o- block1 ..... [/dev/vgSAN/lvSAN1 (500.0MiB) write-thru activated]
```

```

7 | | | o- alua ..... [ALUA Groups: 1]
8 | | |   o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
9 | | o- block2 ..... [/dev/vgSAN/lvSAN2 (520.0MiB) write-thru activated]
10 | |   o- alua ..... [ALUA Groups: 1]
11 | |     o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
12 | o- fileio ..... [Storage Objects: 1]
13 | | o- file1 ..... [/root/diskFile1 (1.0GiB) write-back activated]
14 | |   o- alua ..... [ALUA Groups: 1]
15 | |     o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
16 | o- pscsi ..... [Storage Objects: 0]
17 | o- ramdisk ..... [Storage Objects: 0]
18 o- iscsi ..... [Targets: 1]
19 | o- iqn.2018-01.local.somuvmmnet:target1 ..... [TPGs: 1]
20 |   o- tpg1 ..... [no-gen-acls, no-auth]
21 |     o- acls ..... [ACLs: 1]
22 |       | o- iqn.2018-01.local.somuvmmnet:vmdeux ..... [Mapped LUNs: 3]
23 |         | o- mapped_lun0 ..... [lun0 block/block1 (rw)]
24 |         | o- mapped_lun1 ..... [lun1 block/block2 (rw)]
25 |         | o- mapped_lun2 ..... [lun2 fileio/file1 (rw)]
26 |       o- luns ..... [LUNs: 3]
27 |         | o- lun0 ..... [block/block1 (/dev/vgSAN/lvSAN1) (default_tg_pt_gp)]
28 |         | o- lun1 ..... [block/block2 (/dev/vgSAN/lvSAN2) (default_tg_pt_gp)]
29 |         | o- lun2 ..... [fileio/file1 (/root/diskFile1) (default_tg_pt_gp)]
30 |       o- portals ..... [Portals: 1]
31 |         o- 0.0.0.0:3260 ..... [OK]
32 o- loopback ..... [Targets: 0]

```

2.2.5 Adding a rule to the firewall

Now, we need to allow the TCP connections through port 3260 to use for SAN, using:

```

1 # firewall-cmd --add-port=3260/tcp --permanent
2 success
3 # firewall-cmd --reload
4 success

```

2.2.6 Starting target.service

Even though **targetcli** saves the present configuration to disk, a service called *target.service* must be enabled to ensure that the saved configuration is loaded each time after reboots. This is done with:

```

1 # systemctl start target
2 # systemctl enable target
3 Created symlink from /etc/systemd/system/multi-user.target.wants/target.service to
   ↳ /usr/lib/systemd/system/target.service.
4 # systemctl status target
5 ● target.service - Restore LIO kernel target configuration
6 Loaded: loaded (/usr/lib/systemd/system/target.service; enabled; vendor preset: disabled)
7 Active: active (exited) since Tue 2018-01-02 16:28:20 IST; 25s ago
8 Main PID: 4291 (code=exited, status=0/SUCCESS)
9
10 Jan 02 16:28:19 vmprime.somuvmmnet.local systemd[1]: Starting Restore LIO kern...
11 Jan 02 16:28:20 vmprime.somuvmmnet.local systemd[1]: Started Restore LIO kerne...

```

This particular services instructs the kernel of its responsibilities as a SAN server and how the iSCSI targets are configured, so that it can accept incoming connections from iSCSI initiators and act accordingly.

2.3 Connecting the iSCSI Initiator to an iSCSI SAN

Now that the iSCSI SAN server is setup, we need an iSCSI initiator on a different (remote) server that can use the SAN. For this, the very first requirement is to obtain the software in the `iscsi-initiator-utils` package, which help in creating the initiator. We do this by using:

```
1 # yum -y install iscsi-initiator-utils
```

2.3.1 Setting up an initiator name

Next, we need to setup an **initiator name**, which must be the one we used in the ACL for the iSCSI *target*. To do this, we edit the `/etc/iscsi/initiatorname.iscsi` file. It's contents should be:

```
1 InitiatorName=iqn.2018-02.com.somuVMnet:vmPrime
```

The iSCSI configuration file is located at `/etc/iscsi/iscsid.conf` and this can be used to optimize the iSCSI configuration for the server.

2.3.2 iscsiadm Command

Now, we're going to set up the initiator using the `iscsiadm` command. The syntax of this command can be a bit cryptic, and thus it's recommended to use the man page's example section for a jump start on the syntax of the command.

The primary purpose of the **iscsiadm** command is to discover iSCSI targets and login to them, as well as access and manage the open-iscsi configuration database.

2.3.3 Discovery

Now, the initiator is ready for discovery. For this, we use the command:

```
1 # iscsiadm --mode discoverydb --type sendtargets --portal 10.0.99.12 --discover
2 10.0.99.12:3260,1 iqn.2018-02.local.somuvmmnet:vmdeux
```

The `--mode discoverydb` option instructs the `iscsiadm` command to operate on the *discoverydb* section of the configuration database. It can also be abbreviated to `-m discoverydb`. The `--type sendtargets` (or `-t st`) tells the action to perform, i.e., send a list of targets. The `--portal 10.0.99.12` (`-p 10.0.99.12`) specifies the portal to be used for the action. Finally, the `--discovery (-D)` flag tells the command to perform discovery and add records if necessary. The output returned is a list of the targets on that particular portal. The most important piece of information here is the IQN of the relevant target.

2.3.4 Login

The login is performed on a particular IQN at a particular portal/node. This is achieved using:

```
1 # iscsiadm --mode node --targetname iqn.2018-02.local.somuVMnet:vmDeux --portal
   ↪ 10.0.99.12 --login
2 Logging in to [iface: default, target: iqn.2018-02.local.somuvmmnet:vmdeux, portal:
   ↪ 10.0.99.12,3260] (multiple)
3 Login to [iface: default, target: iqn.2018-02.local.somuvmmnet:vmdeux, portal:
   ↪ 10.0.99.12,3260] successful.
```

Now the *mode* has been changed to **node** since we're dealing with a particular portal to login. The `--targetname iqn.2018-02.local.somuVMnet:vmDeux` option can be shortened to `-T iqn.2018-02.local.somuVMnet:vmDeux` and the portal can have a port specified with `-p 10.0.99.12:3260`.

Note that if the iqname of the initiator hasn't been set properly then login won't succeed with a failure due to authentication message. In that case, probably the IQN of the initiator hasn't been set properly. We need to edit the `/etc/iscsi/initiatorname.iscsi` file again and ensure it's identical to that in the ACL of the target. After the change, the **iscsid** service needs to be restarted for the new IQN to be used by the initiator: `systemctl restart iscsid`.

The presence of the new partitions as locally connected devices can be verified using:

```
1 # cat /proc/partitions
2 major minor #blocks name
3
4 8          0   10485760 sda
5 8          1    1048576 sda1
6 8          2   9436160 sda2
7 11         0    3963904 sr0
8 253        0    8384512 dm-0
9 253        1    1048576 dm-1
10 8          16    520192 sdb
11 8          32    520192 sdc
12 8          48     10240 sdd
13 [root@vmPrime ~]# lsblk
14 NAME                                MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
15 sda                                8:0    0   10G  0 disk
16 └─sda1                            8:1    0    1G  0 part /boot
17 └─sda2                            8:2    0    9G  0 part
18     └─rhel-root 253:0    0    8G  0 lvm  /
19         └─rhel-swap 253:1  0    1G  0 lvm  [SWAP]
20 sdb                                8:16    0   508M  0 disk
21 sdc                                8:32    0   508M  0 disk
22 sdd                                8:48    0    10M  0 disk
```

The sdb, sdc and sdd devices are all LUNs on the iSCSI target. The `lsscsi` tool provides the iSCSI information for the target in even greater depth:

```
1 # lsscsi
2 [0:0:0:0] disk VMware, VMware Virtual S 1.0 /dev/sda
3 [4:0:0:0] cd/dvd NECVMWar VMware SATA CD01 1.00 /dev/sr0
4 [40:0:0:0] disk LIO-ORG block1 4.0 /dev/sdb
5 [40:0:0:1] disk LIO-ORG block2 4.0 /dev/sdc
6 [40:0:0:2] disk LIO-ORG file1 4.0 /dev/sdd
```

All the details of this connection to the IQN via the node is stored in the file `/var/lib/iscsi/nodes/iqn.2018-02.local.somuvmmnet:vmdeux/10.0.99.12,3260,1/default`. Once logged in, after every reboot, the iSCSI initiator will automatically login to the SAN and present the LUNs of the target as locally mounted devices. To prevent that, we need to explicitly logout.

2.3.5 Logout

The logout operation needs the exact same parameters to be passed, other than `--login` which of course gets changed to `--logout`.

```
1 # iscsiadm -m node -T iqn.2018-02.local.somuvmmnet:vmdeux -p 10.0.99.12 --logout
2 Logging out of session [sid: 7, target: iqn.2018-02.local.somuvmmnet:vmdeux, portal:
   ↪ 10.0.99.12,3260]
3 Logout of [sid: 7, target: iqn.2018-02.local.somuvmmnet:vmdeux, portal: 10.0.99.12,3260]
   ↪ successful.
```

2.3.6 Deleting node information

To delete all the information pertaining to an iSCSI target we use:

```
1 # iscsiadm -m node -T iqn.2018-02.local.somuvmmnet:vmdeux -o delete
```

Another option would be to delete the folder with the IQN name of the target from `/var/lib/iscsi/nodes/`.

2.4 Verifying the iSCSI Connection

2.4.1 Verification on the iSCSI Initiator

To verify the iSCSI connection we use the **iscsiadm** command. The `-P` command is used to specify the print-level which means that the information is shown as a tree of varying levels of information (the higher the print level, more information is given).

To verify the iSCSI connection, we need information about the session, acquired using:

```
1 # iscsiadm -m session -P 1
2 Target: iqn.2018-02.local.somuvmmnet:vmdeux (non-flash)
3 Current Portal: 10.0.99.12:3260,1
4 Persistent Portal: 10.0.99.12:3260,1
5 *****
6 Interface:
7 *****
8 Iface Name: default
9 Iface Transport: tcp
10 Iface Initiatorname: iqn.2018-02.local.somuvmmnet:vmprime
11 Iface IPaddress: 10.0.99.11
12 Iface Hwaddress: <empty>
13 Iface Netdev: <empty>
14 SID: 8
```

```

15  iSCSI Connection State: LOGGED IN
16  iSCSI Session State: LOGGED_IN
17  Internal iscsid Session State: NO CHANGE
18  # iscsiadm -m session -P 2
19  Target: iqn.2018-02.local.somuvmmnet:vmdeux (non-flash)
20  Current Portal: 10.0.99.12:3260,1
21  Persistent Portal: 10.0.99.12:3260,1
22  *****
23  Interface:
24  *****
25  Iface Name: default
26  Iface Transport: tcp
27  Iface Initiatorname: iqn.2018-02.local.somuvmmnet:vmprime
28  Iface IPaddress: 10.0.99.11
29  Iface HWaddress: <empty>
30  Iface Netdev: <empty>
31  SID: 8
32  iSCSI Connection State: LOGGED IN
33  iSCSI Session State: LOGGED_IN
34  Internal iscsid Session State: NO CHANGE
35  *****
36  Timeouts:
37  *****
38  Recovery Timeout: 120
39  Target Reset Timeout: 30
40  LUN Reset Timeout: 30
41  Abort Timeout: 15
42  *****
43  CHAP:
44  *****
45  username: <empty>
46  password: *****
47  username_in: <empty>
48  password_in: *****
49  *****
50  Negotiated iSCSI params:
51  *****
52  HeaderDigest: None
53  DataDigest: None
54  MaxRecvDataSegmentLength: 262144
55  MaxXmitDataSegmentLength: 262144
56  FirstBurstLength: 65536
57  MaxBurstLength: 262144
58  ImmediateData: Yes
59  InitialR2T: Yes
60  MaxOutstandingR2T: 1
61  # iscsiadm -m session -P 3
62  iSCSI Transport Class version 2.0-870
63  version 6.2.0.874-2
64  Target: iqn.2018-02.local.somuvmmnet:vmdeux (non-flash)
65  Current Portal: 10.0.99.12:3260,1
66  Persistent Portal: 10.0.99.12:3260,1
67  *****
68  Interface:
69  *****
70  Iface Name: default
71  Iface Transport: tcp
72  Iface Initiatorname: iqn.2018-02.local.somuvmmnet:vmprime
73  Iface IPaddress: 10.0.99.11
74  Iface HWaddress: <empty>
75  Iface Netdev: <empty>

```

```

76  SID: 8
77  iSCSI Connection State: LOGGED IN
78  iSCSI Session State: LOGGED_IN
79  Internal iscsid Session State: NO CHANGE
80  *****
81  Timeouts:
82  *****
83  Recovery Timeout: 120
84  Target Reset Timeout: 30
85  LUN Reset Timeout: 30
86  Abort Timeout: 15
87  *****
88  CHAP:
89  *****
90  username: <empty>
91  password: *****
92  username_in: <empty>
93  password_in: *****
94  *****
95  Negotiated iSCSI params:
96  *****
97  HeaderDigest: None
98  DataDigest: None
99  MaxRecvDataSegmentLength: 262144
100 MaxXmitDataSegmentLength: 262144
101 FirstBurstLength: 65536
102 MaxBurstLength: 262144
103 ImmediateData: Yes
104 InitialR2T: Yes
105 MaxOutstandingR2T: 1
106 *****
107 Attached SCSI devices:
108 *****
109 Host Number: 40          State: running
110 scsi40 Channel 00 Id 0 Lun: 0
111 Attached scsi disk sdb          State: running
112 scsi40 Channel 00 Id 0 Lun: 1
113 Attached scsi disk sdc          State: running
114 scsi40 Channel 00 Id 0 Lun: 2
115 Attached scsi disk sdd          State: running

```

2.4.2 Verification on the iSCSI Target

To verify the iSCSI config on the target, we need only check the contents of the `targetcli` command:

```

1  # targetcli
2  targetcli shell version 2.1.fb46
3  Copyright 2011-2013 by Datera, Inc and others.
4  For help on commands, type 'help'.
5
6  /> ls
7  o- / ..... [...]
8     o- backstores ..... [...]
9        | o- block ..... [Storage Objects: 2]
10       | | o- block1 ..... [/dev/vgSAN/lvSAN1 (508.0MiB) write-thru activated]
11       | | | o- alua ..... [ALUA Groups: 1]
12       | | |   o- default_tg_pt_gp ..... [ALUA state: Active/optimized]

```

```

13 | | o- block2 ..... [/dev/vgSAN/lvSAN2 (508.0MiB) write-thru activated]
14 | |   o- alua ..... [ALUA Groups: 1]
15 | |     o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
16 | o- fileio ..... [Storage Objects: 1]
17 | | o- file1 ..... [/root/diskFile1 (10.0MiB) write-back activated]
18 | |   o- alua ..... [ALUA Groups: 1]
19 | |     o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
20 | o- pscsi ..... [Storage Objects: 0]
21 | o- ramdisk ..... [Storage Objects: 0]
22 o- iscsi ..... [Targets: 1]
23 | o- iqn.2018-02.local.somuvmmnet:vmdeux ..... [TPGs: 1]
24 |   o- tpg1 ..... [no-gen-acls, no-auth]
25 |     o- acls ..... [ACLs: 1]
26 |       | o- iqn.2018-02.local.somuvmmnet:vmprime ..... [Mapped LUNs: 3]
27 |         | o- mapped_lun0 ..... [lun0 block/block1 (rw)]
28 |         | o- mapped_lun1 ..... [lun1 block/block2 (rw)]
29 |         | o- mapped_lun2 ..... [lun2 fileio/file1 (rw)]
30 |       o- luns ..... [LUNs: 3]
31 |         | o- lun0 ..... [block/block1 (/dev/vgSAN/lvSAN1) (default_tg_pt_gp)]
32 |         | o- lun1 ..... [block/block2 (/dev/vgSAN/lvSAN2) (default_tg_pt_gp)]
33 |         | o- lun2 ..... [fileio/file1 (/root/diskFile1) (default_tg_pt_gp)]
34 |       o- portals ..... [Portals: 1]
35 |         o- 10.0.99.12:3260 ..... [OK]
36 o- loopback ..... [Targets: 0]

```

Chapter 3

System Performance Reporting

3.1 Understanding System Performance Parameters

The definition of performance of a system is dependent upon the expectations from a system. For example, **low latency** is desired from *database servers*, while **high throughput** is needed from *file servers*.

Actual performance has to be judged on the basis of performance level agreements. This has to be clearly defined for anyone - "*The web server should always react within 10 seconds*" is better than "*generic load should be less than 60%*", because that's what the end user will care about!

Thus, first we need to decide upon which metrics we want to measure, and then collect baseline data for it via monitoring systems.

3.1.1 Typical Performance Focus Areas

| Factor | Description |
|---------|---|
| Memory | The single most important factor that affects server performance. When enough memory isn't available, swap has to be used to house the excess pages and then the IO performance suffers, thus bogging down the entire system. It even affects the network throughput. |
| Disk | Another very important factor in overall server performance. When the disk is slow, too much memory is wasted to buffer data that's waiting to be written to disk. Processes will also have to wait longer to access data from the disk. |
| Network | Network is no longer a significant bottleneck, since most network connections aren't 10Mbps anymore - enterprise infrastructure uses Gigabit connections as a standard. |
| CPU | While the CPU has many tunables, in general it is not a very significant factor in performance deterioration. It is only for certain workloads that CPU becomes a factor in performance. The gain from CPU optimizations can be expressed in nanoseconds. |

3.1.2 Common Performance Monitoring Tools

| Terms | Description |
|---------------|--|
| top | While it's a very basic tool, it's also very rich in features. It provides an excellent generic overview of everything going on in the system. Typical use case for top is to detect problems and then use a more specialized tool to diagnose further. |
| iostat | A dedicated tool to detect Input/Output problems. It shows statistics about I/O. To detect which process is creating a high I/O load, a valuable tool is iotop . |
| vmstat | This tool shows statistics about virtual memory usage. |
| sar | The System Activity Reporter specializes in providing long-term data about what the system has been doing and long term performance statistics. |

3.2 Understanding top

This is perhaps the single most important performance monitoring utility due to the kind of data it provides. There are alternatives to **top** such as **htop**, but **top** is programmed efficiently and doesn't have too much overhead. Comparing the two - **htop** uses about 5 times as much system resources as **top**!

The first feature of interest in the output of **top** is the **load average**, which consists of three numbers: the load average for the last 1, 5 and 15 minutes. The load average is the average of the number of processes in a runnable state, i.e., currently being executed by the CPU or waiting for CPU, over the concerned period of time. Optimally, all CPUs should be utilized as much as possible, but no process should be waiting for the CPU. The output of the **nproc** command tells us the effective number of CPUs available (= Physical CPUs × logical cores per CPU).

The individual CPU utilization per CPU core can be shown by pressing the **1** key. A typical output is:

```
1 # top
2 %Cpu0 :  5.5 us,  3.3 sy,  0.0 ni, 90.7 id,  0.0 wa,  0.5 hi,  0.0 si,  0.0 st
```

Here, the number after the CPU indicates the core number. The *us* value refers to CPU usage in percentage in user space, i.e., by processes started by the end user without administrative privileges. The *sy* does the same, but for processes started by the users with root privileges. The *id* value is the percentage of time the processor remains idle. The next important metric is the number before *wa* which represents the waiting time, i.e., percentage of time processes spend waiting for I/O. A high value here indicates that there's something wrong with the I/O channel and may indicate imminent disk failure.

Next, the memory statistics are shown, which includes the amount of memory completely free and amount of memory used to cache files that are frequently requested. Buffers contain data that needs to be written to disk during high I/O loads. While these are technically *non-essential*, it's suggested that 30% of the total memory be dedicated to buffers/cache usage.

We can also toggle the fields being shown by pressing the **f** key. If we quit **top** using the **q** key, the edits to the configuration are gone the moment we quit. However, if we quit using **Shift + W**, then the configuration is written to the **.toprc** file.

3.3 Understanding iostat

The `iostat` tool is a part of the `sysstat` package, which needs to be installed to use the `iostat` command. The command by itself provides a snapshot of the I/O statistics at the time of the invocation of the command. However, it takes two arguments in the syntax: `iostat <interval> <loops>`. The interval refers to the gap between displaying statistics and the loops refer to the number of times the command should show its output. Typical output for the command is:

```
1 # iostat 3 2
2 Linux 3.10.0-693.17.1.el7.x86_64 (vmPrime.somuVMnet.local)      Tuesday 27 February
   ↪  2018      _x86_64_      (1 CPU)
3
4 avg-cpu:  %user   %nice %system %iowait  %steal   %idle
5 0.50    0.00    0.64    0.49    0.00   98.37
6
7 Device:            tps    kB_read/s    kB_wrtn/s    kB_read    kB_wrtn
8 sda                 1.20         54.56         3.33     584199     35622
9 scd0                 0.00          0.10          0.00        1054         0
10 dm-0                1.11         51.31         3.13     549442     33537
11 dm-1                0.01          0.21          0.00        2228         0
12 sdb                 0.00          0.10          0.00        1044         0
13 sdc                 0.00          0.10          0.00        1044         0
14 sdd                 0.00          0.03          0.00         336         0
15
16 avg-cpu:  %user   %nice %system %iowait  %steal   %idle
17 5.44    0.00    1.36    0.00    0.00   93.20
18
19 Device:            tps    kB_read/s    kB_wrtn/s    kB_read    kB_wrtn
20 sda                 0.68          0.00         10.88         0         32
21 scd0                 0.00          0.00          0.00         0         0
22 dm-0                 2.04          0.00        32.48         0         95
23 dm-1                 0.00          0.00          0.00         0         0
24 sdb                 0.00          0.00          0.00         0         0
25 sdc                 0.00          0.00          0.00         0         0
26 sdd                 0.00          0.00          0.00         0         0
```

In the output, **tps** refers to the number of transactions per second. The *kB_read/s* and the *kB_wrtn/s* values are self explanatory. The next two columns show the total kBs read and written respectively.

3.3.1 Usage scenario

Let us consider a scenario where `top` shows us that processes spend 60% of their execution time waiting for I/O. Let us consider that the concerned server is connected to 6 different disks or other storage devices. We can use the output of the `iostat` command to determine which disk is so slow.

If we consult the output from the command, we can see that `dm-0` has the greatest tps. To find out which device is `dm-0`, we can simply go to the `/dev/mapper` directory and see what links to it:

```
1 # \ls -l /dev/mapper
2 total 0
3 crw-----. 1 root root 10, 236 Feb 27 20:53 control
4 lrwxrwxrwx. 1 root root    7 Feb 27 20:53 rhel-root -> ../dm-0
5 lrwxrwxrwx. 1 root root    7 Feb 27 20:53 rhel-swap -> ../dm-1
```

3.3.2 iotop

The **iotop** command needs to be installed using `yum -y install iotop`. It shows the processes that are doing the most amount of I/O in descending order. Typical output looks like:

```
1 # iotop
2 Total DISK READ :      45.37 M/s | Total DISK WRITE :      0.00 B/s
3 Actual DISK READ:      45.37 M/s | Actual DISK WRITE:      0.00 B/s
4 TID  PRIO  USER      DISK READ  DISK WRITE  SWAPIN      IO>   COMMAND
5 5696 be/4  root       45.37 M/s   0.00 B/s    0.00 % 73.97 % dd if=/dev/sda of=/dev/null
6 5450 be/4  root        0.00 B/s    0.00 B/s    0.00 % 12.66 % [kworker/0:2]
7 1 be/4  root        0.00 B/s    0.00 B/s    0.00 % 0.00 % systemd --switched-root --system
   ↪ --deserialize 21
8 ...
```

Here we can see that the `dd if=/dev/sda of=/dev/null` is performing the most amount of I/O by copying the entire hard disk to `/dev/null`.

3.4 Understanding vmstat

3.4.1 Virtual Memory

Let us consider the typical output of `top` sorted on the basis of the Virtual Memory being used:

```
1 # top
2 PID USER      PR  NI   VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
3 1920 somu    20   0 1901016 215552 47856 S   0.7  11.6   0:47.79 gnome-shell
4 ...
```

We can see that the `gnome-shell` is using 1901016 KiB of virtual memory, which is ≈ 1.82 GiB of virtual memory. Virtual Memory in Linux is memory that doesn't really exist. If we take a look at the `/proc/meminfo` file, we see:

```
1 VmallocTotal: 34359738367 kB
```

If we convert the `VmallocTotal` (Total amount of virtual memory that is possible for the kernel to allocate) to human readable units, it comes up to 32PB! That's not possible on most enterprise gear, let alone consumer hardware. Thus, the memory here doesn't really exist.

The key point here is that the kernel frequently needs to dish out unique memory address pointers to programs that demand it, but not actually assign any real memory till it's needed, i.e., the program tries to write to that location.

The kernel, instead of assigning real memory locations to programs, assigns memory in a virtual address space, which it then maps on to real memory on demand. The program itself remains blissfully oblivious to the knowledge of whether the memory it is referencing is virtual or real. All the trouble of fetching data on requirement and saving data falls on the kernel.

3.4.2 Resident Memory

A much more important concept is that of Resident Memory. Contrastingly to the Virtual Memory, the Resident memory is really used and is the total amount RAM being assigned to the process.

3.4.3 vmstat

The `vmstat` command when used without arguments shows various statistics pertaining to the resource consumption on the system:

```
1 # vmstat
2 procs -----memory----- --swap-- ----io---- -system-- -----cpu-----
3  r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
4  3  0       0 269904   2116 860396    0    0   176   25  124  129  2  2 95  2  0
```

The significance of each is:

| Terms | Description |
|---------------|---|
| proc | This part shows information about the processes: the r shows the number of running processes, b shows the number of blocking processes. A blocking process is a process that's waiting for something (e.g., I/O). |
| memory | This is the total amount of memory in swap, as well as real physical memory (RAM) used for buffers and cache. |
| swap | The two sub-categories are <i>swap-in</i> (si) and <i>swap-out</i> (so). If at any time we see that the system is utilizing swap memory, we can use <code>vmstat</code> to find out if the swap is being used actively, i.e., whether data is being written to or read from it actively. |
| io | The IO section deals with the number of blocks of I/O that's being performed - <i>blocks-in</i> (bi) and <i>blocks-out</i> (bo) provide a way to measure the real I/O activity at the moment, thus helping us discern if the server is spending a lot of time reading or writing during high I/O waits. |
| system | The metrics shown are <i>interrupts</i> (in) and <i>context switches</i> (cs). Interrupts are generally generated when a piece of hardware demands CPU attention. Context switches occur when the CPU switches the present task it's working on after being triggered by the scheduler. It is critical to the multi-tasking ability of a server since multiple processes need to coordinate and divide the CPU cycles. A high number of context-switches would indicate that the CPU isn't getting enough time per process. |
| cpu | These metrics refer to the percentage of CPU time spent executing programs in the <i>user-space</i> (us), <i>system space</i> (sy), <i>idle</i> (id) or <i>waiting</i> (wa). |

Just like `iostat`, the `vmstat` provides an option to show the information at multiple points in time - the first argument is the time delay and the second the number of loops. To re-run `vmstat` every 2 seconds for 5 times we use:

```
1 # vmstat 2 5
2 procs -----memory----- --swap-- ----io---- -system-- -----cpu-----
3  r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
4  2  0       0 255824   2116 877492    0    0   135   20  118  122  1  1 96  2  0
5  0  0       0 255576   2116 877492    0    0     0     0  174  178  6  3 91  0  0
6  0  0       0 255576   2116 877492    0    0     0     0  152  145  6  2 92  0  0
```

```

7  1  0      0 255384  2116 877492  0  0  0  0 179 161  8  2 90  0  0
8  0  0      0 255452  2116 877492  0  0  0  0 256 406  9  8 83  0  0

```

Just like in `iostat` output, the first line has higher values of certain metrics because it gives a generic overview of the system operations where applicable. The next lines portray the activity within the delay time frame.

For detailed memory utilization statistics, we use `vmstat -s`:

```

1  # vmstat -s
2  1865964 K total memory
3  730672 K used memory
4  902516 K active memory
5  466968 K inactive memory
6  255684 K free memory
7  2116 K buffer memory
8  877492 K swap cache
9  1048572 K total swap
10 0 K used swap
11 1048572 K free swap
12 6813 non-nice user cpu ticks
13 1125 nice user cpu ticks
14 7128 system cpu ticks
15 543996 idle cpu ticks
16 8383 IO-wait cpu ticks
17 0 IRQ cpu ticks
18 180 softirq cpu ticks
19 0 stolen cpu ticks
20 734741 pages paged in
21 106795 pages paged out
22 0 pages swapped in
23 0 pages swapped out
24 667859 interrupts
25 685758 CPU context switches
26 1519798095 boot time
27 4208 forks

```

To change the display unit being used, we use the `-S (k/K/m/M)` option to change units. (where K=KiB and k=kB).

3.5 Understanding sar components

`sar` stands for the **System Activity Reporter**. It is a part of the `sysstat` package (like `iostat` and `vmstat`), and it collects data on an interval of 10 minutes by default. However, it can also be used to collect instantaneous data about the system as well.

What truly distinguishes `sar` from the other tools is the fact that it can be tasked to data collection for an extended period of time and the queried for information about a very specific period.

To make sorting and finding data in `sar` easier, it is recommended to set `LANG=C` before starting `sar`. Every Linux OS has an environment variable called **LANG** that affects the behavior of several utilities as well as setting the language. POSIX standard mandates that a locale called either POSIX or C be defined. Thus, it disables localization and makes the output predictable. Unless the option is set, timestamps are formatted in AM/PM which makes filtering said data harder. With the `LANG=C` option however, the timestamps use the military time format (24-hour format). A handy option is to generate an alias such as:

```
1 # echo "alias sar='LANG=C sar'" >> /etc/bashrc
```

sar data is collected via cron jobs in `/etc/cron.d/sysstat`. The collected data is written to `/var/log/sa`. The file `/etc/sysconfig/sysstat` has a `HISTORY` variable which dictates how long data should be stored. Typically, it's on a monthly basis.

3.5.1 `/etc/cron.d/sysstat`

This cronjob launches two different jobs are launched: **sa1** and **sa2**. The *sa1* job is for collecting short term statistics data while the *sa2* job executes once a day to collect data for long term statistics. Both of these write the results of their monitoring in a file in the `/var/log` directory.

3.5.2 `/var/log/sa/sa[dd]`

These are actually a bunch of files that start with the prefix *sa* and end with the date in *dd* format. Thus, typical file names are: *sa01*, *sa25*, *sa31*, etc. These files are unreadable by typical pagers like `less` and needs to be read by using the `sar` utility itself, by issuing commands like `sar -q` to get information about disk statistics, etc. One common mistake while accustoming to `sar` is to forget to start the `sysstat` services, since without them the data for the `sar` log files aren't populated and the utility has no data to work with.

3.6 Setting up sar

If the **sysstat** package isn't already installed, we first need to install it using `yum -y install sysstat`. Next, we ensure that the cron job for data collection via *sa1* and *sa2* were set up properly in `/etc/cron.d/sysstat` file, which should have the contents:

```
1 # Run system activity accounting tool every 10 minutes
2 */10 * * * * root /usr/lib64/sa/sa1 1 1
3 # Generate a daily summary of process accounting at 23:53
4 53 23 * * * root /usr/lib64/sa/sa2 -A
```

Thus, *sa1* is collecting data every 10 mins and *sa2* executes everyday at 11:53PM to collect long term data for the day. Now `sar` is ready to collect data, but if we were to query the `sar` already, we'd come up empty, since `sar` hasn't had the opportunity to log data yet!

Next, we check the config in `/etc/sysconfig/sysstat` file, which typically looks like:

```
1 # sysstat-10.1.5 configuration file.
2
3 # How long to keep log files (in days).
4 # If value is greater than 28, then log files are kept in
5 # multiple directories, one for each month.
6 HISTORY=28
7
8 # Compress (using gzip or bzip2) sa and sar files older than (in days):
9 COMPRESSAFTER=31
10
11 # Parameters for the system activity data collector (see sadc manual page)
```

```

12 # which are used for the generation of log files.
13 SADC_OPTIONS="-S DISK"
14
15 # Compression program to use.
16 ZIP="bzip2"

```

The primary feature of interest in this file is the value of the HISTORY variable which decides how long the collected data is stored.

Now, we have to wait for the `sadc` (*System Activity Data Collector*) utility to collect data for **sar** to analyse.

3.7 Analyzing sar data

Some of the most common options that print certain categories of the collected data are:

3.7.1 I/O operations

The `sar -b` command shows us the total transfers per second (tps), read tps(rtps), write tps(wtps), blocks read per second (bread)(1 block = 512B) and blocks written to per second (bwrtn). Typical output of the command looks like:

```

1 # sar -b
2 00:00:02      tps      rtps      wtps  bread/s  bwrtn/s
3 00:10:01      2.70      1.15      1.55   64.96   995.40
4 00:20:01      0.47      0.27      0.20   12.97    8.26
5 00:30:01      0.07      0.00      0.07    0.00    0.86
6 00:40:01      0.08      0.00      0.08    0.00    1.04
7 ...

```

Thus, the use of military time makes the output a lot easier to process.

3.7.2 Processor information

It is possible to get the information about a single processor (i.e., a single logical core on a physical CPU, since linux considers each a separate processor) using the `sar -P <processorNumber>` command. Typical usage is (to find the usage of processor 0):

```

1 # sar -P 0
2 00:00:02      CPU      %user      %nice      %system      %iowait      %steal      %idle
3 00:10:01          0      5.21      0.00      3.09      0.59      0.00      91.11
4 00:20:01          0      0.12      0.00      0.24      0.15      0.00      99.48
5 ...
6 04:10:01          0      0.29      0.00      0.36      0.10      0.00      99.25
7 Average:          0      1.04      0.00      1.03      0.31      0.00      97.62

```

3.7.3 Network Statistics

The `sar -n DEV` command shows the network statistics for each interface:

```

1 # sar -n DEV
2 Linux 3.10.0-693.17.1.el7.x86_64 (vmPrime.somuVMnet.local)      02/28/18
   ↪      _x86_64_      (1 CPU)
3
4 00:00:02      IFACE  rxpck/s  txpck/s  rxkB/s  txkB/s  rxcmp/s  txcmp/s
   ↪ rxmcsst/s
5 00:10:01      lo      0.00    0.00    0.00    0.00    0.00    0.00
   ↪ 0.00
6 00:10:01  virbr0-nic      0.00    0.00    0.00    0.00    0.00    0.00
   ↪ 0.00
7 00:10:01      virbr0      0.00    0.00    0.00    0.00    0.00    0.00
   ↪ 0.00
8 00:10:01      ens33    63.83    19.03    90.10    1.19    0.00    0.00
   ↪ 0.00
9 ...
10 16:40:01      ens33     1.28     1.13     0.15     0.36    0.00    0.00
   ↪ 0.00
11
12 Average:      IFACE  rxpck/s  txpck/s  rxkB/s  txkB/s  rxcmp/s  txcmp/s
   ↪ rxmcsst/s
13 Average:      lo      0.00    0.00    0.00    0.00    0.00    0.00
   ↪ 0.00
14 Average:  virbr0-nic      0.00    0.00    0.00    0.00    0.00    0.00
   ↪ 0.00
15 Average:      virbr0      0.00    0.00    0.00    0.00    0.00    0.00
   ↪ 0.00
16 Average:      ens33     1.19     0.92     0.83     0.12    0.00    0.00
   ↪ 0.00

```

To view the statistics for just one interface, we can use `sar -n DEV \ grep <interface-Name>|`:

```

1 # sar -n DEV | grep ens33
2 00:10:01      ens33    63.83    19.03    90.10    1.19    0.00    0.00
   ↪ 0.00
3 ...
4 16:40:01      ens33     1.28     1.13     0.15     0.36    0.00    0.00
   ↪ 0.00
5 Average:      ens33     1.19     0.92     0.83     0.12    0.00    0.00
   ↪ 0.00

```

Chapter 4

System Optimization Basics

4.1 Understanding /proc contents

To optimize a Linux system, we should know in depth about the `/proc` file system. This file system provides an interface to the kernel using which we can take a look at the present state of the system as well as optimize it as per our requirements.

The `/proc` file system is filled with several **status files** which tell us about many operational system parameters. In fact, many performance monitoring utilities get their data from the `/proc` file system itself!

In older versions of Linux, there were several hard-coded parameters that needed to be changed to optimize the system, and thus required a recompilation of the kernel. Modern Linux kernels however, get many of these parameters from the `/proc/sys` file system and thus offer optimization in a flexible manner!

The `/proc/sys` file system offers different interfaces that are related to different aspects of the file system which can be tuned through their corresponding interface.

4.2 Analysing the /proc filesystem

4.2.1 Process Directories

The `/proc` file system has a process directory named with a number corresponding to the *PID* of every running process on the system:

```
1 # ls /proc
2 1      17      2064 2514 333 362 465 7   93      locks
3 10     1759   2070 2530 334 363 466 732 acpi     mdstat
4 1180   1764   2093 26   335 364 479 734 asound  meminfo
5 1186   18     2094 2611 336 365 480 736 buddyinfo misc
6 1187   1856   2109 2616 337 366 481 757 bus      modules
7 1191   1875   2116 2664 338 367 482 758 cgroups mounts
8 1195   1880   212 2665 339 368 483 760 cmdline mpt
9 1196   1884   2128 27   340 369 484 762 consoles mtrr
10 1199   19     2147 273 341 370 485 766 cpuinfo  net
11 12     1903   2150 274 342 371 486 767 crypto  pagetypeinfo
12 1297   1907   2151 275 343 372 487 768 devices partitions
```

| | | | | | | | | | | |
|----|------|------|------|-----|-----|-----|-----|-----|-------------|---------------|
| 13 | 1298 | 1921 | 2164 | 278 | 344 | 373 | 488 | 772 | diskstats | sched_debug |
| 14 | 1299 | 1933 | 2167 | 279 | 345 | 374 | 489 | 773 | dma | schedstat |
| 15 | 13 | 1938 | 2175 | 28 | 346 | 375 | 5 | 779 | driver | scsi |
| 16 | 1331 | 1940 | 2180 | 284 | 347 | 376 | 560 | 791 | execdomains | self |
| 17 | 1336 | 1948 | 2181 | 286 | 348 | 377 | 561 | 793 | fb | slabinfo |
| 18 | 1337 | 1953 | 2186 | 287 | 349 | 378 | 587 | 794 | filesystems | softirqs |
| 19 | 1338 | 1954 | 2188 | 288 | 350 | 379 | 589 | 795 | fs | stat |
| 20 | 1344 | 1961 | 2208 | 295 | 351 | 38 | 590 | 8 | interrupts | swaps |
| 21 | 1383 | 1962 | 2243 | 3 | 352 | 380 | 60 | 800 | iomem | sys |
| 22 | 14 | 1970 | 2250 | 323 | 353 | 381 | 601 | 801 | ioports | sysrq-trigger |
| 23 | 15 | 1973 | 2321 | 324 | 354 | 382 | 616 | 802 | irq | sysvipc |
| 24 | 1569 | 1979 | 2322 | 325 | 355 | 39 | 638 | 803 | kallsyms | timer_list |
| 25 | 1584 | 1983 | 2358 | 326 | 356 | 4 | 640 | 818 | kcore | timer_stats |
| 26 | 1586 | 2 | 2426 | 327 | 357 | 40 | 641 | 827 | keys | tty |
| 27 | 16 | 2001 | 2454 | 328 | 358 | 406 | 642 | 831 | key-users | uptime |
| 28 | 1640 | 2043 | 2460 | 329 | 359 | 407 | 645 | 835 | kmsg | version |
| 29 | 1643 | 2048 | 2462 | 330 | 36 | 41 | 646 | 852 | kpagecount | vmallocinfo |
| 30 | 1686 | 2053 | 2481 | 331 | 360 | 453 | 647 | 869 | kpageflags | vmstat |
| 31 | 1691 | 2059 | 25 | 332 | 361 | 454 | 648 | 9 | loadavg | zoneinfo |

Inside each of these process directories, there are several files that tell us about the present status of the process, and provide other necessary details about it. For example, inside the directory for the process with *PID 881*, we find:

```

1 # ls /proc/561/
2 attr                cpuset    limits    net        projid_map  stat
3 autogroup           cwd       loginuid  ns          root        statm
4 auxv                environ   map_files numa_maps   sched        status
5 cgroup              exe        maps      oom_adj     schedstat    syscall
6 clear_refs          fd         mem        oom_score   sessionid    task
7 cmdline             fdinfo     mountinfo oom_score_adj setgroups    timers
8 comm                gid_map    mounts     pagemap     smaps        uid_map
9 coredump_filter      io         mountstats personality stack         wchan

```

The **cmdline** file in this directory tells us about the command that is being run in the process. For example, the command that spawned the process with PID 561 is:

```

1 # cat /proc/561/cmdline
2 /usr/lib/systemd/systemd-journald

```

4.2.2 Status files

The `/proc` directory also houses several files that tell us about the different aspects of what our Operating System is doing and how it's performing. For example, the `/proc/partitions` file contains a list of all the storage devices that our system can access, as well as all the partitions that are housed on those devices:

```

1 # cat /proc/partitions
2 major  minor  #blocks name
3 8       0      10485760 sda
4 8       1      1048576 sda1
5 8       2      9436160 sda2
6 11      0       1048575 sr0
7 253     0       8384512 dm-0
8 253     1       1048576 dm-1

```

Similarly, the `/proc/cpuinfo` file has information about the CPU configuration, and the `/proc/meminfo` has information about the memory configuration. A list of all the file systems supported by the currently running operating systems (that have a kernel module presently loaded) can be found in `/proc/filesystems`:

```
1 # cat /proc/filesystems
2 nodev      sysfs
3 nodev      rootfs
4 nodev      ramfs
5 nodev      bdev
6 nodev      proc
7 nodev      cgroup
8 nodev      cpuset
9 nodev      tmpfs
10 nodev     devtmpfs
11 nodev     debugfs
12 nodev     securityfs
13 nodev     sockfs
14 nodev     pipefs
15 nodev     anon_inodefs
16 nodev     configfs
17 nodev     devpts
18 nodev     hugetlbfs
19 nodev     autofs
20 nodev     pstore
21 nodev     mqueue
22 nodev     selinuxfs
23          xfs
24 nodev     rpc_pipefs
25 nodev     nfsd
26          fuseblk
27 nodev     fuse
28 nodev     fusectl
```

Thus, if we insert the module for vFat or ext4 into the kernel, it'll show up in the `/proc/filesystems` listing:

```
1 # modprobe vfat
2 [root@vmPrime proc]# modprobe ext4
3 [root@vmPrime proc]# cat /proc/filesystems
4 nodev      sysfs
5 nodev      rootfs
6 ...
7 nodev      selinuxfs
8          xfs
9 nodev      nfsd
10          fuseblk
11 nodev      fusectl
12          vfat
13          ext3
14          ext2
15          ext4
```

4.2.3 `/proc/sys`

The `/proc/sys` file system is our interface to optimization of the system. It has folders related to the different aspects of the kernel's functions:

```
1 # ls /proc/sys
2 abi crypto debug dev fs kernel net sunrpc user vm
```

Among these, the most important ones are **fs** for *file systems*, **kernel** for *kernel optimizations*, **net** for *networking* and **vm** for the *virtual memory*.

For example, the `/proc/sys/vm` directory contains several files that help tune the kernel for memory optimization:

```
1 # ls /proc/sys/vm
2 admin_reserve_kbytes      lowmem_reserve_ratio      oom_dump_tasks
3 block_dump                max_map_count             oom_kill_allocating_task
4 compact_memory            memory_failure_early_kill overcommit_kbytes
5 dirty_background_bytes    memory_failure_recovery   overcommit_memory
6 dirty_background_ratio    min_free_kbytes           overcommit_ratio
7 dirty_bytes               min_slab_ratio            page-cluster
8 dirty_expire_centisecs    min_unmapped_ratio        panic_on_oom
9 dirty_ratio               mmap_min_addr             percpu_pagelist_fraction
10 dirty_writeback_centisecs mmap_rnd_bits              stat_interval
11 drop_caches               mmap_rnd_compat_bits      swappiness
12 extfrag_threshold         nr_hugepages               user_reserve_kbytes
13 hugepages_treat_as_movable nr_hugepages_mempolicy    vfs_cache_pressure
14 hugetlb_shm_group         nr_overcommit_hugepages   zone_reclaim_mode
15 laptop_mode               nr_pdflush_threads
16 legacy_va_layout          numa_zonelist_order
17 # cat swappiness
18 30
```

One such file is the `swappiness` file - which defines the willingness of a server to write data to the swap. The default value of `30` means it isn't very willing. If we were to increase the value set in the file, we would make it more likely for the kernel to use swap in cases where it normally wouldn't have.

4.3 Optimizing through /proc

Optimization through the `/proc/sys` file system is an easy affair that needs us to only change the value contained in certain files to tune the associated aspect. Let us consider the `/proc/sys/net/ipv4` directory, which deals with IPv4 networking on the system:

```
1 # cd /proc/sys/net/ipv4
2 # ls
3 # ls ip_*
4 ip_default_ttl  ip_forward      ip_local_reserved_ports
5 ip_dynaddr      ip_forward_use_pmtu ip_nonlocal_bind
6 ip_early_demux  ip_local_port_range ip_no_pmtu_disc
7 # cat ip_forward
8 0
```

The value of `0` in the `ip_forward` file shows that this system doesn't do IP forwarding, i.e., it doesn't forward specific IP packets for other IPs, i.e., it doesn't act as a router. To change this, we use:

```
1 # echo 1 > ip_forward
2 # cat ip_forward
3 1
```

When we echo values into the proc file system, it is only temporary and is wiped with every reboot - thus giving us an opportunity to test our settings directly before making anything permanent. If we like the results, we can make it permanent - otherwise reboot to restore the original values.

4.3.1 Sysctl

The utility `sysctl` can be used to display as well as control all the tunables that exist for a system:

```
1 # sysctl -a
2 abi.vsyscall32 = 1
3 crypto.fips_enabled = 0
4 debug.exception-trace = 1
5 ...
6 vm.swappiness = 30
7 vm.user_reserve_kbytes = 55126
8 vm.vfs_cache_pressure = 100
9 vm.zone_reclaim_mode = 0
```

Thus, there is no need to manually traverse the file system hierarchy in the `proc` file system. We can directly set the values using `sysctl` command. Further, it can help us find every single tunable that is related to a keyword very easily by piping the output to `grep`. For example, if we want to tune ICMP (the protocol behind Ping and other control messages that can be sent over IP packets) , we can use:

```
1 # sysctl -a | grep icmp
2 sysctl: reading key "net.ipv6.conf.all.stable_secret"
3 sysctl: reading key "net.ipv6.conf.default.stable_secret"
4 net.ipv4.icmp_echo_ignore_all = 0
5 net.ipv4.icmp_echo_ignore_broadcasts = 1
6 net.ipv4.icmp_errors_use_inbound_ifaddr = 0
7 net.ipv4.icmp_ignore_bogus_error_responses = 1
8 net.ipv4.icmp_msgs_burst = 50
9 net.ipv4.icmp_msgs_per_sec = 1000
10 net.ipv4.icmp_ratelimit = 1000
11 net.ipv4.icmp_ratemask = 6168
12 sysctl: reading key "net.ipv6.conf.ens33.stable_secret"
13 sysctl: reading key "net.ipv6.conf.lo.stable_secret"
14 sysctl: reading key "net.ipv6.conf.virbr0.stable_secret"
15 sysctl: reading key "net.ipv6.conf.virbr0-nic.stable_secret"
16 net.ipv6.icmp_ratelimit = 1000
17 net.netfilter.nf_conntrack_icmp_timeout = 30
18 net.netfilter.nf_conntrack_icmpv6_timeout = 30
```

Thus, `sysctl -a` and `grep` together make it extremely easy to discover the tunable we need to optimize any facet pertaining to our current needs.

4.4 Introducing sysctl

The purpose of the **sysctl** utility is that it can act as an interface to control the entirety of the `/proc/sys` file system, including making **permanent** changes to the tunables for system optimization. Thus, it's an invaluable tool to handle kernel runtime parameters.

We can put our parameters in `/etc/sysctl.conf` or the related directory `/etc/sysctl.d`. All of these parameters are read by **sysctl** and then passed on to the `/proc/sys` file system (also known as **procfs**). It is encouraged to thoroughly test out the changes by first echoing the parameter values to the concerned file in `/proc/sys` first so that any errors can be weeded out and then making the changes permanent by creating a `.conf` file in the `/etc/sysctl.d` directory.

4.5 Using sysctl

Any user changes to kernel runtime parameters should be put inside the `/etc/sysctl.d` directory. Anything put in this directory overwrites the vendor presets in the `/usr/lib/sysctl.d` directory, and of course, the default kernel parameter values. One such file containing vendor preset kernel parameters is the `/usr/lib/sysctl.d/00-system.conf`:

```
1 # Kernel sysctl configuration file
2 #
3 # For binary values, 0 is disabled, 1 is enabled. See sysctl(8) and
4 # sysctl.conf(5) for more details.
5
6 # Disable netfilter on bridges.
7 net.bridge.bridge-nf-call-ip6tables = 0
8 net.bridge.bridge-nf-call-iptables = 0
9 net.bridge.bridge-nf-call-arptables = 0
```

The files in the `/usr/lib/` directory in general shouldn't be touched since they're mostly vendor presets and thus get overwritten with every update. The only way to have our settings overwrite theirs is via configuring the settings in the concerned directory in the `/etc` directory.

The kernel parameters controllable through the `sysctl` utility are all named according to the convention : if the file is `/proc/sys/net/ipv4/ip_local_port_range`, then its equivalent `sysctl` setting will be `net.ipv4.ip_local_port_range`, i.e., each directory in the path name of the file after `/proc/sys` separated by a '.', terminating with the file name. This can be demonstrated with:

```
1 # cat /proc/sys/net/ipv4/ip_local_port_range
2 32768          60999
3 # sysctl -a | grep net.ipv4.ip_local_port_range
4 net.ipv4.ip_local_port_range = 32768          60999
```

Thus, to change the swappiness permanently, we write the setting in `/etc/sysctl.conf` (or a new configuration file in `/etc/sysctl.d`) and then reboot to see the changes:

```
1 vm.swappiness = 60
```

4.5.1 sysctl -w

To directly load a `sysctl` setting immediately without reading it from the `/proc/sys` file system, we use: `sysctl -w <parameter=value>`.

4.5.2 sysctl -p

To load an entire configuration file and set all of the kernel parameters mentioned in it, we use `sysctl -p <configFile.conf>`.

4.6 Modifying Network Behaviour through /proc and sysctl

We can see our IP address(es) and ping it via:

```
1 # ip a
2 1: ens33: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP qlen 1000
3 link/ether 00:0c:29:3b:b9:1c brd ff:ff:ff:ff:ff:ff
4 inet 10.0.99.11/24 brd 10.0.99.255 scope global ens33
5 valid_lft forever preferred_lft forever
6 inet6 fe80::f408:1ebf:7742:9fd8/64 scope link
7 valid_lft forever preferred_lft forever
8 # ping -c 2 10.0.99.11
9 PING 10.0.99.11 (10.0.99.11) 56(84) bytes of data.
10 64 bytes from 10.0.99.11: icmp_seq=1 ttl=64 time=0.420 ms
11 64 bytes from 10.0.99.11: icmp_seq=2 ttl=64 time=0.135 ms
12
13 --- 10.0.99.11 ping statistics ---
14 2 packets transmitted, 2 received, 0% packet loss, time 1001ms
15 rtt min/avg/max/mdev = 0.135/0.277/0.420/0.143 ms
```

Now, ping is an ICMP control message, and ICMP control messages can be blocked by changing an associated kernel parameter. To find the parameter responsible for ignoring ICMP echo requests that ping needs, we use:

```
1 # sysctl -a | grep icmp
2 net.ipv4.icmp_echo_ignore_all = 0
3 net.ipv4.icmp_echo_ignore_broadcasts = 1
4 net.ipv4.icmp_errors_use_inbound_ifaddr = 0
5 net.ipv4.icmp_ignore_bogus_error_responses = 1
6 net.ipv4.icmp_msgs_burst = 50
7 net.ipv4.icmp_msgs_per_sec = 1000
8 net.ipv4.icmp_ratelimit = 1000
9 net.ipv4.icmp_ratemask = 6168
10 net.ipv6.icmp.ratelimit = 1000
11 net.netfilter.nf_conntrack_icmp_timeout = 30
12 net.netfilter.nf_conntrack_icmpv6_timeout = 30
```

The very first displayed parameter is exactly what we need. So, we first test it with:

```
1 # echo 1 > /proc/sys/net/ipv4/icmp_echo_ignore_all
2 [root@vmPrime ~]# ping 10.0.99.11
3 PING 10.0.99.11 (10.0.99.11) 56(84) bytes of data.
4 ^C
5 --- 10.0.99.11 ping statistics ---
6 4 packets transmitted, 0 received, 100% packet loss, time 3000ms
```

Now that our host has stopped replying to pings, we can make the changes permanent by adding to the `/etc/sysctl.conf` file (and then rebooting):

```
1 net.ipv4.icmp_echo_ignore_all = 1
```

Chapter 5

Configuring Logging

5.1 Understanding Logging In RHEL 7

Due to the introduction of Systemd several services of the older Unix System V now have a counter part in their systemd equivalent. Such is the case for rsyslogd and journald. The former is responsible for logging in System V systems while journald does it in systemd systems. However, due to the concern of backwards compatibility (i.e., being able to use tools written for older versions of Linux which may have used System V utilities), modern distros like RHEL 7 also have rsyslogd installed.

There can be thus, three different ways of logging application information in RHEL 7:

- Directly write to a log file somewhere - no standardized way of accessing logs.
- Write to Systemd's *Journald* - logs are accessible via **journalctl**
- Write to *rsyslogd* - logs are accessible via `/var/log`.

An important point to note here is that **rsyslog** is still the central logging authority, but journald simply adds features to the way that logging is organized. Thus, journald doesn't really replace rsyslog.

This however means that there's scope for confusion on part of the user (or admin) who's handling the system - to understand exactly where a certain programs might write it's logs to. Thus, we can connect the two together to show the same information.

5.2 Connecting Journald to Rsyslogd

We merely need to add a few lines of configuration to have both services report their own logs to each other:

```
1 # To connect Journald to rsyslogd:
2 ## In /etc/rsyslog.conf:
3 $ModLoad imuxsock
4 $OmitLocalLogging off
5 ## In /etc/rsyslog.d/listend.conf:
6 $SystemLogSocketName /run/systemd/journal/syslog
7
```

```

8  # To connect rsyslogd to journald:
9  ## In /etc/rsyslog.conf:
10 $ModLoad omjournal
11 *.* :omjournal:

```

rsyslog messages are sent to **journald**, and vice versa. However, sending to **journald** is disabled by default in **rsyslog.conf**. To fix this we add the load the module **omjournal** (*output module journal*) using `$ModLoad omjournal`. Next, we use **rsyslog**'s notation for indicating the facility, priority and destination. The statement `*.* :omjournal:` means for any facility and any priority, we want the log to be forwarded to *omjournal*.

Receiving from **journal** is enabled by default in **rsyslog.conf**. This is done via: `$ModLoad imuxsock` (Input Module UniX SOCKet), which instructs **rsyslog** to listen to a socket. Now, local logging has to be enabled using the `$OmitLocalLogging off` option. Finally, the socket name on which **rsyslog** will listen will have to be specified in the `/etc/rsyslog.d/listend.conf` file, and has to be set to the value `/run/systemd/journal/syslog`. Everything on the **systemd** end is already configured and needs no manual intervention. This completes the integration of the two.

5.2.1 Modules

Thus, the connection between **rsyslog** and **journald** is facilitated by modules. There are several types of modules, which can be identified and classified by:

| Prefix | Type | Description |
|------------|---------------|--|
| im* | Input Module | Source of information for the rsyslog journal; Loaded in <code>/etc/rsyslog.conf</code> and socket name specified in <code>/etc/rsyslog.d/listend.conf</code> . |
| om* | Output Module | Destination to which data from rsyslog will be sent; Configured in <code>/etc/rsyslog.conf</code> |
| | | Other modules such as parser modules, message modification modules, etc. |

Together these modules lets us manipulate the log messages any way we want.

5.2.2 Importing text files to log : httpd error log

To import the HTTPD error log to **rsyslog**, the following needs to be added to the file `/etc/rsyslog.conf`:

```

1  $ModLoad imfile
2  $InputFileName /var/log/httpd/error_log
3  $InputFileTag apache-error:
4  $InputFileState state-apache-error
5  $InputRunFileMontitor

```

This takes the error log read and maintained by **apache** and inserts the data into **rsyslog**. The `$InputRunFileMonitor` enables monitoring of the specified file.

5.2.3 Exporting data to an output module : exporting to a database

Since rsyslog writes the data to a simple text file, and for managing log information the ability to query is very important, we may choose to export the data to a database. Assuming we're using a MySQL database:

```
1 $ModLoad ommysql
2 $ActionOmmysqlServerPort 1234
3 *.* :ommysql:database-serverName,database-name,database-userid,database-password
```

The first line loads the MySQL Output module for rsyslog. Then, we define the server port on which the logs will be forwarded. Then, finally, we configure the output module using the *facility, priority :destination* method where we send everything (all facilities and every priority[*.*]) to the output module, while also providing the database details.

Part II

Networking and Apache

Chapter 6

Configuring Advanced Networking

Chapter 7

Managing Linux Based Firewalls

Chapter 8

Configuring Apache Virtual Hosts

Chapter 9

Managing Advanced Apache Features

Part III

DNS and File Sharing

Chapter 10

Configuring a Cache-only DNS Server

Chapter 11

Configuring NFS File Sharing

Chapter 12

Managing SMB File Sharing

Part IV

Essential Back-end Services

Chapter 13

Setting up an SMTP Server

Chapter 14

Managing SSH

Chapter 15

Managing MariaDB

Chapter 16

Managing Time Services

Chapter 17

Shell Scripting