LVM Thin Provisioning

NYRHUG September 2015

Patrick Ladd
Technical Account Manager
pladd@redhat.com

#whatis TAM

- Premium named-resource support
- Proactive and early access
- Regular calls and on-site engagements
- Customer advocate within Red Hat and upstream
- Multi-vendor support coordinator
- High-touch access to engineering
- Influence for software enhancements
- NOT Hands-on or consulting

Agenda

- System Setup - Before
- LVM Thin Provisionng
  - The easy way
  - What it did
  - The long way
- System after Thin Provisioning

System Before

Basic 7.1 Install

```
[root@rhel71 ~]# lsblk
NAME          MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
sda             8:0    0   20G  0 disk
├─sda1          8:1    0  500M  0 part /boot
└─sda2          8:2    0 19.5G  0 part
  ├─rhel-swap 253:0    0    2G  0 lvm  [SWAP]
  └─rhel-root 253:1    0 17.5G  0 lvm  /
sdb             8:16   0    4G  0 disk
sdc             8:32   0    2G  0 disk
sr0            11:0    1 1024M  0 rom
```

```
[root@rhel71 ~]# pvs -a
  PV              VG   Fmt  Attr PSize  PFree
  /dev/cdrom                ---      0      0
  /dev/rhel/root            ---      0      0
  /dev/rhel/swap            ---      0      0
  /dev/sda                  ---      0      0
  /dev/sda1                 ---      0      0
  /dev/sda2       rhel lvm2 a--  19.51g 40.00m
  /dev/sdb                  ---      0      0
  /dev/sdc                  ---      0      0
```

Basic 7.1 Install

```
[root@rhel71 ~]# lvs -a
  LV   VG   Attr       LSize  Pool Origin Data%  Meta%  Move Log Cpy%Sync Convert
  root rhel -wi-ao---- 17.47g
  swap rhel -wi-ao----  2.00g
```

```
[root@rhel71 ~]# findmnt
TARGET                           SOURCE     FSTYPE     OPTIONS
/                                /dev/mapper/rhel-root
                                            xfs        rw,relatime,seclabel,attr2,inode64,noquota
├─/proc                          proc       proc       rw,nosuid,nodev,noexec,relatime
│ └─/proc/sys/fs/binfmt_misc     systemd-1  autofs     rw,relatime,fd=33,pgrp=1,timeout=300,minproto=5,maxproto=5,direct
├─/sys                           sysfs      sysfs      rw,nosuid,nodev,noexec,relatime,seclabel
│ ├─/sys/kernel/security         securityfs securityfs rw,nosuid,nodev,noexec,relatime
│ ├─/sys/fs/cgroup               tmpfs      tmpfs      rw,nosuid,nodev,noexec,seclabel,mode=755
│ │ ├─/sys/fs/cgroup/systemd     cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,xattr,release_agent=/usr/lib/systemd/systemd-cgroups-
│ │ ├─/sys/fs/cgroup/cpuset      cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,cpuset
│ │ ├─/sys/fs/cgroup/cpu,cpuacct cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,cpuacct,cpu
│ │ ├─/sys/fs/cgroup/memory      cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,memory
│ │ ├─/sys/fs/cgroup/devices     cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,devices
│ │ ├─/sys/fs/cgroup/freezer     cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,freezer
│ │ ├─/sys/fs/cgroup/net_cls     cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,net_cls
│ │ ├─/sys/fs/cgroup/blkio       cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,blkio
│ │ ├─/sys/fs/cgroup/perf_event  cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,perf_event
│ │ └─/sys/fs/cgroup/hugetlb     cgroup     cgroup     rw,nosuid,nodev,noexec,relatime,hugetlb
│ ├─/sys/fs/pstore               pstore     pstore     rw,nosuid,nodev,noexec,relatime
│ ├─/sys/kernel/config           configfs   configfs   rw,relatime
│ ├─/sys/fs/selinux              selinuxfs  selinuxfs  rw,relatime
│ └─/sys/kernel/debug            debugfs    debugfs    rw,relatime
├─/dev                           devtmpfs   devtmpfs   rw,nosuid,seclabel,size=1931784k,nr_inodes=482946,mode=755
│ ├─/dev/shm                     tmpfs      tmpfs      rw,nosuid,nodev,seclabel
│ ├─/dev/pts                     devpts     devpts     rw,nosuid,noexec,relatime,seclabel,gid=5,mode=620,ptmxmode=000
│ ├─/dev/mqueue                  mqueue     mqueue     rw,relatime,seclabel
│ └─/dev/hugepages               hugetlbfs  hugetlbfs  rw,relatime,seclabel
├─/run                           tmpfs      tmpfs      rw,nosuid,nodev,seclabel,mode=755
└─/boot                          /dev/sda1  xfs        rw,relatime,seclabel,attr2,inode64,noquota
```

LVM Thin Provisioning

Thin volumes

- Over-allocate current storage available
- Needs to be specified at creation time
  - Steps:
    - Create thin pool logical volume (LV)
    - Create thin LVs with -V instead of -L
- Scenarios:
  - Don't know how much I'll need or where
  - Thin snapshots!

Setting up a thin volume - Step 1: Add more disk space

```
[root@rhel71 ~]# pvcreate /dev/sdb
  Physical volume "/dev/sdb" successfully created
[root@rhel71 ~]# vgextend rhel /dev/sdb
  Volume group "rhel" successfully extended
[root@rhel71 ~]# vgdisplay
  --- Volume group ---
  VG Name               rhel
  System ID
  Format                lvm2
  Metadata Areas        2
  Metadata Sequence No  4
  VG Access             read/write
  VG Status             resizable
  MAX LV                0
  Cur LV                2
  Open LV               2
  Max PV                0
  Cur PV                2
  Act PV                2
  VG Size               23.50 GiB
  PE Size               4.00 MiB
  Total PE              6017
  Alloc PE / Size       4984 / 19.47 GiB
  Free  PE / Size       1033 / 4.04 GiB
  VG UUID               3NPodf-TaMS-tatT-C812-dmdR-UHWP-V8JgHC
```

Setting up a thin volume - Step 2: Create Thin Pool

```
[root@rhel71 ~]# lvcreate --type thin-pool
--name mypool -L 4G rhel

  Logical volume "mypool" created.
```

Setting up a thin volume -
Step 3: Create Thin Volume

```
[root@rhel71 ~]# lvcreate -V 50G
--thinpool mypool rhel --name thinvol
```

```
   WARNING: Sum of all thin volume sizes (50.00 GiB) exceeds
the size of thin pool rhel/mypool and the size of whole volume
group (23.50 GiB)!
  For thin pool auto extension
activation/thin_pool_autoextend_threshold should be below 100.
  Logical volume "thinvol" created.
```

Setting up a thin volume - Step 4: Profit!

```
[root@rhel71 ~]# lvs
  LV      VG   Attr       LSize  Pool   Origin Data%  Meta%  Move Log Cpy%Sync Convert
  mypool  rhel twi-aotz--  4.00g               0.00   1.37
  root    rhel -wi-ao---- 17.47g
  swap    rhel -wi-ao----  2.00g
  thinvol rhel Vwi-a-tz-- 50.00g mypool        0.00

[root@rhel71 ~]# mkfs.xfs /dev/rhel/
root     swap     thinvol
[root@rhel71 ~]# mkfs.xfs /dev/rhel/thinvol
meta-data=/dev/rhel/thinvol      isize=256    agcount=16, agsize=819184 blks
         =                       sectsz=512   attr=2, projid32bit=1
         =                       crc=0        finobt=0
data     =                       bsize=4096   blocks=13106944, imaxpct=25
         =                       sunit=16     swidth=16 blks
naming   =version 2              bsize=4096   ascii-ci=0 ftype=0
log      =internal log           bsize=4096   blocks=6400, version=2
         =                       sectsz=512   sunit=16 blks, lazy-count=1
realtime =none                   extsz=4096   blocks=0, rtextents=0
[root@rhel71 ~]# mount /dev/rhel/thinvol /mnt

[root@rhel71 ~]# df -h /mnt/
Filesystem                Size  Used Avail Use% Mounted on
/dev/mapper/rhel-thinvol   50G   33M   50G   1% /mnt
```

OK – What did we just do???

```
[root@rhel71 ~]# lvs -a
  LV              VG   Attr       LSize  Pool   Origin Data%  Meta%  Move Log Cpy%Sync Convert
  [lvol0_pmspare] rhel ewi-------  4.00m
  mypool          rhel twi-aotz--  4.00g               0.64   1.66
  [mypool_tdata]  rhel Twi-ao----  4.00g
  [mypool_tmeta]  rhel ewi-ao----  4.00m
  root            rhel -wi-ao---- 17.47g
  swap            rhel -wi-ao----  2.00g
  thinvol         rhel Vwi-aotz-- 50.00g mypool        0.05
```

The “`lvcreate --type thin-pool`” command created:

- LV mypool_tdata – thin data block pool
- LV mypool_tmeta – metadata for thin volume
  - Default (Pool_LV_size / Pool_LV_chunk_size * 64)
  - Min 2MB / Max 16GB
- LV lvol0_pmspare – recovery area for metadata areas
  - Used in case metadata needs repair – copied to the _pmspare and worked on, overwrites original if successful
  - As large as the largest metadata area

Keep an eye on your pool

```
[root@rhel71 mnt]# ls -lh foo; lvs; df -h /mnt
-rw-r--r--. 1 root root 4.0G Sep  9 17:03 foo
  LV      VG   Attr       LSize  Pool   Origin Data%  Meta%  Move Log Cpy%Sync Convert
  mypool  rhel twi-aotzD-  4.00g               100.00 48.83
  root    rhel -wi-ao---- 17.47g
  swap    rhel -wi-ao----  2.00g
  thinvol rhel Vwi-aotz-- 50.00g mypool        8.00
Filesystem                 Size  Used Avail Use% Mounted on
/dev/mapper/rhel-thinvol    50G  8.1G   42G  17% /mnt
```

If you have thin pools – monitor the lvs status information!

It doesn't fail gracefully!

```
[root@rhel71 mnt]# ls -lh; lvs -a; df -h /mnt
total 8.0G
-rw-r--r--. 1 root root 3.2G Sep  9 17:07 bar
-rw-r--r--. 1 root root 4.0G Sep  9 17:03 foo
  LV              VG   Attr       LSize  Pool   Origin Data%  Meta%  Move Log Cpy%Sync Convert
  [lvol0_pmspare] rhel ewi-------  4.00m
  mypool          rhel twi-aotzM-  4.00g               100.00 48.83
  [mypool_tdata]  rhel Twi-ao----  4.00g
  [mypool_tmeta]  rhel ewi-ao----  4.00m
  root            rhel -wi-ao---- 17.47g
  swap            rhel -wi-ao----  2.00g
  thinvol         rhel Vwi-aotz-- 50.00g mypool        8.00
Filesystem                Size  Used Avail Use% Mounted on
/dev/mapper/rhel-thinvol   50G  8.1G   42G  17% /mnt
```

Setting up a thin volume – the long way Step 1: Math :(

- If you're planning on allocating the full disk size, you need to plan ahead
- Easiest to backtrack data size from after allocating metadata areas
  - _pmpspare area needs to be as large as largest metadata
  - _metadata needs to be at least 2MB, should be (Pool_LV_size / Pool_LV_chunk_size * 64)
- Use extents (-l) to allocate LVs, not human (-L) sizes
- Example:
  - Device has 1000 extents
  - Allocate 16 extents for _pmpspare
  - Allocate 16 extents for _metadata
  - Allocate 1000 – 16 – 16 = 968 for data

Setting up a thin volume – the long way
Step 2: Build sub-volumes

```
[root@rhel71 ~]# lvcreate -l 16 --name pool_meta rhel
  Logical volume "pool_meta" created.
[root@rhel71 ~]# lvcreate -l 968 --name pool_data rhel
  Logical volume "pool_data" created.
```

Setting up a thin volume – the long way Step 3: Convert data LV to pool

```
[root@rhel71 ~]# lvconvert --type thin-pool --poolmetadata
/dev/rhel/pool_meta /dev/rhel/pool_data

  WARNING: Converting logical volume rhel/pool_data and
rhel/pool_meta to pool's data and metadata volumes.

  THIS WILL DESTROY CONTENT OF LOGICAL VOLUME (filesystem etc.)
Do you really want to convert rhel/pool_data and rhel/pool_meta?
[y/n]: y

  Converted rhel/pool_data to thin pool.
```

Setting up a thin volume – the long way
Step 4: Profit!

```
[root@rhel71 ~]# lvs -a
  LV                VG   Attr       LSize  Pool Origin Data%  Meta%  Move Log Cpy%Sync Convert
  [lvol0_pmspare]   rhel ewi------- 64.00m
  pool_data         rhel twi-a-tz--  3.78g             0.00   0.08
  [pool_data_tdata] rhel Twi-ao----  3.78g
  [pool_data_tmeta] rhel ewi-ao---- 64.00m
  root              rhel -wi-ao---- 17.47g
  swap              rhel -wi-ao----  2.00g
```

Documentation

- LVM Administration guide:

  https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Logical_Volume_Manager_Administration/index.html redhat.

THANK YOU

plus.google.com/+RedHat

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHatNews