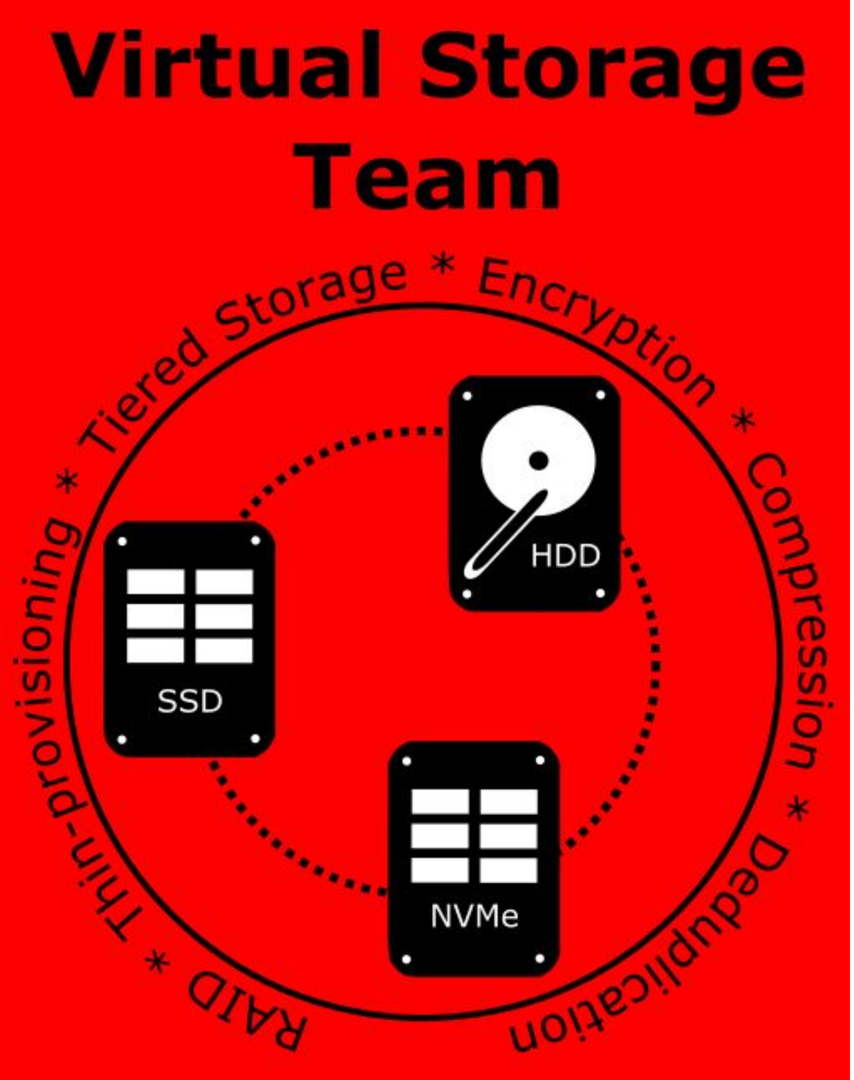

Virtual Storage Team
Tiered Storage * Encryption * Compression * Deduplication * RAID * Thin-provisioning *
HDD
SSD
NVMe

Virtual Storage at Red Hat

RHUG - Feb 4th, 2020

A Bit of History...

- MD / Software RAID
  - Does storage aggregation
  - Provides RAID 0, 1/10/1E, 4/5/6/
  - Metadata stored on-disk (label and operational)
  - Administered via *mdadm*
- LVM
  - Does storage virtualization
  - Provides Linear, stripe, mirror, snapshot, RAID, thin-p, caching
  - Manages label metadata
  - Relies on device-mapper for runtime, kernel “targets”

Virtual Storage Team

A bit more history

- Device-mapper
    - Reasonably simple interface for software storage targets
    - non-LVM targets include: dm-crypt, dm-multipath, dm-zoned, dm-delay, dm-dust...
    - Target specific metadata only
    - Labels written by admin layer (e.g. LVM2, cryptsetup)
- VDO
    - Acquired by RHT, open-sourced shortly after
    - Compression, deduplication, thin-p

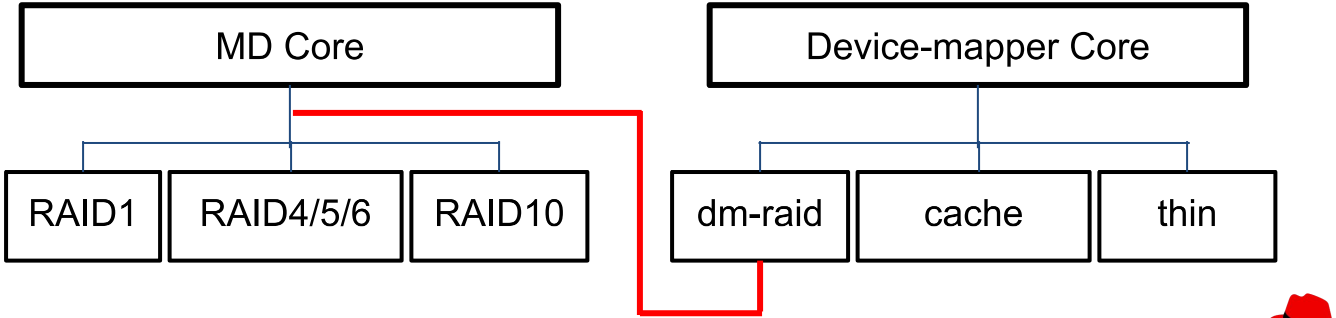
A digression on metadata...

- Some device-mapper targets keep data separate from metadata
  - Pro: allows placement on different HW with different characteristics
  - Pro: allows metadata to be shared w/o data (e.g. for bug fixing and recovery)
  - Pro: isolation of writes - impossible for one to write to another
  - Con: more complex setup
  - Con: resize operations now involve two pieces
  - Con: confusing to users

Virtual Storage Team

Joining forces

- LVM interface to RAID
  - dm-raid456 target was in the works, but abandon
  - dm-raid was created as a shim layer between MD and DM
  - Most, but not all features are in
    - In: all RAID types, per-device bitmaps, reshaping, writeback, writemostly, sync throttling, scrubbing
    - Out: bad block remapping, raid5 journaling



Virtual Storage Team

LVM RAID

$> lvcreate -m 1 -L 500G vg

$> lvcreate --type raid1 -m 1 -L 500G -n lv vg

$> lvcreate --type raid1 --mirrors 1 --size 500G --name lv vg /dev/sd[bc]1

```
[root@bp-02 ~]# lvs -a -o name,vgname,attr,size,syncpercent,devices vg
  LV            VG Attr       LSize   Cpy%Sync Devices
  lv            vg rwi-a-r--- 500.00g 38.03    lv_rimage_0(0),lv_rimage_1(0)
  [lv_rimage_0] vg Iwi-aor--- 500.00g          /dev/sdb1(1)
  [lv_rimage_1] vg Iwi-aor--- 500.00g          /dev/sdc1(1)
  [lv_rmeta_0]  vg ewi-aor---   4.00m          /dev/sdb1(0)
  [lv_rmeta_1]  vg ewi-aor---   4.00m          /dev/sdc1(0)
```

- See lvmraid(7)

LVM - changing RAID types

$> lvconvert --type raid1 vg/lv

```
[root@bp-02 ~]# lvs -a -o name,vgname,attr,size,syncpercent,devices vg
  LV             VG Attr       LSize Cpy%Sync Devices
  lv             vg mwi-a-m--- 5.00g 100.00   lv_mimage_0(0),lv_mimage_1(0)
  [lv_mimage_0] vg iwi-aom--- 5.00g          /dev/sdb1(0)
  [lv_mimage_1] vg iwi-aom--- 5.00g          /dev/sdc1(0)
  [lv_mlog]     vg lwi-aom--- 4.00m          /dev/sdc1(1280)
[root@bp-02 ~]# lvconvert --type raid1 vg/lv
Are you sure you want to convert mirror LV vg/lv to raid1 type? [y/n]: y
  Logical volume vg/lv successfully converted.
[root@bp-02 ~]# lvs -a -o name,vgname,attr,size,syncpercent,devices vg
  LV             VG Attr       LSize Cpy%Sync Devices
  lv             vg rwi-a-r--- 5.00g 100.00   lv_rimage_0(0),lv_rimage_1(0)
  [lv_rimage_0] vg iwi-aor--- 5.00g          /dev/sdb1(0)
  [lv_rimage_1] vg iwi-aor--- 5.00g          /dev/sdc1(0)
  [lv_rmeta_0]  vg ewi-aor--- 4.00m          /dev/sdb1(1280)
  [lv_rmeta_1]  vg ewi-aor--- 4.00m          /dev/sdc1(1281)
```

Virtual Storage Team

LVM RAID - features in development

- RAID1E
  - Like RAID10, but with elastic # of stripes

```
2 drives          3 drives             4 drives
========          ==========           ==============
A1  A1            A1  A1  A2           A1  A1  A2  A2
A2  A2            A2  A3  A3           A3  A3  A4  A4
A3  A3            A4  A4  A5           A5  A5  A6  A6
A4  A4            A5  A6  A6           A7  A7  A8  A8
..  ..            ..  ..  ..           ..  ..  ..  ..
========          ==========           ==============
```

LVM RAID - features in development

- Dm-integrity enhancement
  - Allows for self-healing of soft-corruption (e.g. adjacent track erasure, cosmic rays, etc)
  - Will be able to add or remove while volume is active
  - Comes with a performance penalty

Virtual Storage Team

Clustering

- Clvmd is out (RHEL8), shared LVM is in
  - See lvmlockd(7)
- Cluster “mirror”ing is out (RHEL8)
- Cluster RAID1 / 10 / 1E is in development
- Cluster snapshots, thin-p, caching - not coming

Thin-provisioning

```
$> lvcreate -T -L 5G -V 10G -n thinLV vg/thinpool
$> lvcreate -T -V 10G -n thinLV2 vg/thinpool
$> lvcreate -s -n thinLV_snap vg/thinLV
```

- LVM thin-p allocates blocks from physical storage only when used - see lvmthin(7)
  - Can create LV larger than backing store (over-provisioning)
  - Multiple “thinLV”s can share the same phy device
  - Supports thousands of non-COW snapshots
  - Running out of back-end space can hurt!

Virtual Storage Team

Virtual Data Optimizer (VDO)

- VDO provides deduplication, compression, zero-block elimination
  - Also a form of thin-provisioning
  - Allows over-provisioning
  - Can run out of space even if writing to previously allocated blocks
- LVM integration is in development
  - Currently managed by ‘vdomgr’
  - Styled after thin-p with vdoPool and vdoLVs

Caching

$> lvcreate -L 100G -n lv vg /dev/slow
$> lvcreate -H -L 10G -n cachepool vg/lv /dev/fast

```
$> lvs -a -o name,vgname,attr,size,syncpercent,devices vg
  LV                VG Attr       LSize   Cpy%Sync Devices
  [cachepool]       vg Cwi---C---  10.00g 0.00     cachepool_cdata(0)
  [cachepool_cdata] vg Cwi-ao----  10.00g          /dev/sdb1(25606)
  [cachepool_cmeta] vg ewi-ao----  12.00m          /dev/sdb1(25603)
  lv                vg Cwi-a-C--- 100.00g 0.00     lv_corig(0)
  [lv_corig]        vg owi-aoC--- 100.00g          /dev/sdb1(0)
```

- See lvmcache(7)

Caching cont.

- Current cache implementation based on dm-cache
  - Functions as a hot-spot cache
  - Takes time to warm, adapts to changing workloads
  - Separate data and metadata area
- Secondary cache type in development
  - Based on dm-writecache
  - Interleaved metadata
  - Speeds writes, reads generally serviced from page cache

Virtual Storage Team

Other development

- Storage Instantiation Daemon (SID)
- Boot Entry Manager (BOOM)
- Snapshot Manager
- Multipath
- Encryption
- Stratis