Western Digital®

Zoned Storage Support in Fedora: Current Status and Future

Damien Le Moal

Western Digital Research, System Software Group

Fedora Nest 2021

Outline

- Background
  - Zoned storage device overview
  - Kernel support history
- Zoned Storage in Fedora: current status
  - Kernel, system utilities, and applications
  - Missing, but needed, things
- Future improvements
  - New kernel features, new applications

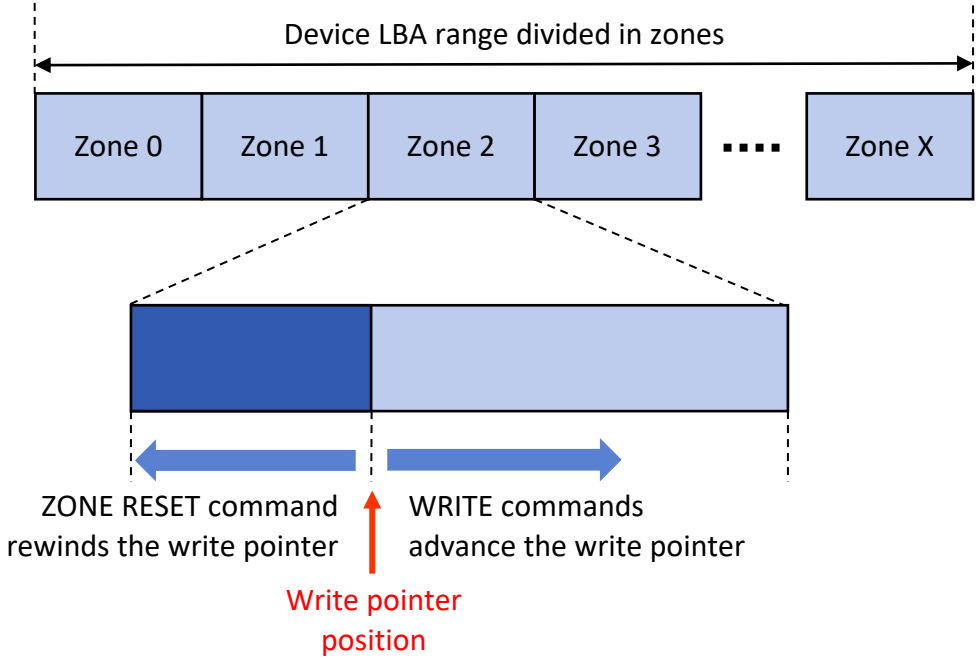
Background

Zoned Storage Overview and Linux Kernel Support History

Zoned Storage Devices

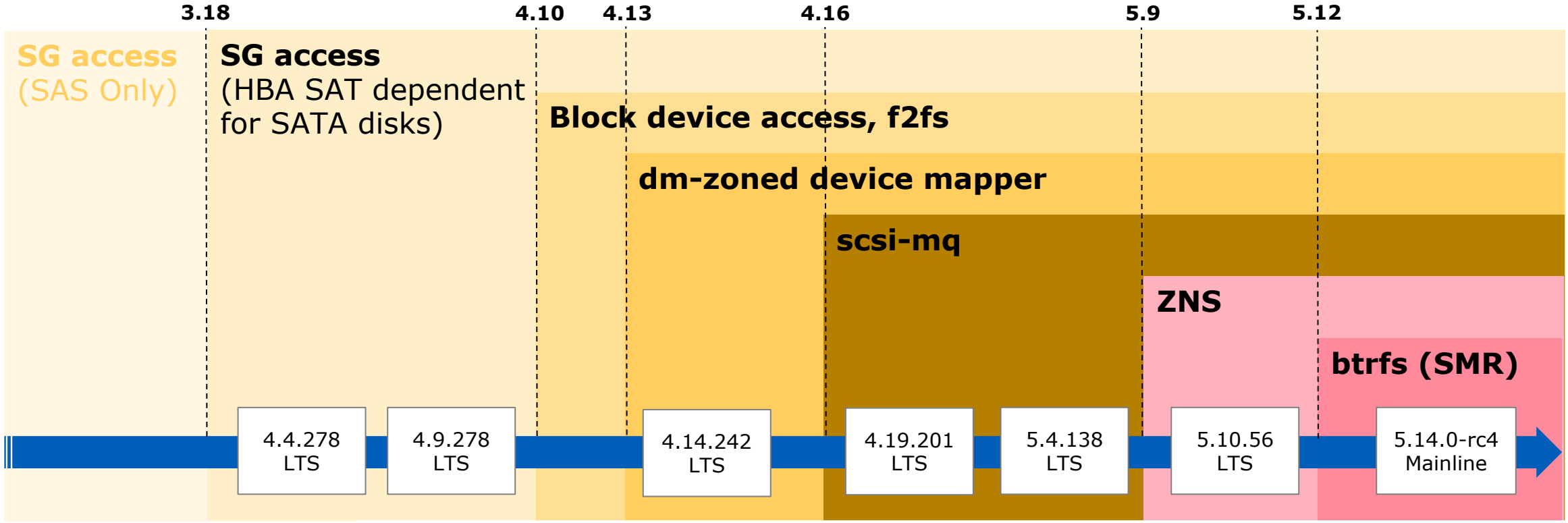
SMR HDDs and NVMe ZNS SSDs

- Shingled Magnetic Recording (SMR) hard-disks
  - Interface defined by the ZBC (SCSI) and ZAC (ATA) standards
  - Increase capacity without increasing device cost
- NVMe Zoned Namespace (ZNS)
  - Better command latency behavior (no device internal GC)
  - Reduced device cost (controller DRAM, over provisioning)
- Principle: LBA range divided into zones
  - Conventional zones: accept random writes
  - Sequential write required zones
    - Writes must be issued sequentially starting from the “write pointer”
    - Zones must be reset before rewriting (“rewind” write pointer to beginning of the zone)
- Users of zoned devices must be aware of the sequential write rule
  - Device fails write command not starting at the zone write pointer



Linux Kernel Support History

Started with SMR SCSI device type 0x14 support in kernel 3.18



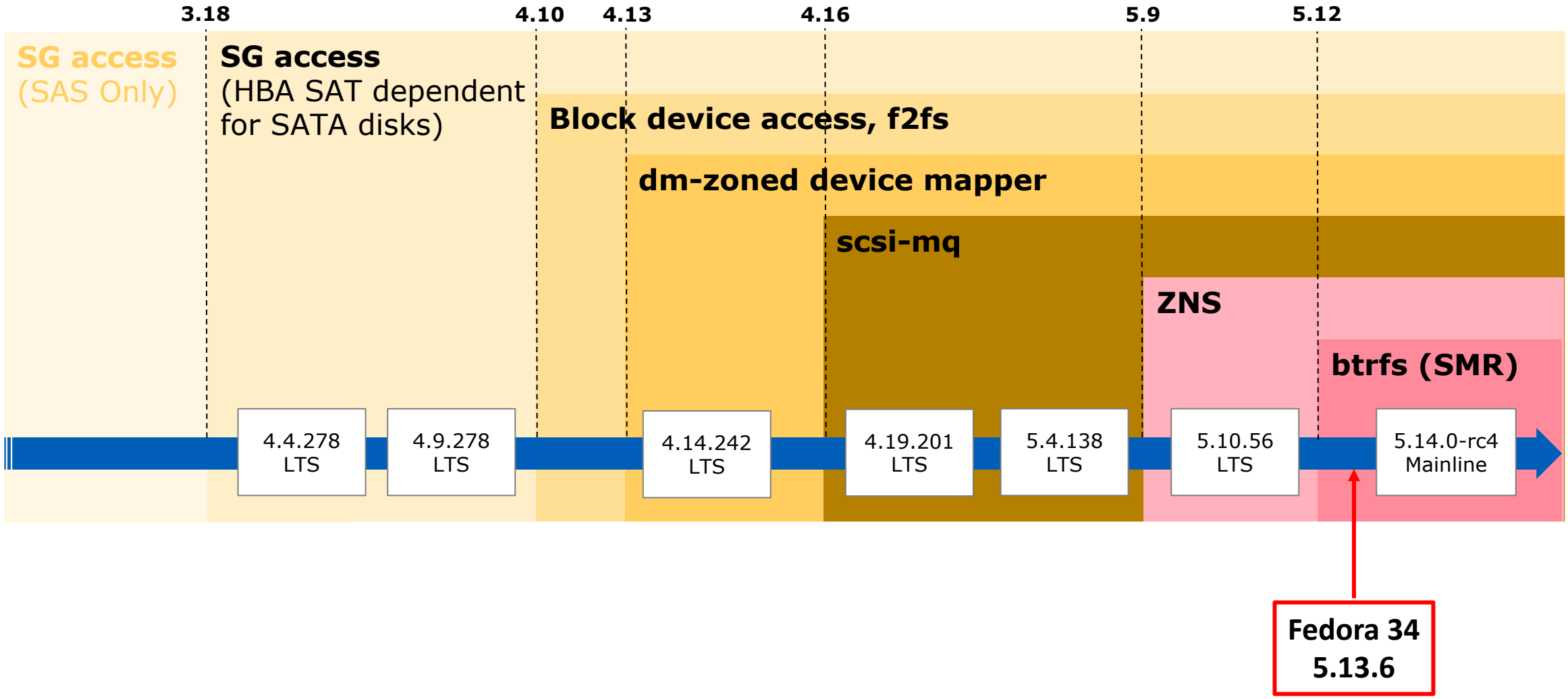
- Other notable milestones
  - Performance improvements in 5.0, zonefs file system in 5.6, zone append write emulation for SCSI in 5.8, dm-crypt support in 5.9
  - Constant maintenance, bug fixes and small improvements in almost every release

Zoned Storage in Fedora

Current Status

Fedora Kernel

Latest stable (almost) provides all supported kernel features



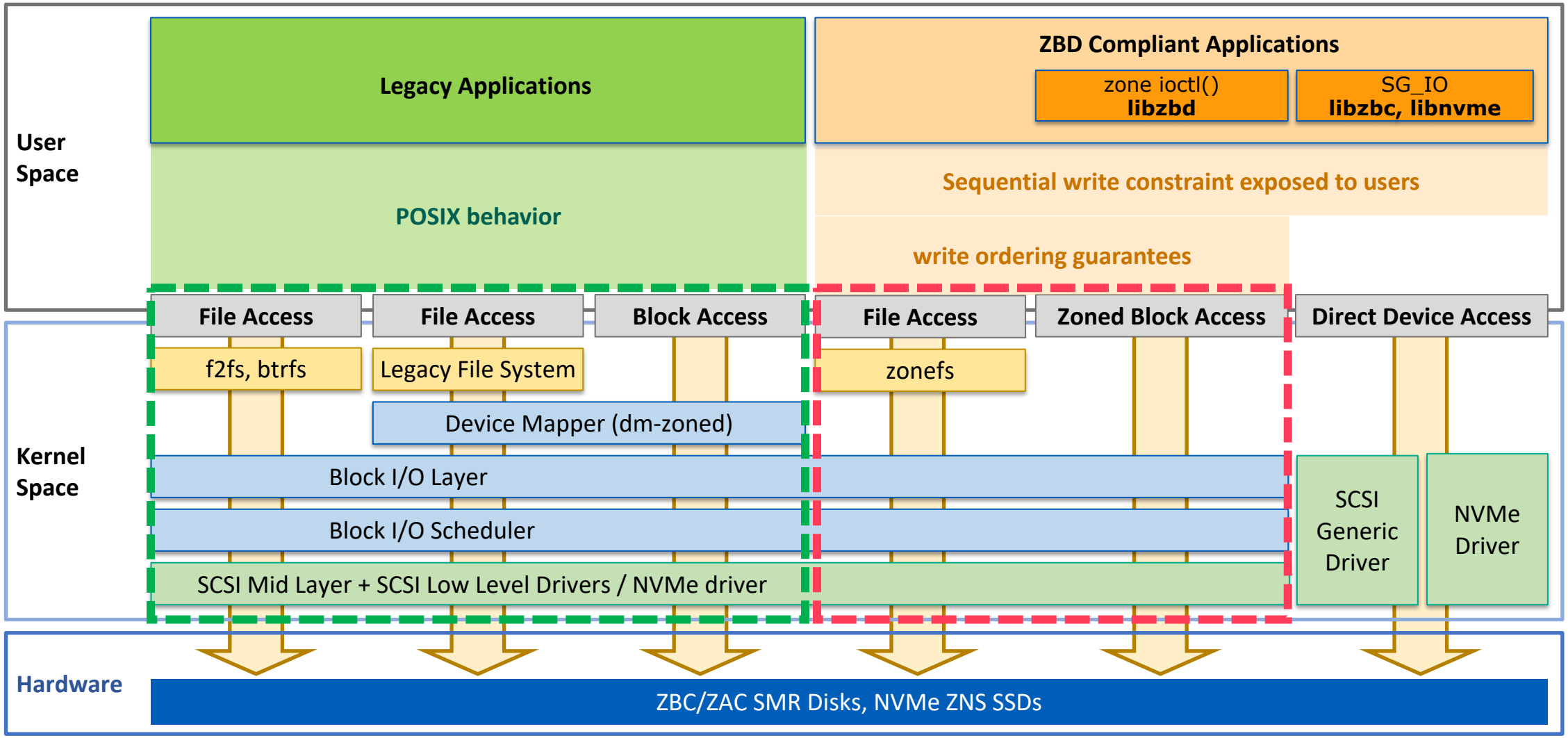
Fedora Kernel

Zoned block device support is enabled by default

- Fedora kernel is compiled with CONFIG_BLK_DEV_ZONED enabled
  - Since Fedora 26
  - No need to recompile the kernel !
- This enables:
  - Block layer user API (ioctls) and all kernel internal zone API
    - Zone append emulation in scsi disk driver
  - Device mapper targets support
    - dm-linear, dm-crypt and dm-zoned
  - f2fs and btrfs native zone support
- Zonefs (CONFIG_FS_ZONEFS) recently enabled by default
  - Fedora 33, 34 and 35
- Beware that zoned-btrfs is still in stabilization phase
  - Functional, but some problems remain

Fedora Kernel: User Application View

Fedora Linux enables all kernel supported device access paths



System Utilities

Recently improved with new packages added

- Current util-linux version 2.36.2 includes zoned device support
  - lsblk
  - blkzone
    - But still lacking ZNS zone capacity report added in util-linux 2.37
  - libblkid
    - zonefs support since version 2.36
    - But no zoned-btrfs support (patches in upstream tree but no libblkid tagged version yet)
- Current btrfs-progs 5.13.1 includes zoned-btrfs support
  - Added in version 5.12
- Recent package addition to Fedora 33, 34 and 35 (Thanks Neal !)
  - dm-zoned-tools
    - Provides the dmzadm utility for dm-zoned device mapper target format, check and control
  - zonefs-tools
    - Provides mkfs.zonefs (mkzonefs)

Applications

Not many things, yet ☺

- fio: Current version shipped is 3.26
  - Contains full support for zoned block devices benchmarking with “--zonemode=zbd” option
    - Added in fio 3.9
- No other packaged application that we know of support zoned storage
- But, thanks to the top-level kernel support, Fedora provides a great environment for developing, compiling and testing zoned storage support in existing or new applications
  - We use it in out lab and all our test racks ☺

Things Still Missing

Help welcome !

- libblkid version 2.38
  - For zoned btrfs
  - Waiting for Karel to release ☺
- nvme-cli ZNS support
  - Added to version 1.13, but 1.11.1 is currently shipping
  - Only need to request package upgrade !
- LUKS format for zoned dm-crypt
  - “cryptsetup luksFormat ...” is not writing the LUKS super block sequentially
  - Start offset must align to a zone start sector
- LVM integration of zoned device mapper targets
  - No automatic device mount
    - dm-zoned

Zoned Storage in Fedora

What is coming

Linux Kernel

Stable (production grade) zoned-btrfs is the primary target

- Complete stabilization of zoned-btrfs
  - Zone reclaim through automatic block group rebalancing
    - Extent write issue identified, fix coming !
  - Other bugs and various annoying problems need to be addressed
    - Issue tracker at https://github.com/naota/linux/issues
- btrfs improvements
  - De-clustered parity: erasure coded volumes
    - Include support for all RAID levels + stronger erasure coding schemes
      - Solve write-hole problem with CoW and journaling
    - Allows supporting RAIDed zoned-btrfs volumes
      - De-clustered parity needed because of zone append write use
- Other projects
  - De-clustered parity erasure coded DM target
    - For other applications and file systems than btrfs

Libraries and Applications

Many things coming !

- Fully functional btrfs tools and zoned-btrfs mount
  - The release of libblkid version 2.38 will solve all current problems
- libzbc (passthrough for SMR) and libzbd development libraries
  - Deployed in the field already (e.g. Dropbox uses libzbc)
    - https://github.com/westerndigitalcorporation/libzbc
    - https://github.com/westerndigitalcorporation/libzbd
- RocksDB ZNS support with ZenFS
  - Released on github https://github.com/westerndigitalcorporation/zenfs
    - Needs some cleanups and improvements (installer, man pages) before packaging
  - Depends on RocksDB 6.19.3 (plugin support)
    - RocksDB 6.15.5 is shipping currently
- Ceph
  - SMR native support in development with the new Bluestore engine
  - Work by Abutalib Aghayev with the Ceph team (Sage Weil)

Conclusion

We are in good shape !

- Overall, the current zoned storage support in Fedora is very good
  - Things work without needing to recompile everything from source
  - We always recommend Fedora for test & development related to zoned storage !
- Many things can still be improved or are still missing
  - Help is welcome ☺
  - We can mentor beginners too !
    - Contact us: damien.lemoal@wdc.com, johannes.thumshirn@wdc.com, matias.bjorling@wdc.com
- Zoned storage in Linux is extensively documented in zonedstorage.io
  - https://zonedstorage.io
    - Web-site content is open source on GitHub at https://github.com/westerndigitalcorporation/zonedstorage.io
  - Includes a Linux distribution page
    - https://zonedstorage.io/distributions/linux/
  - All examples shown are using Fedora ☺

Questions ?

Western Digital®