Zoned Storage MC

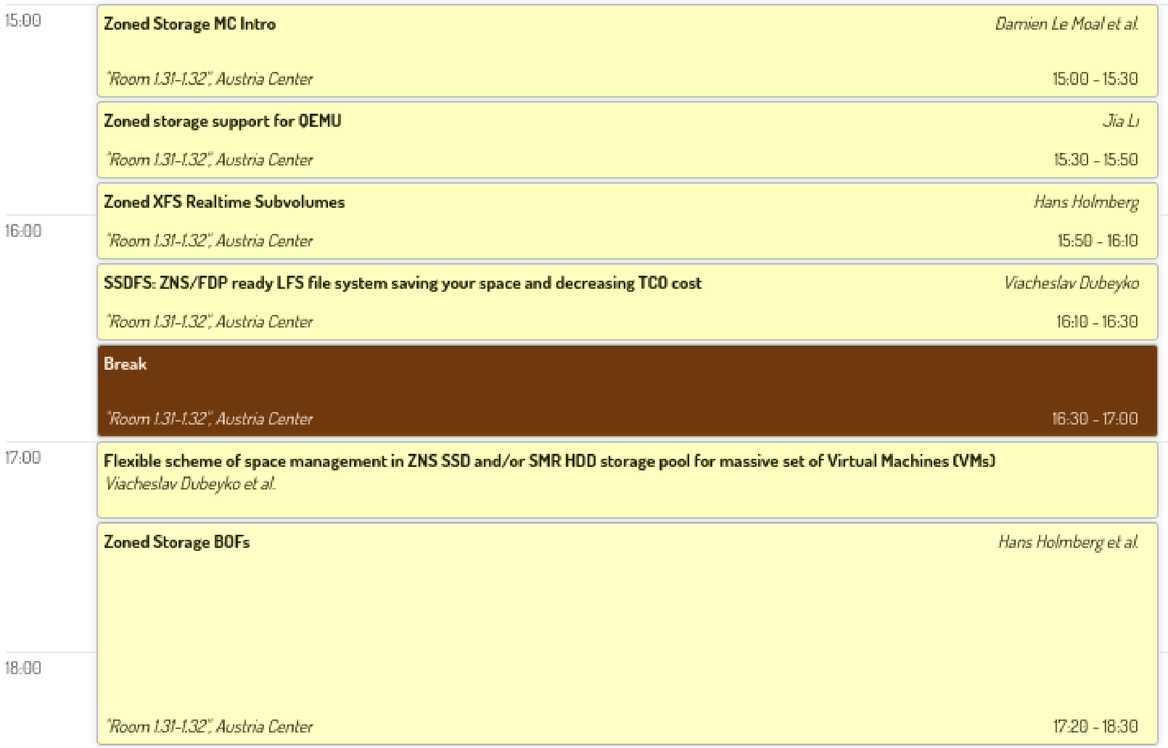
Hans Holmberg, Johannes Thumshirn

LINUX
PLUMBERS
CONFERENCE Vienna, Austria / Sept. 18-20, 2024

Overview

- First half
  - Retrospective and current upstream state
  - Quick updates
    - Damien Le Moal - IO Stack & zonefs
    - Johannes Thumshirn - BTRFS
    - Dennis Maisenbacher - Cloud & virtualization
  - In-depth presentations of ongoing efforts
- Second half
  - BOF / discussions
  - Feel free to suggest topics

- Remote participants
  - Raise hand / turn on video for comments/questions



State of Zoned Storage

LINUX
PLUMBERS
CONFERENCE Vienna, Austria / Sept. 18-20, 2024

The Evolution of the Zoned Storage Ecosystem

Research

[2]

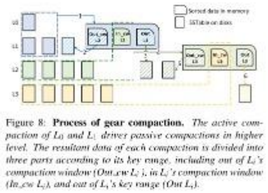


Figure 8: **Process of gear compaction.** *The active compaction of L_0 and L_1 drives passive compactions in higher level. The resultant data of each compaction is divided into three parts according to its key range, including out of L_i's compaction window (Out_{cw} L_i), in L_i's compaction window (In_{cw} L_i), and out of L_i's key range (Out L_i).*

[3]

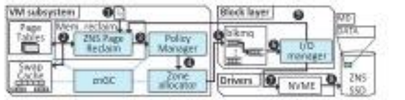


Figure 7: ZNSwap overview. Shaded shapes are internal ZNSwap components.

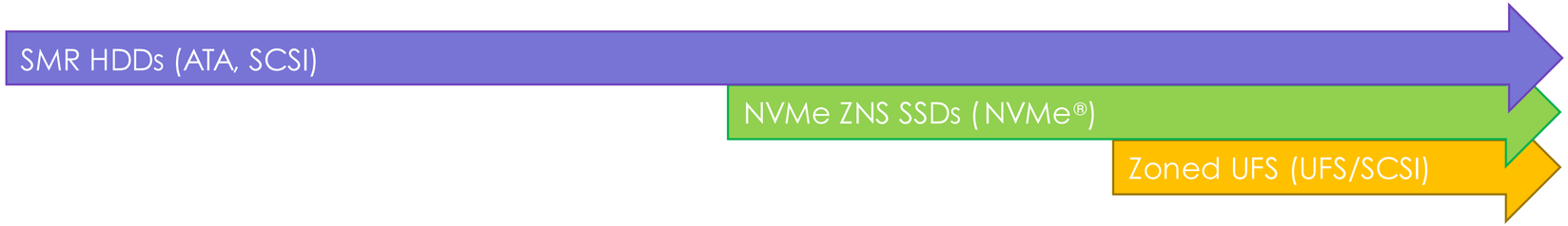
[4]

ZWAL: Rethinking Write-ahead Logs for ZNS SSDs with Zone Appends [5]

Linux(tm) eco system

[1]

Standards



[1] https://upload.wikimedia.org/wikipedia/commons/c/c2/Human_evolution_scheme.svg
[2] https://www.usenix.org/system/files/conference/fast15/fast15-paper-aghayev.pdf
[3] https://www.usenix.org/system/files/fast19-yao.pdf
[4] https://www.usenix.org/system/files/atc22-bergman.pd
[5] https://dl.acm.org/doi/abs/10.1145/3642963.3652203

Current state of the Zoned Storage Stack

Library/Tools support

- Libzbd, libnvme, xNVME, SPDK, fio, qemu, blkzone, blktests,

End-to-end Application Enablement

- MySQL, RocksDB, TerarkDB, ..

Local File-system support

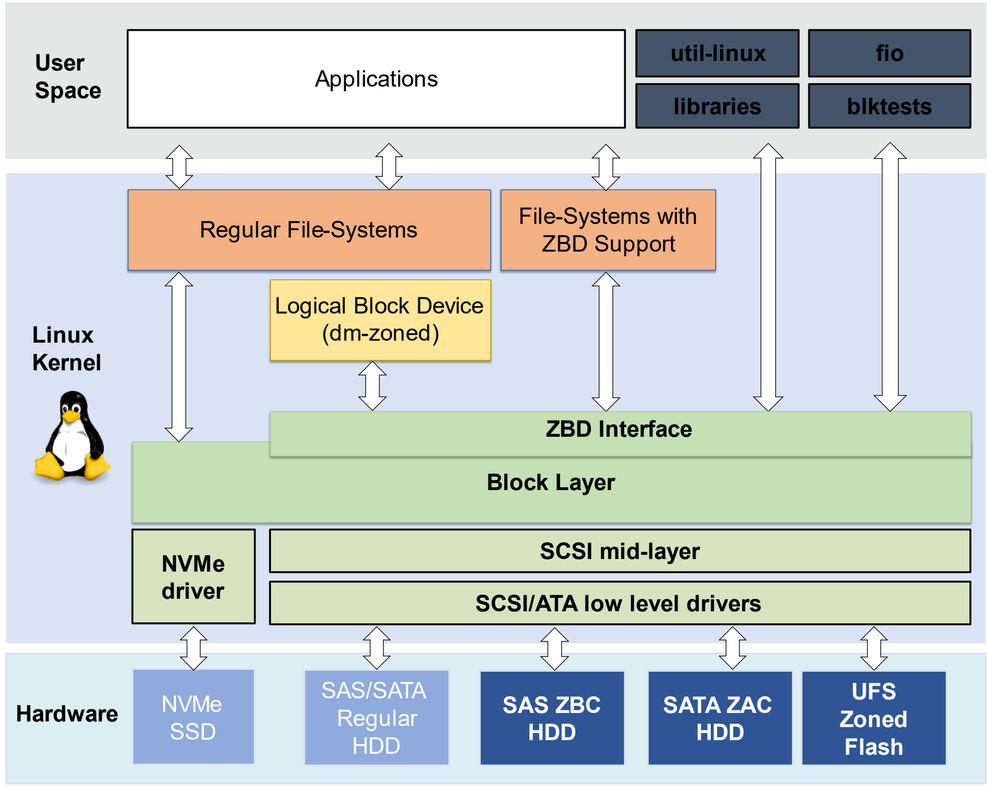
- f2fs (client – UFS)
- btrfs (enterprise – ZNS/SMR)
- xfs (enterprise, under development)

Storage Systems

- Ceph, OpenEBS, Mayastor, Saunafs, SPDK's CSAL, ,,

Distributions

- RHEL 9+, Fedora 33+, Debian 11+ and Ubuntu 21.04+



Mature, robust, and used in production by some of the biggest consumers of storage

I/O Stack (Block Layer, DM, SCSI/NVMe)

- New per-zone sequential write ordering control with zone write plugging (v6.10)
  - Same ”at most one write per zone in flight at any time” principle as with zone write locking (mq-deadline) but control is applied to BIOs instead of requests
- Advantages:
  - Remove dependency on mq-deadline (zone write locking): any block I/O scheduler can be used, included “none”
  - Generic zone append emulation implementation
    - Simplifies DM and SCSI sd code
  - Significant performance improvements in some cases, including for read operations
- Drawbacks
  - Not many so far (identified performance degradation with some SMR drives on zone boundary crossing)

zonefs

- Some small changes
  - Error recovery bug fix (v6.8)
  - Conversion to new mount API (v6.9)
  - Large folio support added (v6.10)

- Started exploring zone append user interface through io_uring
  - Changes to io_uring and iomap needed
  - New write append operation will return the written offset to the user
    - Applicable to regular O_APPEND writes

BTRFS

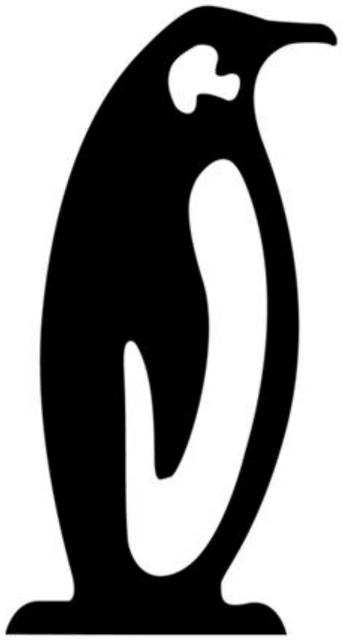
- Initial support upstream v5.12
- Garbage collection added in v5.13
- NVMe ZNS support added in v5.16
- (experimental) RAID0/1/10 support added in v6.7

- Stabilization is still on-going
  - Early ENOSPC errors
    - Metadata overcommit
    - Space accounting
  - RAID support

Cloud storage & Virtualization

	SPDK	ceph	Mayastor OpenEBS	QEMU
Upstream	• Initial ZNS support since v20.10 with ongoing effort[1]	• Crimson – seastore ships with initial ZBD support (Reef stable release)[3][4]		• Initial ZNS emulation[6] • Attaching ZNS devices via vfio PCI passthrough or virtio-blk[7] • Attaching SMR disks via virtio-scsi or vhost-scsi[8]
In progress	• CSAL - ZNS patches to be upstreamed[2]	• Fix for active zone resource exhaustion	• Open PR for initial ZNS support under discussion[5] • CSAL Integration?	• qcow2 support for zoned storage emulated devices[9]

[1] https://spdk.io/release/2020/10/30/20.10_release/
[2] https://dl.acm.org/doi/abs/10.1145/3627703.3629566
[3] https://docs.ceph.com/en/reef/dev/zoned-storage/
[4] https://docs.ceph.com/en/latest/releases/reef/
[5] https://github.com/openebs/mayastor/pull/1298
[6] https://lists.nongnu.org/archive/html/qemu-block/2020-06/msg00720.html
[7] https://www.qemu.org/docs/master/devel/zoned-storage.html
[8] https://zonedstorage.io/docs/tools/qemu#qemu-virtio-scsi
[9] https://lore.kernel.org/all/20240122184830.40094-1-faithilikerun@gmail.com/



Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024