OCP SUMMIT

March 20-21
2018
San Jose, CA

OPEN. FOR BUSINESS.

Linux Kernel Support for Hybrid SMR Devices

Damien Le Moal, Director, System Software Group,
Western Digital

Outline

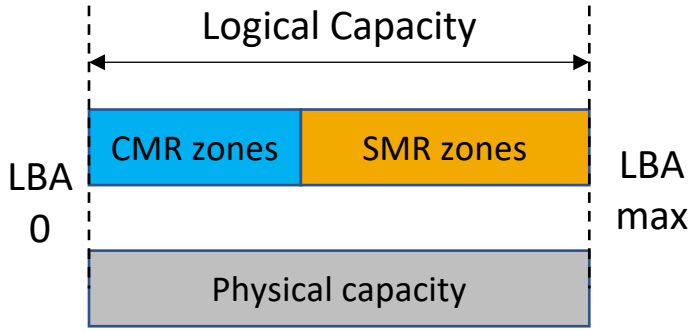
- Hybrid SMR device host view
  - Changes from ZBC
- Kernel block I/O stack support
  - Background: ZBC/ZAC support
  - Hybrid SMR devices identification and initialization
  - I/O path (sequential write constraint)
- Other software
  - File systems and device mapper
  - User level libraries and tools

Hybrid SMR Devices Host View

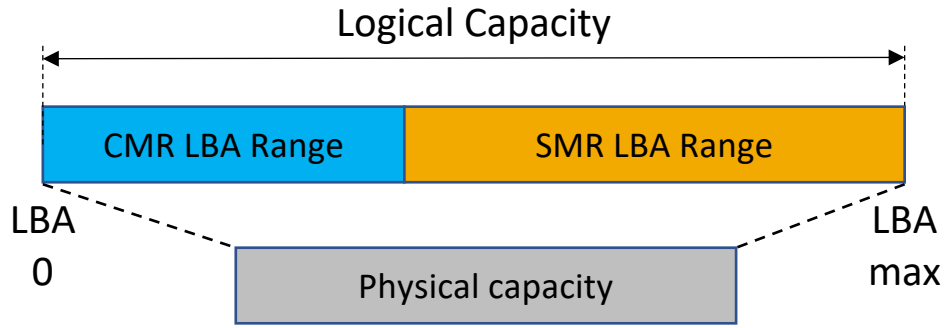
Fixed capacity with thin provisioned LBA space

- LBA space is fixed and larger than the effective physical capacity
  - Thinly provisioned fixed logical capacity
- LBA space is split between CMR space and SMR space
  - CMR space: conventional recording space starting at LBA 0
  - SMR space: shingled sequential recording space starting after the CMR space
    - SMR space size is the maximum physical capacity of the device
  - Physical sectors maps either CMR LBAs **or** SMR LBAs

Regular SMR disk (ZBC/ZAC)



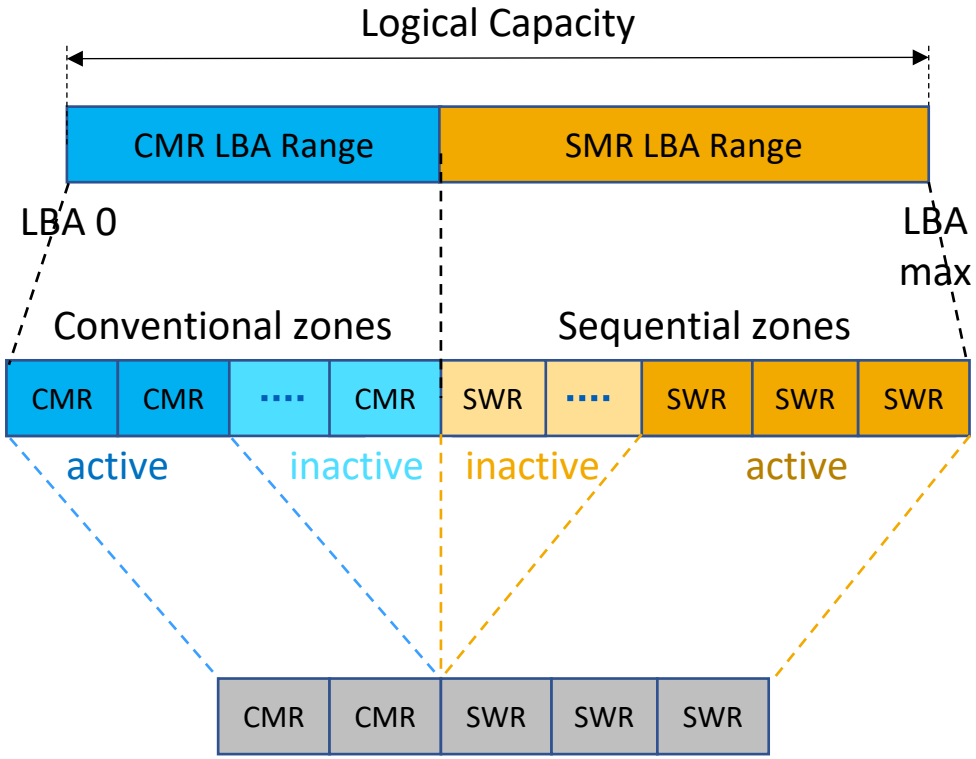
Hybrid SMR disk



Hybrid SMR Devices Host View

Zones and zones mapping

- LBA space is fully described by zones
  - Similarly to regular ZBC/ZAC devices
  - The CMR LBA space is divided into a set of conventional (CMR) zones
  - The SMR LBA space is divided into sequential write required (SWR) zones
- Not all zones have physical mapping
  - "inactive" condition
    - No physical storage mapping
  - Zone conversion operation changes zone condition from inactive to active
    - Changes physical storage mapping between CMR LBAs and SMR LBAs
    - Zone conversion is a data destructive operation
  - The total of the size of all active zones is equal to the usable storage capacity
    - Always lower than the device logical capacity



Hybrid SMR Standard

Block device type, new feature set, new commands

- Standard development is on-going
- Device type/signature is 0x00
  - Regular block device
- 3 new commands defined
  - **Media Convert***: change the mapping state of a set of contiguous zones
  - **Media Report***: report the list of contiguous zones that constitute an optimal conversion unit
    - For devices that have conversion constraints
  - **Media Query***: report the result of an hypothetical conversion
    - For devices with a zone conversion granularity of one zone
- Zone information handling remains mostly unchanged
  - Zone types are constant
  - Zone conditions change based on the zone usage
  - Report zones command gives zone information
  - Zone state machine unchanged for SWR zones
- New ”inactive” (not provisioned) zone condition added for all zone types
  - Indicates that a zone currently has no physical storage assigned

*Not fixed, being discussed in the standard groups

Hybrid SMR Standard

Conventional zones

- Activation of conventional zones requires re-format of the physical storage
  - Ex: change from shingled tracks formatting previously used for SWR zones to un-shingled tracks
  - Requires a full sequential write pass of the zone tracks
- 2 possibilities
  - Drive managed approach: conventional zones initialization is done automatically on activation
    - "foreground" conversion
    - Can take time
  - Host managed approach: conventional zones are activated in an uninitialized state
    - Fast activation, but host must do a first sequential write pass to fully re-format the zones
- Preserve ZBC backward compatibility by allowing implicit initialization on write*
  - Enforce or not write-at-write-pointer checks (WPC) for initial sequential write pass
  - User configurable bit
  - With WPC off: device handles writes past the write pointer automatically by filling the gap from the WP position
    - Allows random writes anywhere in the zone
    - Can increase write latency

*Being discussed in the standard groups

Hybrid SMR and ZBC/ZAC Comparison

Host-managed vs hybrid host-managed summary

Feature	ZBC/ZAC (HM)	Hybrid SMR	Note
Device type	0x14	0x00	Allows operation under regular block device rules with all storage mapped into CMR LBA space (RC_BASIS=0)
LBA space	Fully described with zone descriptors from report zones		New "inactive" zone condition to indicate physical mapping state
SMR zones	Sequential write required		Unchanged state machine for active zones
CMR zones	Conventional	Conventional with WP check off	May incur higher write latencies during zone initialization
		Conventional* with WP check on	Random writes allowed only below WP position

* May be introduced as a new zone type

Background: Linux Kernel ZBC/ZAC Support

Added to kernel 4.10

- Current support restricts device characteristics
  - All zones must be the same size, except for a last eventual runt zone
  - The zone size must be a power of 2 number of LBAs
  - Reads in SWR zones must be unrestricted
    - No errors returned for read after a zone WP
  - sysfs files indicate device model and zone size
    - "host-managed" or "host-aware"
- Sequential write constraint of SWR zones is exposed as is to the device user
  - File systems, device mappers or applications
- In-kernel I/O path guarantees in-order write command delivery to the HBA
  - Per SWR zone write lock limits in-flight write commands to one at most
  - Conventional zones not affected

```
[    3.687797] scsi 5:0:0:0: Direct-Access-ZBC ATA      HGST HSH721414AL TE8C PQ: 0 ANSI: 7
[    3.696359] sd 5:0:0:0: Attached scsi generic sg4 type 20
[    3.696485] sd 5:0:0:0: [sdd] Host-managed zoned block device
[    3.865072] sd 5:0:0:0: [sdd] 27344764928 512-byte logical blocks: (14.0 TB/12.7 TiB)
[    3.873046] sd 5:0:0:0: [sdd] 4096-byte physical blocks
[    3.878343] sd 5:0:0:0: [sdd] 52156 zones of 524288 logical blocks
[    3.884591] sd 5:0:0:0: [sdd] Write Protect is off
[    3.889440] sd 5:0:0:0: [sdd] Mode Sense: 00 3a 00 00
[    3.889458] sd 5:0:0:0: [sdd] Write cache: enabled, read cache: enabled, doesn't support DPO or FUA
[    4.253140] sd 5:0:0:0: [sdd] Attached SCSI disk
```

```
> cat /sys/block/sdd/queue/zoned
host-managed

> cat /sys/block/sdd/queue/chunk_sectors
524288
```

Hybrid SMR Kernel Core Support Plan

Drive identification, initialization and I/O path

- Device type 0x00 combined with hybrid SMR feature set enabled identifies a hybrid SMR device
  - Defines a new zone model "hybrid-host-managed"
  - Mostly backward compatible with regular host-managed
- Zone size and other kernel restrictions still apply
  - Unchanged from ZBC
- Zone write locking changes needed to include conventional zones when conventional WP check is enabled
  - Enable write locking of conventional zones to ensure successful first sequential write pass
  - As the kernel does not track zone conditions nor zones WP, zone write locking will be enabled for all zones
    - Zone write locking applies to ALL zones of the device
- No strong use case for in-kernel implementation of Hybrid SMR commands
  - No API for media report, media query and media convert, at least initially
    - User level tool

Other Kernel Components Support

Device mapper drivers and file systems

- DM core changes needed to handle new zone models and conditions
  - Similarly to ZBC disks, filter zone models stacking
  - Allow use of dm-linear driver
    - Expose fully provisioned zoned block device
    - Create zoned block devices fully compatible with hybrid-host-managed model
- dm-zoned driver changes to handle new zone types and conditions
    - Create regular block device on top of any disk zone configuration
    - Strong candidate for in-kernel convert support to dynamically adjust zone mapping based on I/O activity and data aging
- F2fs support still valid
  - Can handle conventional zones first initial sequential write pass with no overhead
  - But initialized conventional zones needed for in-place metadata updates
    - Conventional zone initialization can be handled at format time for WPC on cases
  - However, 16 TB limit on device size is a problem
    - Will need extensive changes to overcome
    - Limit can by temporarily bypassed using dm-linear
      - Fully provisioned compact LBA space
- BtrFS support (work on going) unchanged
  - Larger LBA space not a problem (16 EiB limit)
  - Conventional zones initialization pass can be supported
    - Similarly to f2fs, only few initialized conventional zones needed

User Level Tools

Libzbc, sg3-utils and tcmu-runner

- Libzbc will include hybrid SMR device support
  - New commands API and tooling for zone conversion
  - Probably will not include fake driver support
    - tcmu-runner better emulation solution
- Sg3-utils tools
  - Media report, query and convert
- Device emulation with tcmu-runner
  - Built on top of ZBC emulation device already included
  - Useful for testing kernel software and user tools on top of various drive configurations

- To get involved with or to learn more about Hybrid SMR, join the OCP Storage Project Group!!!
  http://www.opencompute.org/projects/storage/
- Project meets every 2nd Thursday of the Month
  - Rotates between 9am PT and 4pm PT
  - Next Project Call is on April 12 @ 4pm PT
- To Register:
  - Access Link: https://attendee.gotowebinar.com/register/4800086622601834753
  - Webinar ID: 542-942-379

Questions ?

OCP SUMMIT