SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2017

ReFS Support For SMR Drives

Raj Das
Microsoft

Agenda

- ReFS V1 Primer
- ReFS V2 Intro
- Features and Motivation
- SMR Support

ReFS V1

- Windows allocate-on-write file system.
  - Metadata stored in B+ tables.
- Merkle trees verify metadata integrity.
  - Optional for data.
- Inline recovery from single copy corruption.
- Inline chkdsk.
- Periodic Scrubbing of data/metadata.

ReFS V2

- Windows Server 2016.
- Focused on Server Workloads.
  - VMs, Private Cloud.
- Parity Friendly IO pattern.
- Dynamic Tiering in IO Codepath.
- Redo logging.

ReFS V2 (Contd.)

- Sparse VDL
  - Efficient tracking of initialized/uninitialized data on disk.
- Parallel efficient allocator.
  - ~150K allocations/sec sustained.

SMR

- Three Kinds of SMR disks.
  - Drive Managed.
  - Host Aware.
  - Host Managed.
- SMR disk broken up into bands, of 256MB.
- Host managed SMR requires strictly sequential write pattern within a band.

ReFS SMR Support (Requirements)

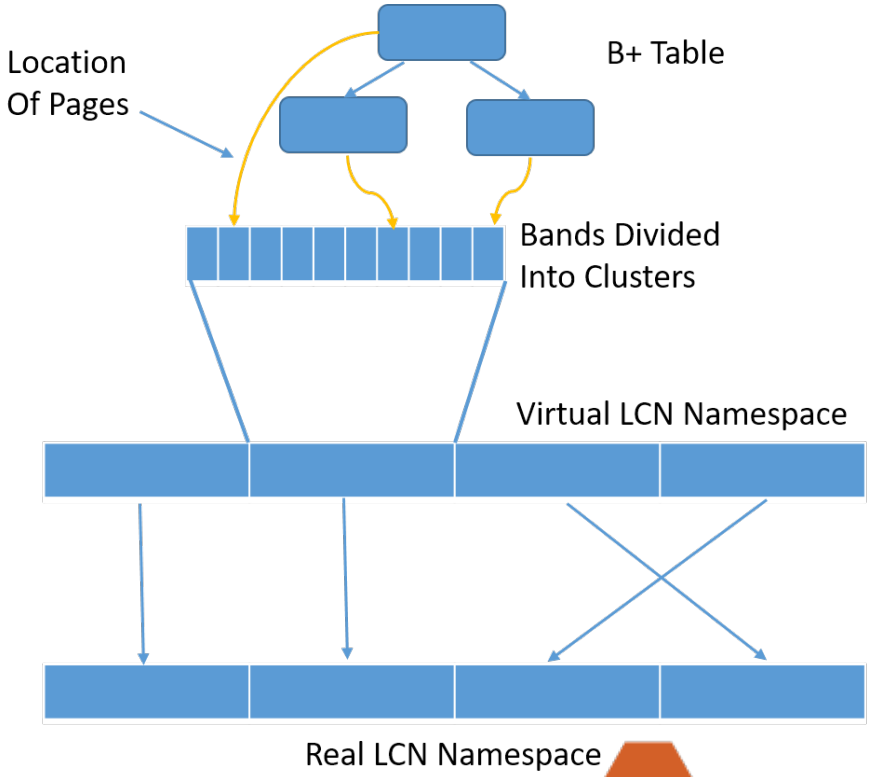
- Minimal Change to application, should just work.
- Multiple files written simultaneously.
  - Serialized write preferred but not required.
- Must give serialized IO throughput of disk.
- Application retains control.

ReFS V2 Features

- Policy Driven Allocation.
  - Policy Module guides where to allocate from.
  - Sets boundaries.
- Tiering Aware Allocation.
- Deferred Allocation Logic.
  - LCNs assigned when needed, i.e. IO codepath.

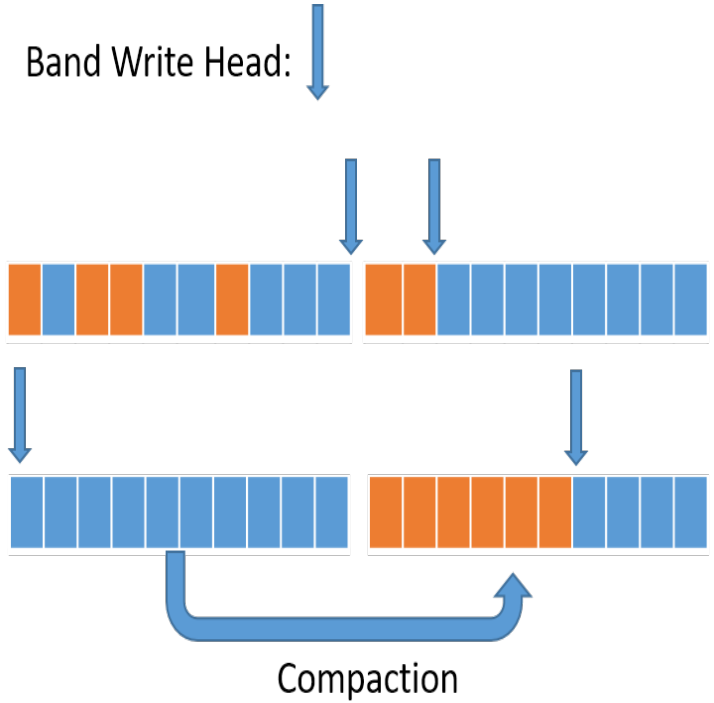
ReFS V2 Features

- Cluster Bands
- 64MB/256MB in size.
- Can swap Bands.
- Per Band matadata tracked independently.



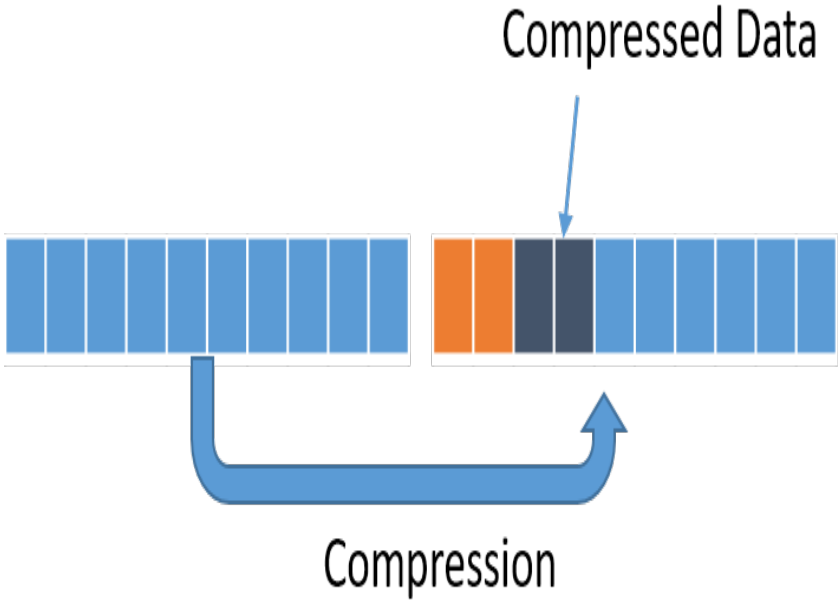
ReFS V2 Features

- Compaction
  - Read Band with holes.
  - Write only allocated regions.
  - Update Index.



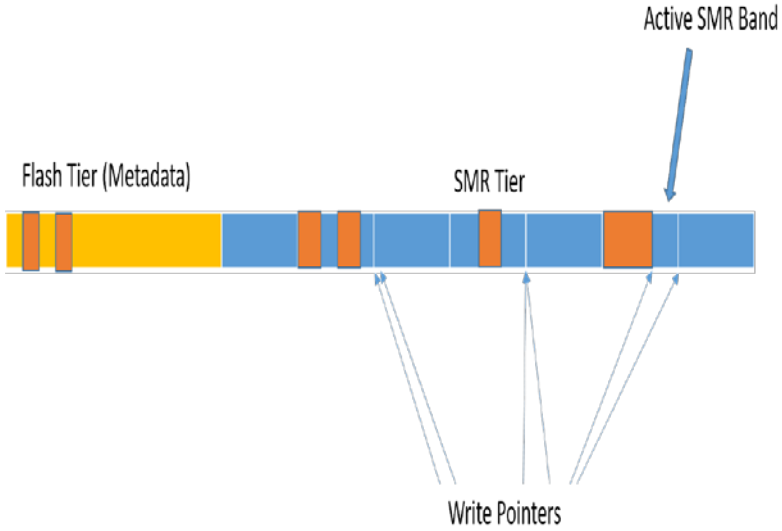
ReFS Features

- Compression
- Compaction + Compress data before writing.
- Supports:
  - LZNT1
  - LZX
  - XPRESS
  - XPRESS_HUFF



ReFS SMR Support

- 256MB sized cluster band.
  - Maps 1:1 to SMR zone.
- Storage Spaces.
  - Create Tiered Disk.
  - SSD Tier, SMR Tier.



ReFS SMR Support (Contd.)

- All metadata in SSD(Flash) tier.
- All Data in SMR Tier.
- Deferred Allocation Logic.
- Small data IOs redirected to single band.

ReFS SMR Support (Contd.)

- Cluster band tracks it's SMR Zone write pointer.
- Strict IO serialization logic per band.
- Next IO sent after preceding IO queued to HW.

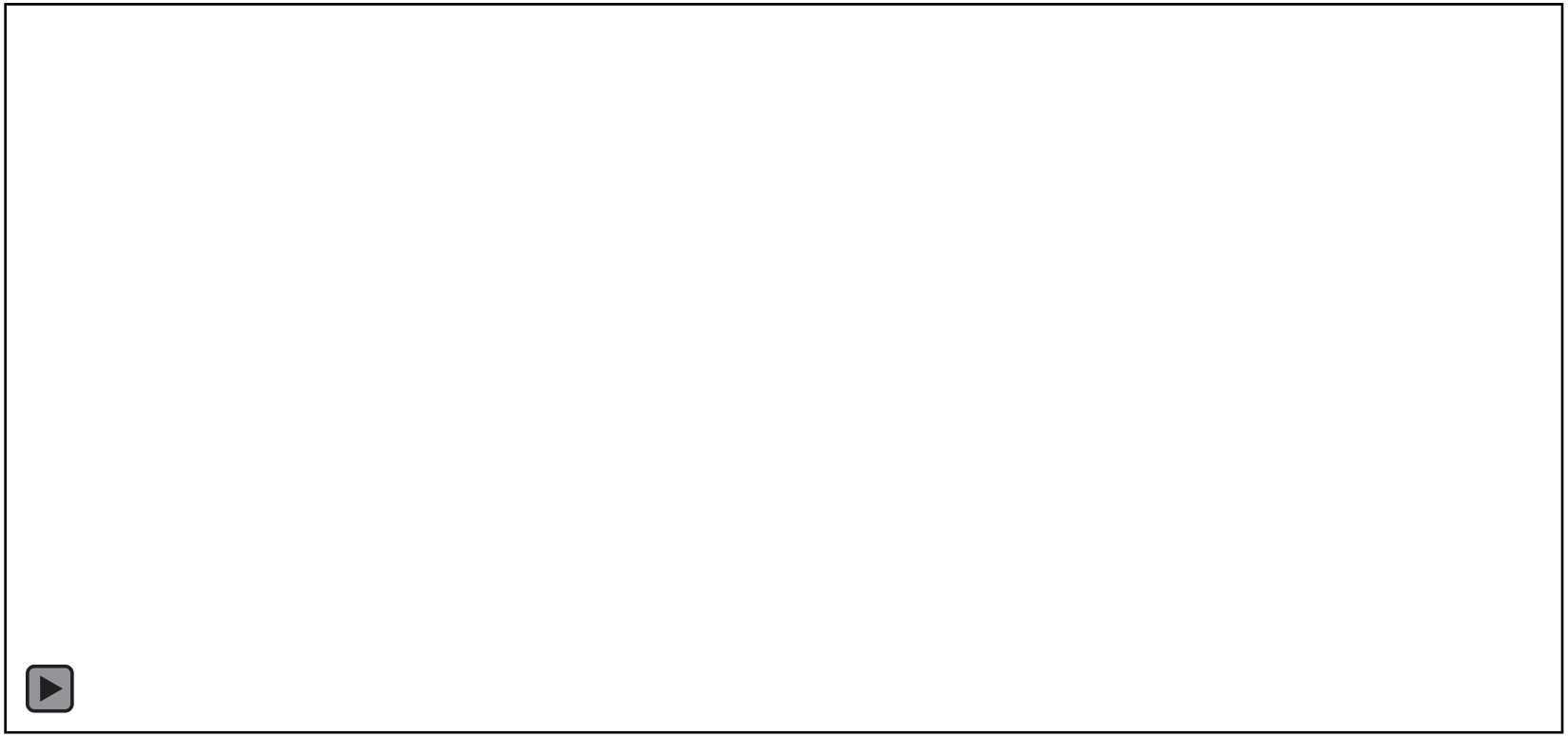
ReFS SMR Support (Contd.)

- ReFS Tracks
  - Free Space in SSD Tier.
  - Free Space in SMR Tier.
  - Writable Free space in SMR Tier.
  - GC space in SMR Tier = (FreeSMR – WritableSMR)
- Values reported by FSCTL.
- GC controlled by application.
  - GC pausable/resumable, and abortable.

ReFS SMR Support (Contd.)

- GC Overhead Concerns:
  - Steals bandwidth from applications.
- Large Files get their own dedicated Bands.
- Small Files share bands.

Demo



- Questions?