FileSystems - Some Observations

Sanjeev Bagewadi

Filesystems : What are they ?

- A software (stack ?) which help store/retrieve data
- Ondisk filesystems : UFS, FFS, XFS, ZFS… *FS…
- Network Filesystems (Protocols ?)
  - NFS, SMB/CIFS
- Distributed FileSystems :
  - Andrew File System(DFS), IBMs GPFS and many more

What do i work on ?

- Work for Nutanix
  - A HyperConverged solution
  - Has its own distributed filesystem : **NDFS**
  - More details : https://nutanixbible.com
- In Nutanix Files team
  - A software defined scale-out NAS providing NFS/SMB access to clients
  - A scale-out filesystem layered on top of NDFS
  - Uses a derivative of OpenZFS

Problem: APIs for Backup Vendors

- Needed APIs for Backup Vendors to do the following :
  - Take snapshots of the filesystem
  - Report changes between snapshots
  - Should scale for millions of changed files/directories and PetaBytes of storage
- Scalability Requirements :
  - Constant time or O(n) at max
  - Least amount of metadata churn

‘zfs diff’- Overview and Some bottlenecks

- Outline :
  - Reports blocks (holding dnodes) which have changed
  - Not all dnodes reported are modified (false-positives)
    - Filtering applied to identify changed dnodes
    - Translate inode to filename/path
      - Sequentially search for inode through all entries
- Bottlenecks :
  - ‘zfs diff’ - Single threaded
  - Translating ‘inode’ to file-path is expensive
    - Sequential search
- Detailed in the talk : OpenZFS Dev Summit 2020

Some observations

- Ondisk layouts play a major role
  - Blkptr : birth time (TXG) plays a major role in identifying changed blocks
  - Indexed directory entries help search quickly - O(1).
- Typical challenges of Filesystems:
  - Not all metadata will fit in memory.
    Need intelligence to pick the right metadata to load.
  - Changes to Ondisk layouts last really long...
    - They come back to haunt you :-) much later :-)
    - Once, error is introduced... fixing it on the fly may be challenging.
  - Ensuring ondisk crash consistency along with perf
  - Problems don't go away with reboots :-)
  - Ondisk corruptions can be challenging to debug