Operations and attributes related to Virtualization

draft-iyer-nfsv4-space-reservation-ops-00.txt 2010-03-23

Rahul Iyer
Mike Eisler
Deepak Kenchammana
James Lentini

Go further, faster™

Introduction

- Rapid uptake in Virtualization technology
- Virtual disks often stored on NFS volumes
- However, virtualization workflows not fully supported
  - No way to reserve space backing a file
  - No way to tell how much space will be freed on deletion
    - Dedupe complicates this
  - No way to release sub regions of the file

Space reservation

- Hypervisors often want to preallocate virtual disk files
  - Prevents ENOSPC errors during VM operation
- Current practice is to zero out the file
  - Inefficient
  - Not guaranteed to work on deduped filesystems
- Need a space reservation op

space_reserve attribute

- Specifies whether the blocks backing the file have been allocated
- Readwrite attribute
- Attribute is per file
- Set via SETATTR
- MUST ensure every byte in the file can be written to/overwritten
  - Writes can never return ENOSPC
- SETATTR fails with NFS4ERR_NOSPC if allocation cannot be guaranteed

Space freed by deletes

- Virtual machine migration common in virtualized workflows
  - Storage/CPU load balancing, capacity etc.
- Need to know which is the “best” VM to migrate
  - E.g: VM that frees up most space
- Current NFS attributes
  - size
  - space_used
- Inadequate in filesystems with snapshots, dedupe etc.
  - Problems with accounting

space_freed attribute

- Specifies the number of bytes freed if the file is deleted
- Count of the number of blocks unique to the file multiplied by the block size
- Read only attribute
- Attribute is per file
- Might be hard to compute within an NFS timeout
  - Alternatively, can be an NFS op with the result in a callback

Hole Punch

- Virtual disks are essentially filesystems within files
- Deleting files within the guest filesystem is not communicated to the file server
- SCSI introduced a TRIM command to communicate deletions to the storage device
- Virtual SCSI layer can translate TRIM into an NFS op

HOLE_PUNCH op

```
struct HOLEPUNCH4args {
    /* CURRENT_FH: file */
    offset4     hpa_offset;
    length4     hpa_count;
};

struct HOLEPUNCH4res {
    nfsstat4    hpr_status;
};
```

Hole Punch Considerations

- Hole Punch is a hint
  - Server need not support it
  - Space deallocation is optional too
- Subsequent reads to the region MUST return zero if HOLEPUNCH succeeds
- Block deallocation can be deferred
  - But subsequent reads MUST get zeroes
- Decreases space_used
  - space_freed might not change
- Should not affect space reservation policy
  - Writes to the HOLEPUNCH'ed region cannot fail with ENOSPC