VFS/Filesystems Discussion

Suparna Bhattacharya
suparna@in.ibm.com

(with inputs from linux-fsdevel, Zach Brown, Chris Mason, Mike Fasheh, Ram Pai, Dipankar Sarma, Badari Pulavarthy, Dave Howells)

Commonality, Maintainability, Abstractions

- filemap.c maintainability and path lengths - simplify ?
  - e.g. 9 generic_file_xxx routine variations for write !
    - nolock X DIO X AIO X vectored
    - better fsync/O_SYNC abstraction
- Is there a better way to deal with DIO vs buffered ?
  - O_DIRECT as a mount or chattr option
    - alignment ?
    - avoids simultaneous DIO and buffered
  - Preallocation w/o instantiation
    - zero-filled returns (high water mark generalization)
- Are buffer heads still a problem ?
  - Used in fallback paths like block_size != page_size
    - alternative: io count per page
  - Used by ext3 journalling
    - alternative : introduce a new ordered mode
  - Concerns
    - low mem usage on x86, SLB misses on ppc64 ?
    - additional code paths to maintain
- Alternative to bmap ?
  - Efficient handling of sparse files

Enhancements in Generic Code ?

- Generic delayed and multiblock allocation, extents, nobh
  - How many filesystems would benefit - aka how generic ?
    - transaction vs file as unit of writeout
    - ABISS requirements ?
    - space reservation - PG_DELALLOC
    - ext3 journal lock and multiple lock_page ?
- Lock ordering & concurrency - fs specific and generic
  - copy_from_user deadlocks (mmap + write)
    - ordering locks by superblock+inode+offset (cluster fs)
  - i_sem concurrency ?
  - Cluster filesystems
    - more locking hooks - cluster locks across operations
    - sleeping ->drop_inode()
- Zero-copy, DIO and caching
  - sendfile on O_DIRECT, pipe_buffer type move pages between fds
  - filesystem caching, NFS superblock caching (dhowells ?)
  - limiting page cache memory usage (mem mgmt topic ?)
- AIO ?
  - BOF at OLS to consolidate AIO users
- Error handling ?
  - Removable media, ENOMEM, EIO, ENOSPC

Namespace and Dcache

- Shared sub-trees
  - "Mirror" bind capability that is active in nature
  - RFC from Al Viro, elaborated by Bruce J Fields
  - Unclonable feature to avoid exponential increase in vfsmounts
  - Patch from Ram Pai on linux-fsdevel
    - http://www.sudhaa.com/~ram/readahead/sharedsubtree
    - review needed : pnode traversal, attach-recursive_mnt
    - testcases in development
    - todo: util-linux support, better visual tools
- Parallel updates in dcache
  - dcache lock contention on some workloads
- Dcache memory fragmentation
  - Better reclamation logic
    - rbtree/treap based on dentry address instead of LRU ?
    - needs parallel updates as it increases dcache lock contention

Disclaimers and Trademarks

This work represents the view of the authors and does not necessarily represent the view of IBM.