Background

- **Dawning Cluster File System**
  - A file-server based cluster file system for Linux clusters, especially for Dawning 4000L
  - Based on local file system (logical file system), e.g. ext2
  - Multiple meta data servers and storage servers
  - File data striping across storage servers.

Motivation

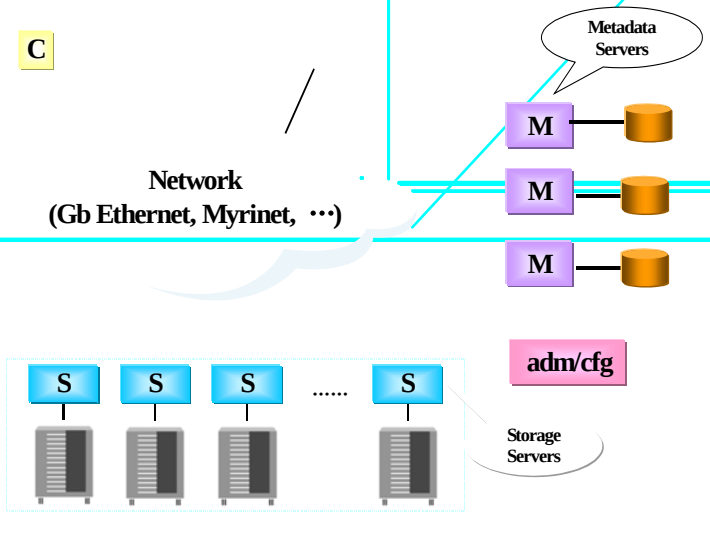
- Improving metadata processing performance,include file/directory creation, removal, lookup, etc
- Enable DCFS to support large directories

Key issues

- Limitation imposed by local 32-bit Linux system
  - A single local file can only contain limited number of directory entries
- No uniformly distributed hash function can be found
  - The maximum file size limitation can sooner be reached by skew data
  - Low space utilization of directory entry file
- Directory sequential scan operation
  - How to efficiently avoid getting same entry repeatedly

Our Implementation

- Limited Multi-level Extendible Hashing (LMEH)
  - Two separated files assigned to each single directory, in different role
    - ✓ Index file for hash table
    - ✓ Dentry file store directory entry blocks
  - New directory entry block allocated at the tail of dentry file
    - ✓ Dentry file will be more denser and reduce number of holes
    - ✓ Increase the ratio of space utilization.
  - Delayed-splitting scheme and multi-level hash tree
    - ✓ Different buckets indexed by different number of hash value bits
    - ✓ Reduce the times of hash table splitting
  - Design a special searching regulation for sequential directory scan operation.
    - ✓ No additional context needed in the entry —— compatible with POSIX standard.
    - ✓ Avoid repeated getting same entry
  - Limit the maximum bits number of hash value
    - Avoid splitting operation breaking through the local system file size limitation
    - Chained all blocks with same hash value when the maximum bits number reached



Dawning Cluster File System Architecture

Effect of LMEH

- Test environment
  - 22 client nodes, 8 meta data servers
  - Nodes: Dawning Tiankuo R220XP server
    - ✓2 2.4GHz Intel Xeon Processors
    - ✓ 2GB memory
    - ✓ Redhat 7.2, Linux-2.4.18-3smp
  - Network
    - ✓Gigabit Ethernet: 106.2MB/sec, 97.1μsec latency
  - Disks
    - ✓ SCSI disk: Seagate Ultra320 (ST373307LC)
    - ✓ 8MB data buffer, 2.99 msec average latency
    - ✓ 60MB/sec by iozone on EXT2
- Scenario
  - Empty file creation and deletion
  - Each client operates on different directory
  - Measurement unit: files / sec