Scaling Distributed File System Metadata Throughput

Kai Ren, Swapnil Patil, Kartik Kulkarni, Adit Madan, Garth Gibson (CMU)

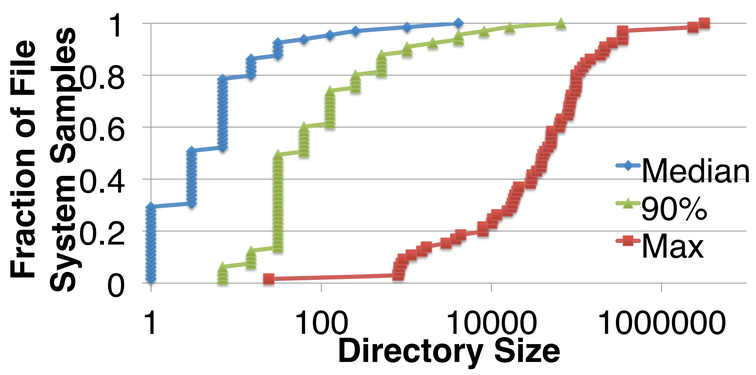
Overview

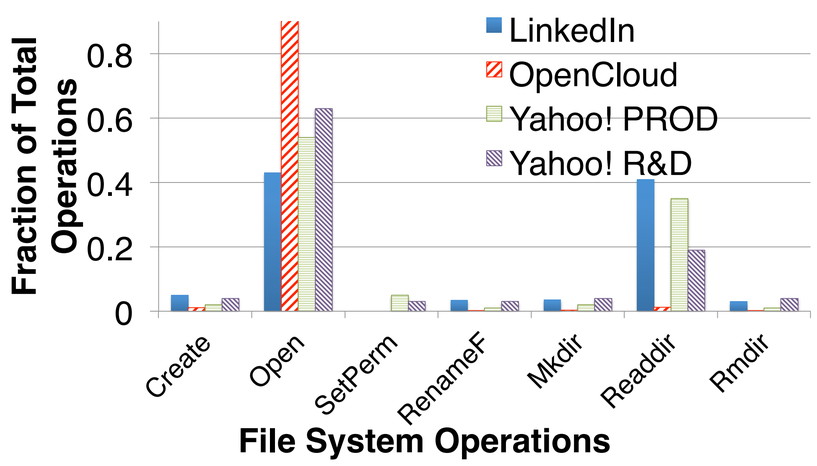
•**Problem**: The growing size of modern storage systems is expected to soon achieve and exceed billions of objects. Accesses to *metadata* (directory entries and file attributes) and *small files* are becoming a performance bottleneck.

• **Approach**: A middleware design provides scalable metadata path for existing distributed file systems
- •Partition the namespace at per-directory basis
- •Incrementally partition large directories
- •Represent metadata in log-structured merge tree

Design and Implementation

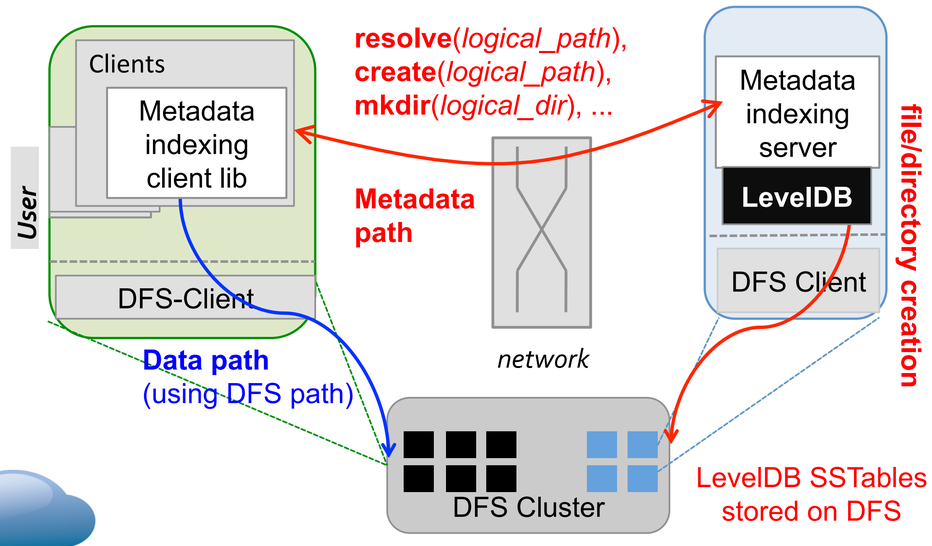
•**Workload Analysis**:
- •The depths of 90% directories are between 6 to 12
- •About 90% directories have fewer than 1000 entries [MSST13]
- •Most popular operations are *open*(), *stat*(), and *readdir*().





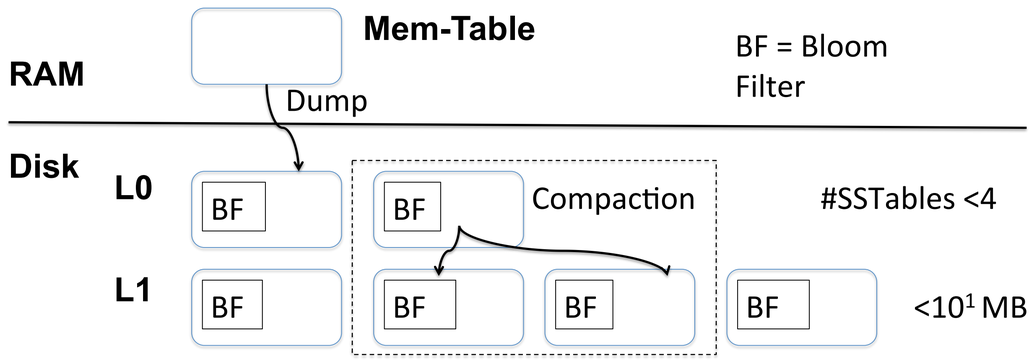
•**Namespace Distribution**:
- •Newly created directory is randomly assigned to a server
- •Binary splitting a directory partition using GIGA+ [FAST11]
- •Packing attributes and small files into LevelDB:

[Parent dir. Inumber, Object name] → [Attributes, File data| Link]



Log-structured Merge Tree

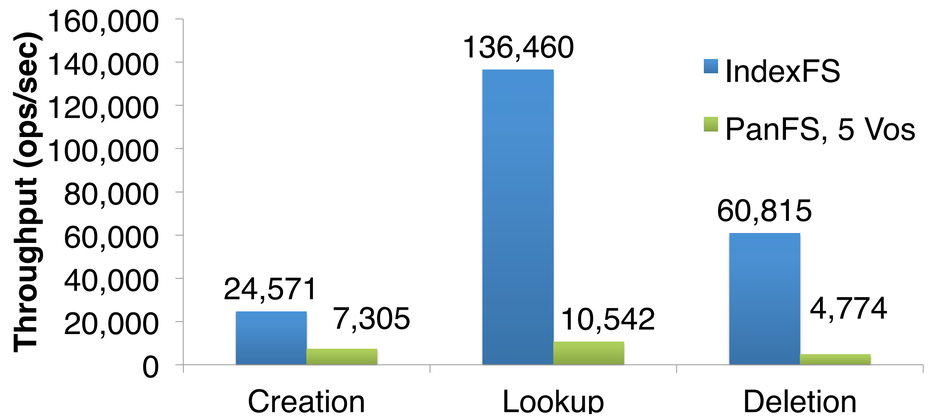
- •NoSQL databases adopt write optimized data structures like LSM (Log Structured Merge) Tree for high ingestion rate
- •Insertion: buffered in memory, later dumped to disk
- •Compaction: merging SSTables in L_i to L_{i+1}
- •Use Bloom-filters to reduce negative lookups



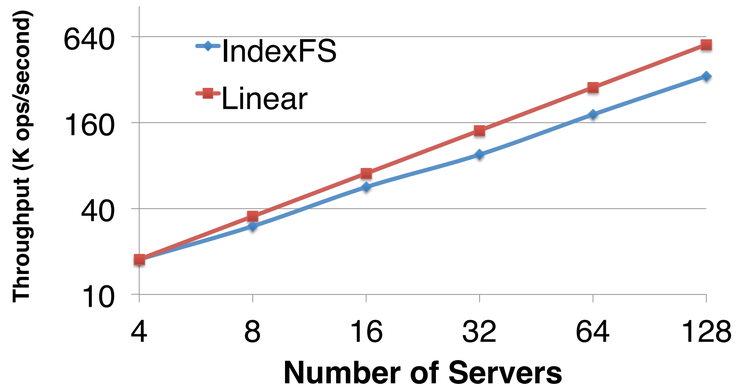
Evaluation of IndexFS

•**IndexFS Experiment Platform**:
- •Parallel file system: Panasas [FAST08] AS12 file system, 5 shelves with 5 metadata servers, 20Gbps each, 50 data servers
- •5 test machines with AMD Opteron 6272 64-core, 128GB memory, 40GE NIC, Mellanox 40GE switch

•**Mdtest on PanFS:** Three-phases HPC benchmark
- •Create / Stat / Delete 5 million files in a single directory



•**Scalability Test on HDFS:** Replay Linked-In one-day HDFS traces by partitioning the traces in round-robin fashion.
- •Scale the number of server/client machines from 4 to 128
- •Each machine has dual core, 8GB memory and 1GE NIC



Conclusion

- •Sustaining high metadata throughput for many machines
- •Delay file creates until file is non-trivial in size
- •Portable to a variety of file systems such as HDFS and PanFS

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley · University of Washington