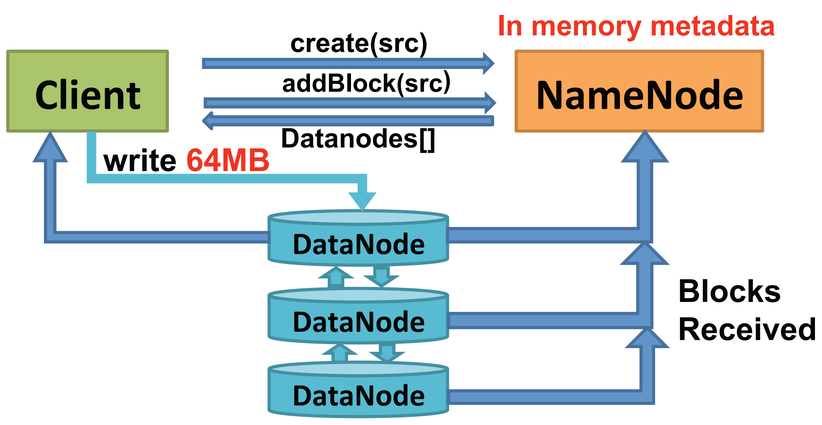
Scaling the Metadata Service in DISC storage

Lin Xiao, Wittawat Tantisiriroj, Garth Gibson

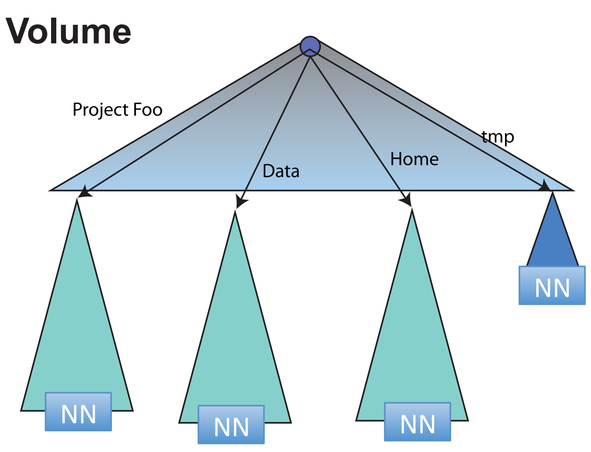
Metadata Server in Hadoop/Google FS

- Single metadata server is responsible for:
  - Maintaining namespace hierarchy
  - Managing data blocks and datanodes mapping
- For each 64MB data block, assign datanodes to hold the data blocks
- Clients write data directly to the datanodes to reduce metadata server load



What are GFS / HDFS Doing?

- Colossus doing rewrite of GFS, said to use BigTable for GFS metadata
- Yahoo! Federation
  - Static partition, like Volume
  - Hot Spot / Hard to load balance



Challenges

- Namespace Partition
  - Full path name: one table lookup for each entry
  - Directory depth + path name (Dep+Path)
  - Parent Inode + file name (Inode + FN)
  - Hash(path name): Better load balance
- Rename transactions
- Distributing block allocation
- Quota
  - Single node
  - Decentralized with row locking
  - Partitioned
- Repair process complexity

Single Metadata Server Limitation

- Memory Space Limitation
  - Keep all namespace and block information in memory when running
  - Limit number of files/folders and blocks
  - Small files are not rare in real system
- Performance
  - Write ahead log limits throughput of file creation
    - Maybe extend limit with SSD still does not scale
  - Only takes a handful of clients to saturate HDFS NameNode @ 5600 file creation per second [Shvachko 2011]

Should We Put a Scalable DB in HDFS?

- Stuff the 900lb gorilla into the VW Beetle?
- Scalable DBs are seen to be less robust than HDFS?
- Google may publish all they learn

But

- IcyTable is more robust
- Hbase keeps getting better
- Need to try if only to enumerate problems
- Google may not publish all they learn

So scale NameNode with scalable database

- Store files, directories and block mapping in scalable database
- Richer queries for admin and user

Preliminary Result

- Modified HDFS java plugin to store namespace in Hbase
- Two quad-core 2.83GHz Xeon, 16GB Memory, four 7200 rpm SATA 1TB disks, 10Gigabit Ethernet
- Hbase : 4 RegionServers
- Namenode: one disk for write-ahead-log(for comparison)

Schema	File creations/sec
NN with WriteAheadLog on disk	~110
NN with WriteAheadLog in tmpfs	~4361
Dep + Path	~4350
Inode + FN	~3800

Carnegie Mellon University

Parallel Data Laboratory