HDFS2, 一种分布式 NN 实现

2011-12-02 孙桂林

Why?

- Scalability
  - 10,000 Nodes
  - 1,000,000,000 Files
- Availability
  - No Single Point
  - Failover

Baidu 百度

The limits to growth

- Memory of 1,000,000,000 files
  - 380GB (1.1 blocks per file)
  - 860GB (3 blocks per file)
  - 1300GB (5 blocks per file)
- Performance
  - Latency and/or Throughput

HDFS Federation?

- Scale of cluster VS Scale of namespace
- Do we need a larger namespace?
  - Scale of computing & Locatily
  - Easy to use (directory moving)
  - Operation cost (data migration)

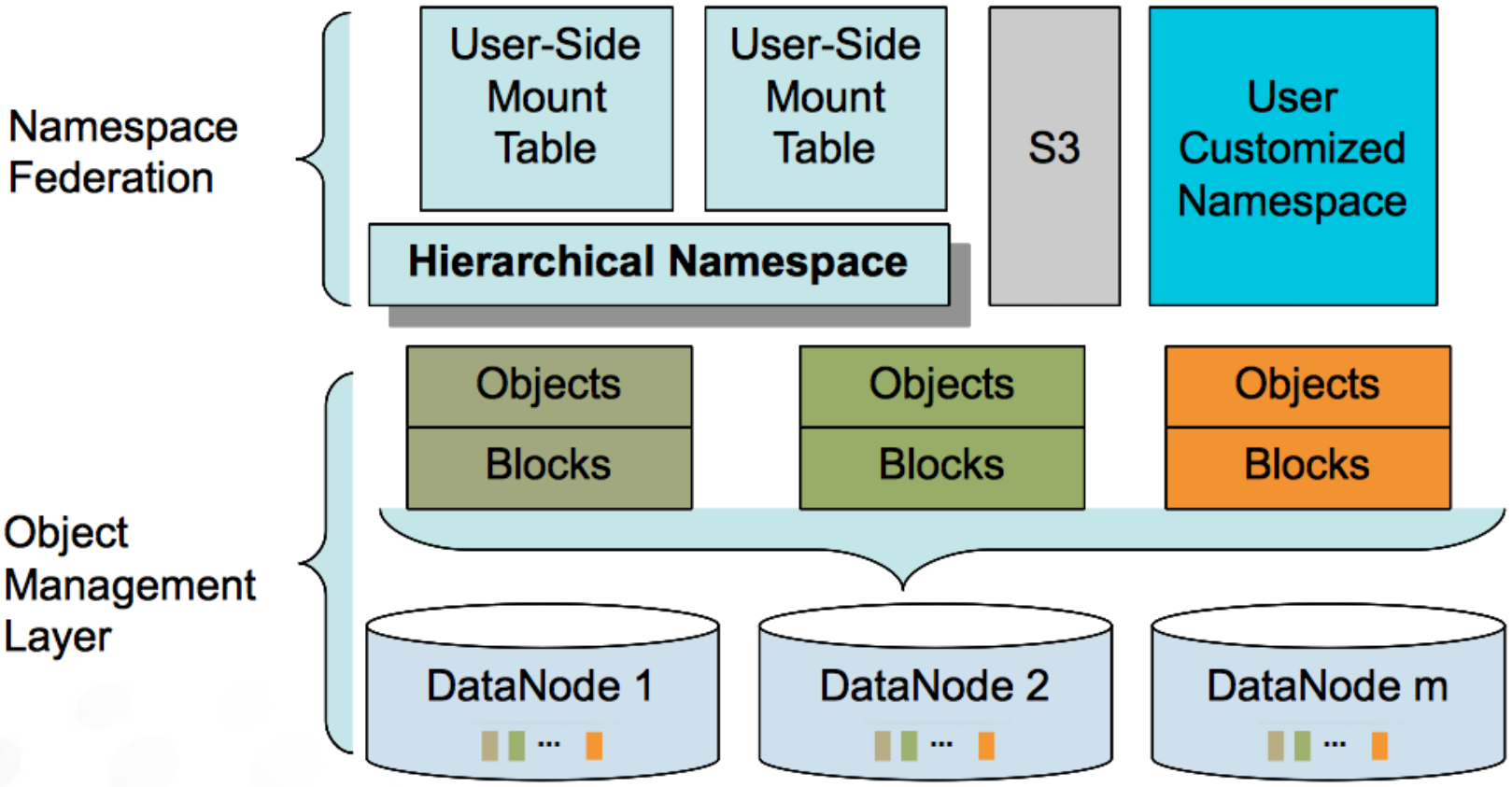
Baidu 百度

Solution

- Shared Object Management Layer
- Light-weight namespace
  - Without blocks management
  - Without file attributes
  - Move infrequently accessed data to disk (optional)
- Namespace Federation (optional)
- Shared storage (optional)

Baidu 百度

Architecture



Grow Up

- Online VS Offline
- Data VS Metadata

Baidu 百度

Memory

- Namespace
  - 1,000,000,000 files 66GB
  - 2,000,000,000 files 132GB
  - …
- Object Management Layer
  - 100,000,000 files per node (60~80GB).

Performance

- Latency
  - Batch operation (Group by Node)
- Throughput
  - Meta operation
  - Internal/External block operation
  - Overall throughput increased by 500% ~ 1000%

Bai du 百度

Availability

- Single point failure @ HDFS2
- HDFS2 HA VS AvatarNode HA

And...

- Fast copy & snapshot
- Trash with intelligence
- Startup parallelization
- Computing locality
- Locking optimization

Baidu 百度

Q & A

Baidu 百度