Multiprotocol Locking and Lock Failover in OneFS

Aravind Srinivasan

EMC, Isilon Storage Division

Agenda

- Overview
- OneFS Overview
- Overview of the DLM in OneFS
- Multiprotocol Locking in OneFS
- Lock Failover in OneFS

Overview

- Any clustered file system (like Isilon's OneFS) needs a robust Distributed Lock Manager (DLM) to synchronize resources accessed from different protocol clients (such as SMB and NFS).
- Also we need a failover mechanism to implement the failover semantics of these protocols so that the locks are not lost even when a node in the cluster goes down.

OneFS Overview

EMC - Isilon OneFS Cluster

- NAS file server
- Scalable
  - Add more storage in 5 mins
- Reliable
  - 8x mirror / +4 parity
  - Striped across nodes
- Single volume file system
- 3 to 144 nodes
- Fully symmetric peers
  - No metadata servers
- Commodity hardware
  - CPU, Mem, Disks

EMC - Isilon OneFS File System

- Concurrent access to all files with all protocols
  - SMB1/SMB2
  - NFSv3/NFSv4
  - SSH
  - HTTP/FTP

OneFS – High Level Overview

- OneFS is EMC-Isilon's seventh-generation operating system that provides the intelligence behind all EMC-Isilon scale-out storage systems.
- It combines the three layers of traditional storage architectures—file system, volume manager and RAID—into one unified software layer, creating a single intelligent file system that spans all nodes within a cluster.

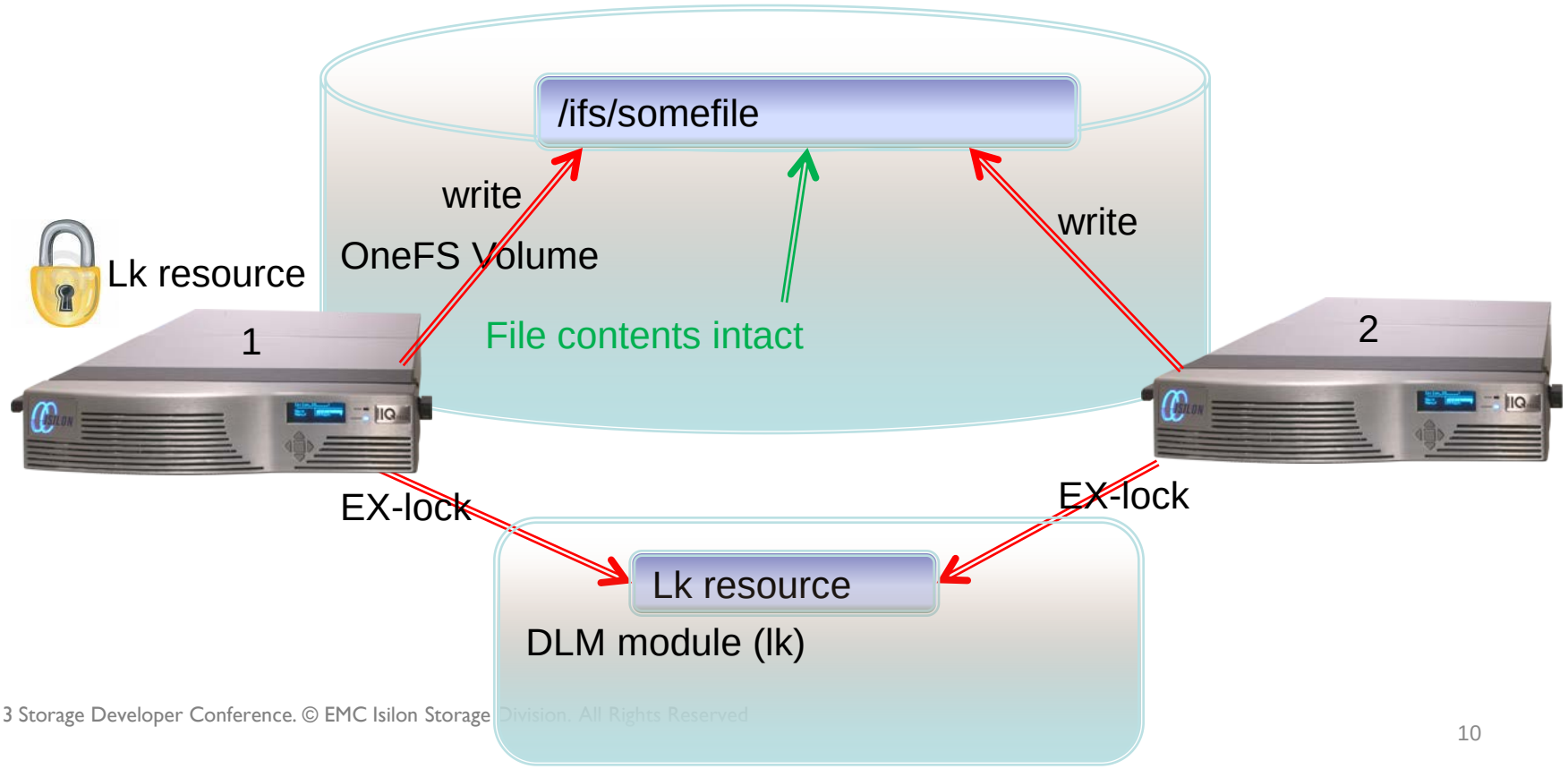
OneFS – High Level Overview

- Isilon's OneFS enables:
  - Independent or linear scalability of performance and capacity
  - A single point of management for large and rapidly growing repositories of data
  - Mission-critical reliability and high availability with state-of-the-art data protection

OneFS DLM Overview

DLM In OneFS

- Goal of DLM



DLM in OneFS - Overview

- The DLM in OneFS is called LK and is split into two distinct roles:
  - Initiator and
  - Coordinator

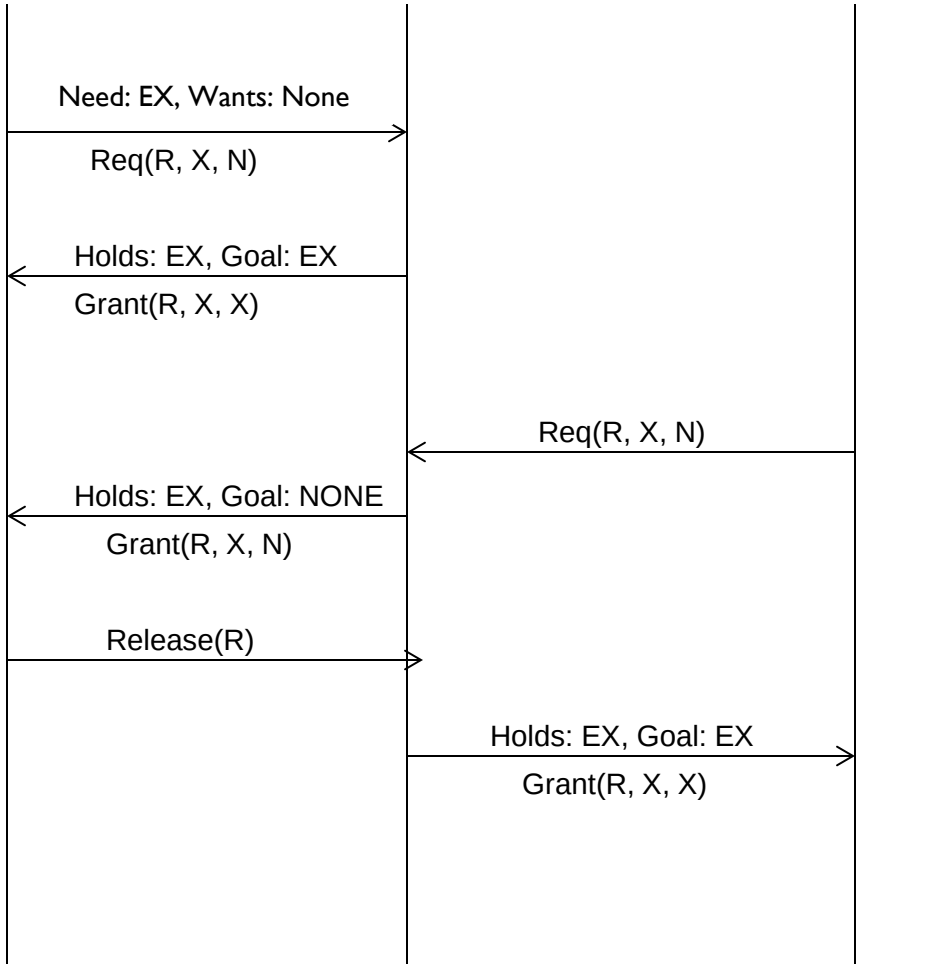
LK - Coordinator

- Locks are coordinated in lk.
- Each resource is coordinated by a particular node. The lk coordinator node arbitrates locking within the cluster for a particular subset of resources.
- The coordinator is chosen by a numeric transformation of the resource ID, in the simplest case, the ID modulo the number of nodes in the cluster.

LK - Initiator

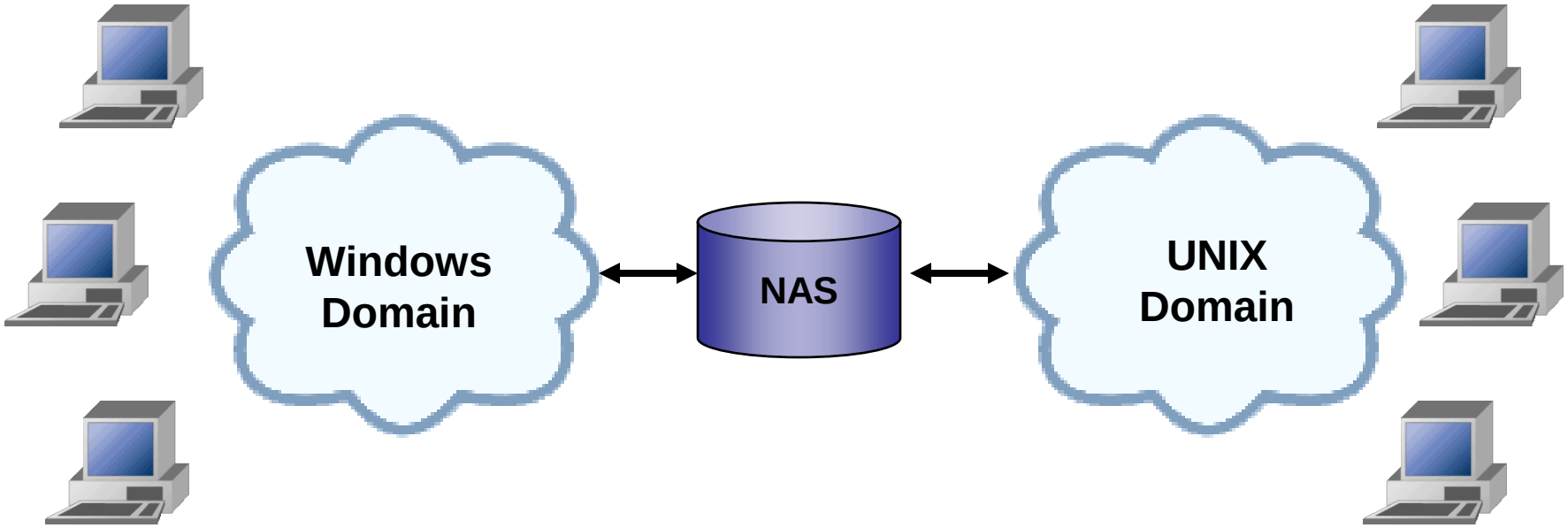
- The initiator is the one requesting the lock.
- On the initiator side, there is one entry for each resource for which there is a local owner or waiter.

LK – Coordinated Two-Tier Locking



Multiprotocol File Sharing Environments

- The same file data is accessed by both UNIX/Linux and Windows users concurrently



Multiprotocol Locking in OneFS

- Locks in LK are coordinated per domain
  - There can be multiple lock domain in existence at any time, each one controlling locks for a different aspect of the system. Eg: OPLOCK domain/CBRL domain
- Locks within a domain contend with each other
- This concept of domain enables OneFS to implement multiprotocol locking support

Multiprotocol Locking in OneFS

- We can define any LK domain and share it between protocols to make them contend with each other.
- OneFS tries to coordinate the share modes from NFSv4 and SMB clients by having a domain shared by both the protocols.

OneFS Lock Failover Overview

Lock Failover in OneFS

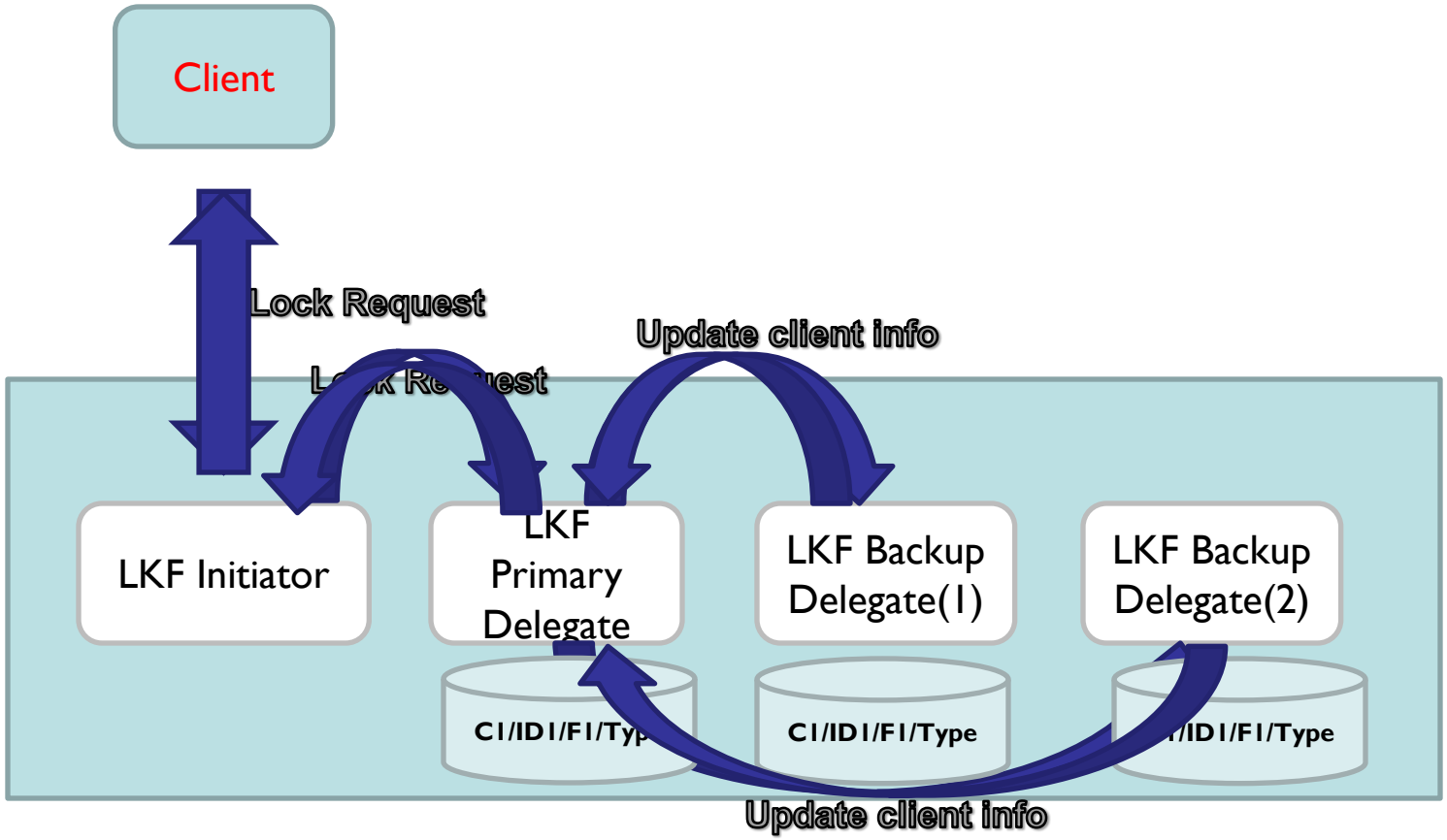
- There are protocols like NFS which require locks to stay even when a node in the cluster goes down.
- LK on its own, is a pure DLM without any failover semantics. So if a node goes down all it's LK locks will be lost.
- In order to implement lock failover, OneFS has a component called LKF, which is a consumer of LK with failover support.

LKF in OneFS

- LKF Terms
  - LKF Initiator – The node to which the client is connected.
  - LKF Primary Delegate – The node which talks to LK to get the locks on behalf of the client. This is chosen by hashing the client name with the number of nodes in the cluster.
  - LKF Backup Delegate(s) – The node(s) which stay in sync with the Primary Delegate

Note: These are different from the LK Coordinator and initiator

OneFS LKF - Overview



LKF in OneFS

- Failover Scenario
  - Node with the lock (Primary Delegate) goes down.
  - As part of the group change:
    - The API is first suspended to prevent any new requests from coming in.
    - Backup Delegates take over as primary and get the locks for the client.
    - Once this is done, the API is resumed.

LKF in OneFS - Contd

- The Primary and the Backup Delegates must always stay in sync.
- When a node goes down, one of the backup delegates will take over as the primary and get the locks held by the client.
- Currently used only by NFS clients in OneFS.

LKF Challenges

- LKF can be extended to support other protocols like SMB3.
- The main challenge is to confirm to the protocol specific requirements and tweak the system accordingly.
- This can be extended to failover other information in addition to locks as well by having a blob of data to be failed over rather than just the lock.

LKF Challenges

- LKF can be extended to support other protocols like SMB3.
- The main challenge is to confirm to the protocol specific requirements and tweak the system accordingly.
- This can be extended to failover other information in addition to locks as well by having a blob of data to be failed over rather than just the lock.

Summary

- Distributed locking in OneFS is achieved by using a OneFS specific DLM called LK
- The domain concept in LK allows the potential to enable multiprotocol locking support in OneFS
- Lock Failover in OneFS is achieved using a system called LKF which is a consumer of LK.

Questions?

Contact

Aravind Srinivasan

asrinivasan@isilon.com