SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2015

SMB 3.0 Transparent Failover for EMC Isilon OneFS

John Gemignani

EMC – Emerging Technologies Division

Isilon

Clusters may be capable of offering continuous availability to files by moving workloads from one node to another.

Some protocols can do this seamlessly

- HTTP, HDFS, NFS3

- Some protocols can do this with proper support
  - NLM, NFS4, SMB3
- Others simply cannot
  - SSH, FTP

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Agenda

- OneFS Overview
- SMB CA and Witness
  - What SMB CA Is Intended to Do
  - What SMB Witness Can Do To Help
  - Intended Workflows for CA
- Implementation in OneFS
- Experiences

2015 Storage Developer Conference. © EMC. All Rights Reserved.

OneFS Overview

2015 Storage Developer Conference. © EMC. All Rights Reserved.

OneFS Overview

SDC 15

2015 Storage Developer Conference. © EMC. All Rights Reserved.

OneFS Features

- Scalable performance and capacity
- Data integrity and protection
- High availability
- All nodes are fully-functional, symmetric peers
- Client-facing protocols entirely in user-mode
- Protocols supported by a common, high-performance infrastructure

2015 Storage Developer Conference. © EMC. All Rights Reserved.

OneFS Features (2)

- Concurrent access to all files from all protocols:
  - SMB1/SMB2/SMB3
  - NFSv3/NFSv4/NLM/NSM
  - HDFS
  - SSH
  - HTTP
  - FTP
- Protocols supported within “zones” and “pools”

2015 Storage Developer Conference. © EMC. All Rights Reserved.

SMB CA and Witness

2015 Storage Developer Conference. © EMC. All Rights Reserved.

What SMB CA Is Intended To Do

- Address applications that aren't resilient to issues relating to connectivity:
  - I/O errors
  - Unexpected closure of file handles
  - Long access outages
- Resolve ugly complications arising from outages when clients cache data under a lease
- Do so in an automated and transparent manner

2015 Storage Developer Conference. © EMC. All Rights Reserved.

How SMB CA Accomplishes This

- Support file open requests for persistent handles
- Persistent handles backed by persistent data
- Persistent handles are available for reclaim from any server within the cluster, for a bounded time
- For protection and continuity, while disconnected, the file cannot be opened by anyone else (subject to bounded time)

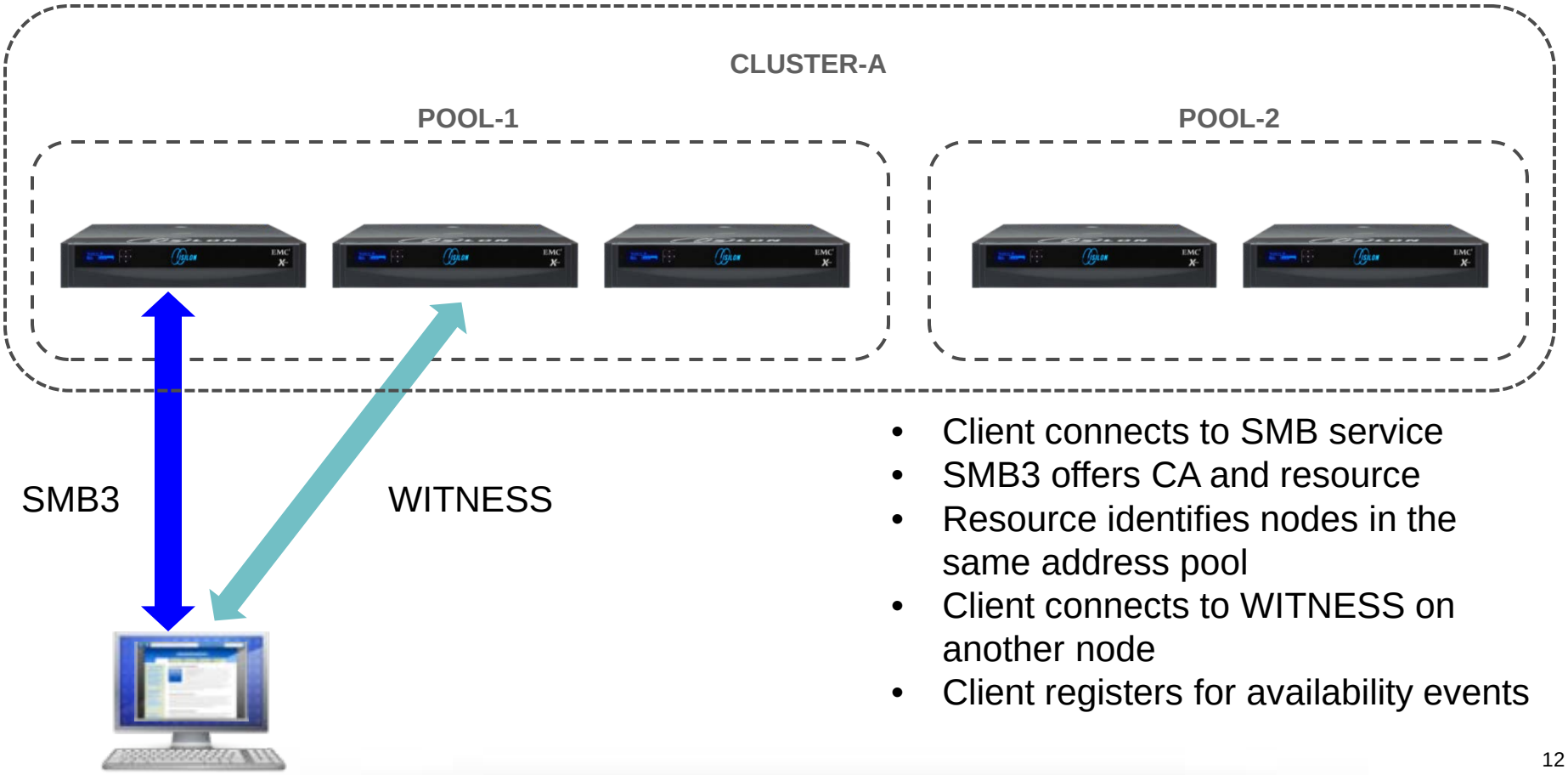
2015 Storage Developer Conference. © EMC. All Rights Reserved.

What SMB Witness Can Do To Help

- Identify paths to a resource
- Provide feedback to clients about availability
- Expedite the transfer of the workflow
  - No TCP keep-alive dependencies
  - No SMB timeouts needed
- Outages minimized, even nearly indiscernible
- Supported by any node in the pool

2015 Storage Developer Conference. © EMC. All Rights Reserved.

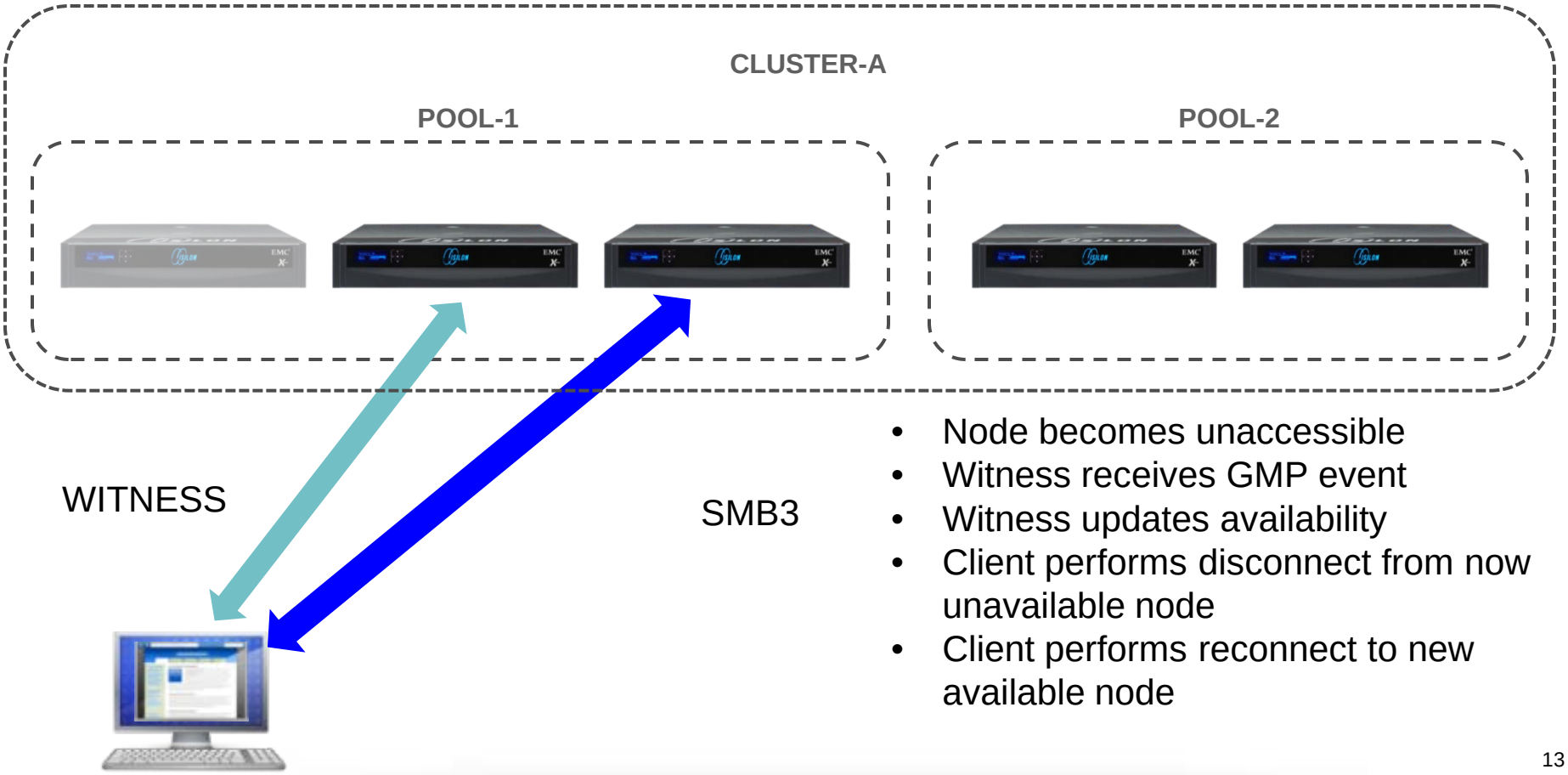
SMB CA and WITNESS



- Client connects to SMB service
- SMB3 offers CA and resource
- Resource identifies nodes in the same address pool
- Client connects to WITNESS on another node
- Client registers for availability events

2015 Storage Developer Conference. © EMC. All Rights Reserved.

SMB CA and WITNESS (2)



- Node becomes unaccessible
- Witness receives GMP event
- Witness updates availability
- Client performs disconnect from now unavailable node
- Client performs reconnect to new available node

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Intended Workflows for CA

- Node maintenance – planned
  - Hardware servicing
  - Software updates
    - Simple: updates without node reboot
    - Complex: updates with node reboot
- Cluster reconfiguration – planned

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Intended Workflows for CA (2)

- Node failure – unplanned outage
  - SMB service outage
  - Transient cluster-related issues
  - Node downtime
- Non-disruptive home directories

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Intended Workflows for CA (3)

- Workload migration – future opportunity
  - Ability to move workload across nodes
  - Potential for load balancing
  - Potential recovery from various pool-related infrastructure problems

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Implementation in OneFS

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Implementation In OneFS

- The Parts
  - Administration
  - Supporting cluster infrastructure
  - CA in the SMB service
  - The Witness protocol

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Administration

- This is, by far, the easy part
- CA is a share option
- Web UI
- Commands

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Supporting Cluster Infrastructure

- Hands-down the most difficult and sensitive part
- Lock subsystem was chosen as it provides:
  - Cluster-coherent management of resources
  - Ownership (registrations)
  - Manages contention, distribution and recovery
  - State survives total loss of the server node

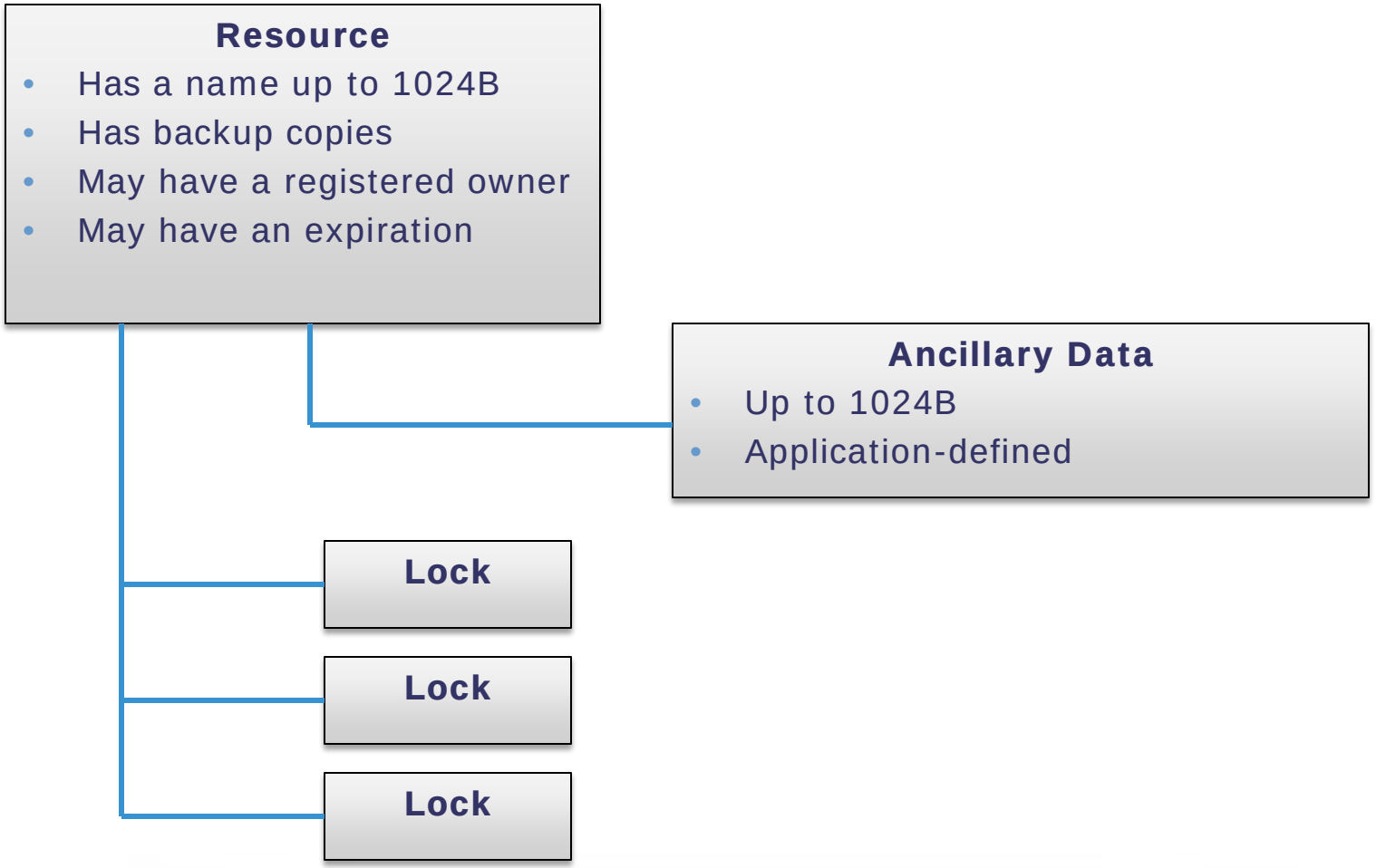
2015 Storage Developer Conference. © EMC. All Rights Reserved.

Supporting Cluster Infrastructure (2)

- Now supports persistence of ancillary file data
- Persistent handle gets us to persistent data
- Persistent data can be up to 1024 bytes and is application-defined
- State may have an associated expiration
- Leases are also managed this way

2015 Storage Developer Conference. © EMC. All Rights Reserved.

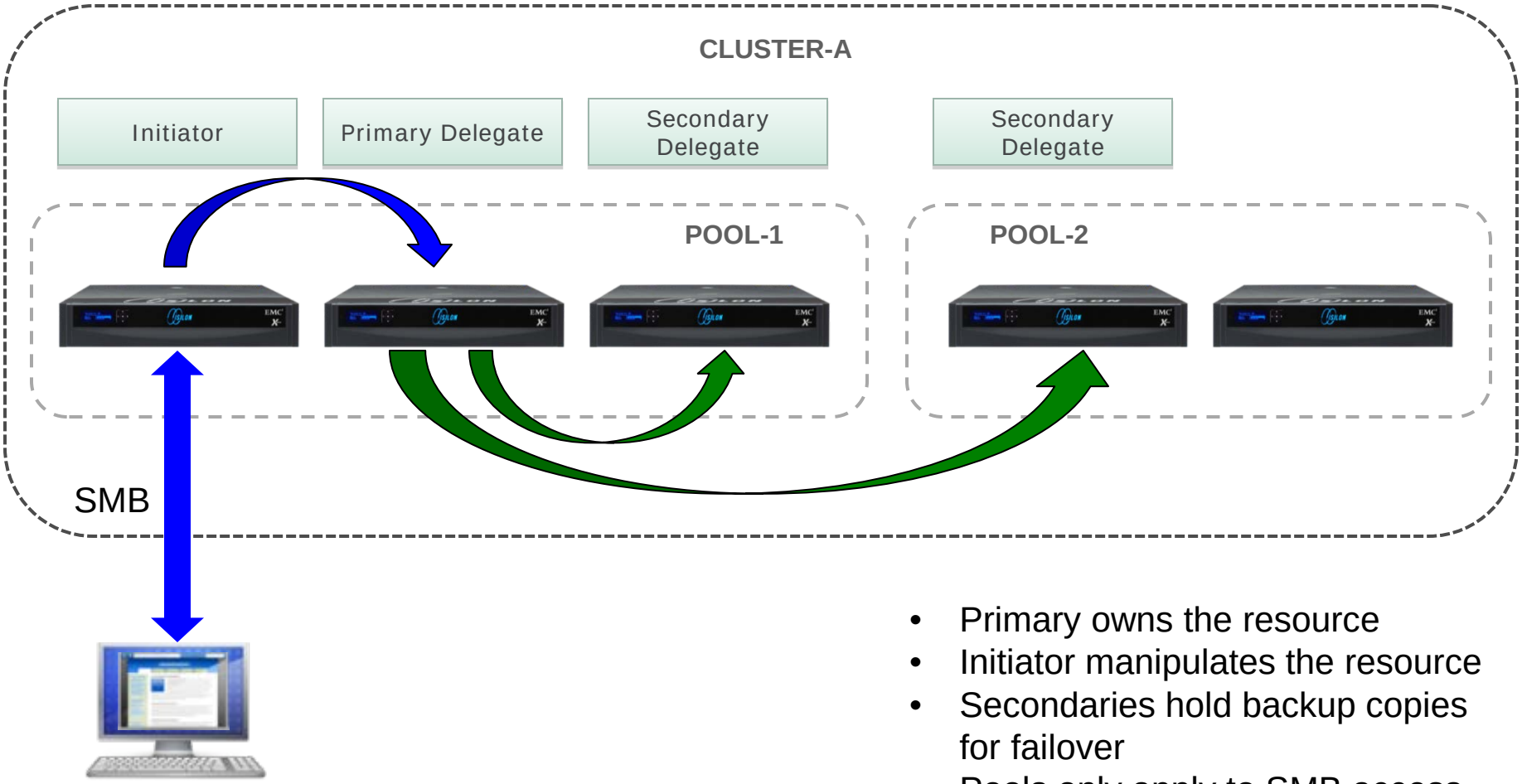
Supporting Cluster Infrastructure (3)



SDC 15

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Supporting Cluster Infrastructure (4)



- Primary owns the resource
- Initiator manipulates the resource
- Secondaries hold backup copies for failover
- Pools only apply to SMB access

SDC 15

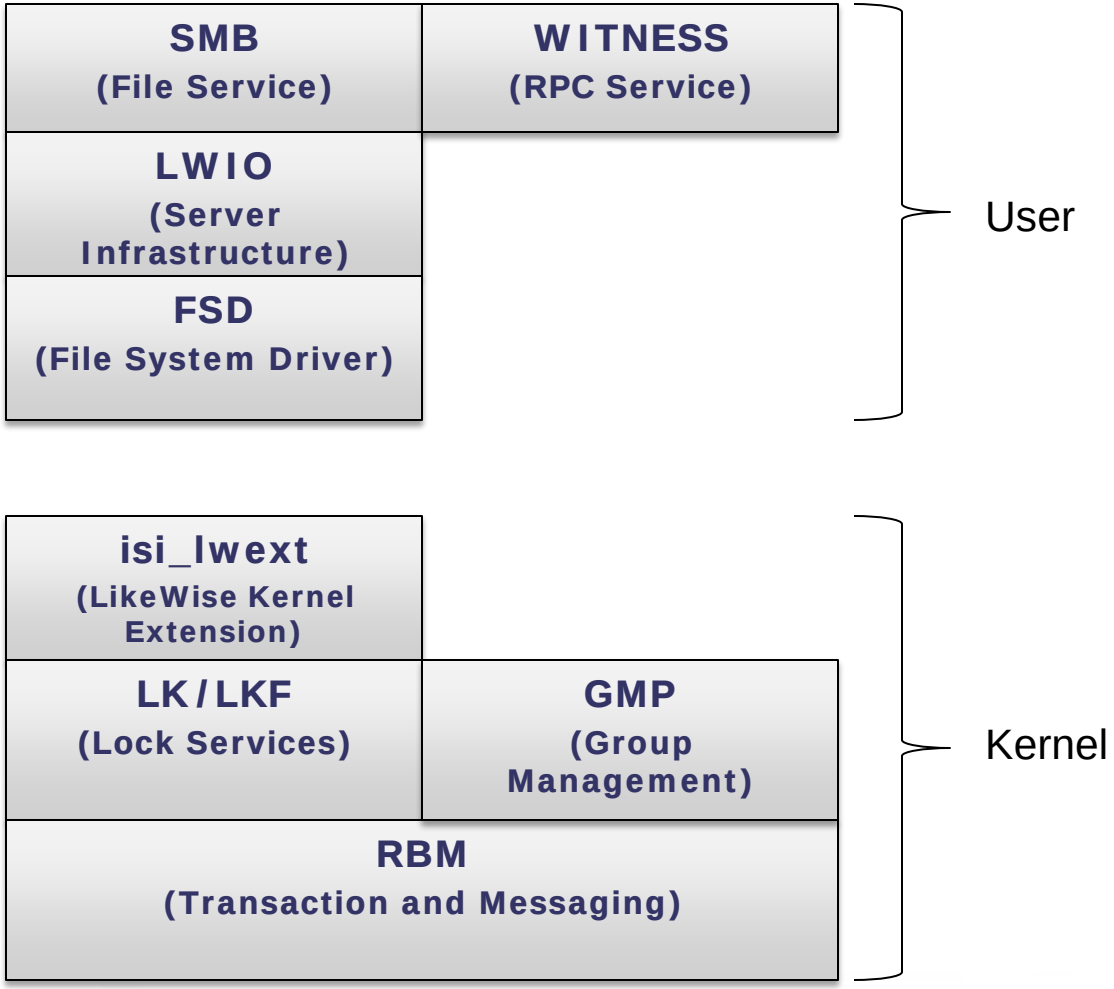
2015 Storage Developer Conference. © EMC. All Rights Reserved.

CA in the SMB Service

- Moderately difficult
  - The right tinker toys need to be in place
- Built upon several layers of both improvements and enhancements
- Support client requests for persistent handles
- Required a cluster-wide persistent handle
  - Must be globally accessible
  - Must be unique

2015 Storage Developer Conference. © EMC. All Rights Reserved.

CA in the SMB Service (2)



SDC 15

2015 Storage Developer Conference. © EMC. All Rights Reserved.

CA in the SMB Service (3)

- SMB – File services
- WITNESS – RPC service for availability
- LWIO – High performance server infrastructure
- FSD – OneFS user-mode personality driver
- LWEXT – OneFS kernel-mode personality system service loadable module
- LKF – OneFS persistent lock/state subsystem
- GMP – OneFS Cluster group management
- RBM – OneFS transaction and message subsystem

2015 Storage Developer Conference. © EMC. All Rights Reserved.

The Witness RPC

- Not too difficult
- Two types of responses to notification requests
  - Status update (available, unavailable)
  - Please move (to IP address)
- OneFS supports the Witness V1 interface
- Only events related to status updates sent
  - OneFS already has cluster event facility

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Experience

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Experience

- Witness and client reaction is reasonably fast
- Simple tree-connect restored in 1-2 seconds
- Other times are related to the number of file reconnect/reclaim operations sent from the client
- Original design treated all reconnects the same
  - Same node case caches state for returns
  - Other node case relies on stored state

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Experience (2)

- Our SMB3 session IDs are not cluster-wide
  - Reconnects “steal” the original state
  - Previous node is notified to invalidate its copy
- With home directories lockout may be a problem
  - Administrator may allow conflicting opens to break through the lockout

2015 Storage Developer Conference. © EMC. All Rights Reserved.

Questions?

Contact Information

john.gemignani@emc.com

2015 Storage Developer Conference. © EMC. All Rights Reserved.