NFS INDUSTRY CONFERENCE

DAFS Extensions for NFS

Mark Wittle

Technical Director

Network Appliance

mwittle@netapp.com

NFS INDUSTRY CONFERENCE

Presentation Outline

- DAFS Overview
- Direct Access Networks
- DAFS Performance Operations
- DAFS Local File Sharing Enhancements
- Applicability for NFS

NFS INDUSTRY CONFERENCE

DAFS Overview

NFS INDUSTRY CONFERENCE

DAFS Derived from NFSv4

- Many things are the same
  - Client-server, request-response messages
  - Basic file & directory operation set
  - File Attributes, ACLs
  - Locks, leases, delegations
  - Authentication types, internationalisation

NFS INDUSTRY CONFERENCE

DAFS - NFS Differences

- Some things are different
  - Session-based (with authentication & attributes)
  - Credit-based flow-control (request throttling)
  - Transport channel aware
  - Different wire protocol format (XDR, Endian)
- Two new areas of focus
  - High performance
  - Local file sharing semantics

NFS INDUSTRY CONFERENCE

DAFS Performance Focus

- RDMA Support
  - Direct data placement
  - Pre-allocate transfer buffers
  - Split protocol header from application data
  - User-space I/O support
- Sophisticated I/O Controls
  - Batched and chained I/O operations
  - File access pattern hints for cache efficiency

NFS INDUSTRY CONFERENCE

DAFS Local File Sharing

- Application Cluster Environment
  - Access control for cooperating processes
  - Atomic file append operations
- Failure Recovery
  - Extended locking semantics
  - Fencing to control partial cluster failures
  - Server response cache management

NFS INDUSTRY CONFERENCE

Direct Access Networks

NFS INDUSTRY CONFERENCE

Direct Access Networks

- Technology protocols
  - InfiniBand, VIA, iWARP
- Capabilities
  - Multiple connections multiplexed in HW
  - Memory pre-registration
  - Send/receive messages
  - Remote DMA read/write
  - Asynchronous I/O completion

NFS INDUSTRY CONFERENCE

Direct Access Benefits

- No data packet fragmentation or reassembly
  - Benefits similar to IP Jumbo Frames, but with larger packets
  - Less transmission overhead with fewer HW interrupts
  - No ordering & space management issues
  - No data copying to recreate contiguous buffers

Direct Access Benefits (2)

- No realignment of data copies
  - Protocol headers and data buffers transmitted separately
  - Allows data alignment to be preserved
- No user/kernel boundary crossing
  - Less system call overhead
- No user/kernel data copies
  - Data transferred directly to application buffers

NFS INDUSTRY CONFERENCE

DAFS Performance Operations

NFS INDUSTRY CONFERENCE

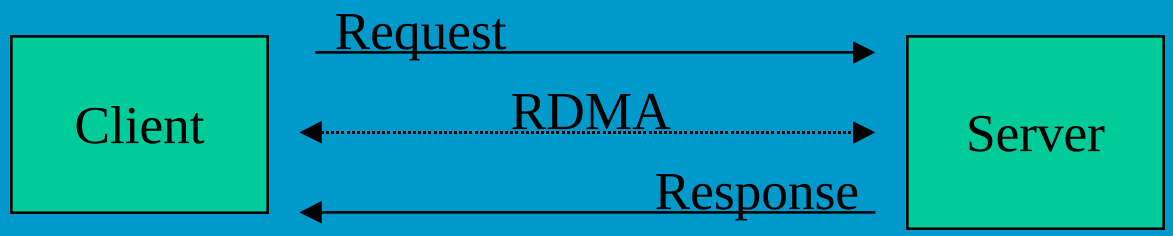
DAFS Inline I/O Operations

- Inline Read and Write
  - Just like NFS

  - Pre-allocated buffers, pre-registered with the transport
  - Configurable message size limit
  - Configurable data offset
  - Low transport latency, simple model

DAFS Direct I/O Operations

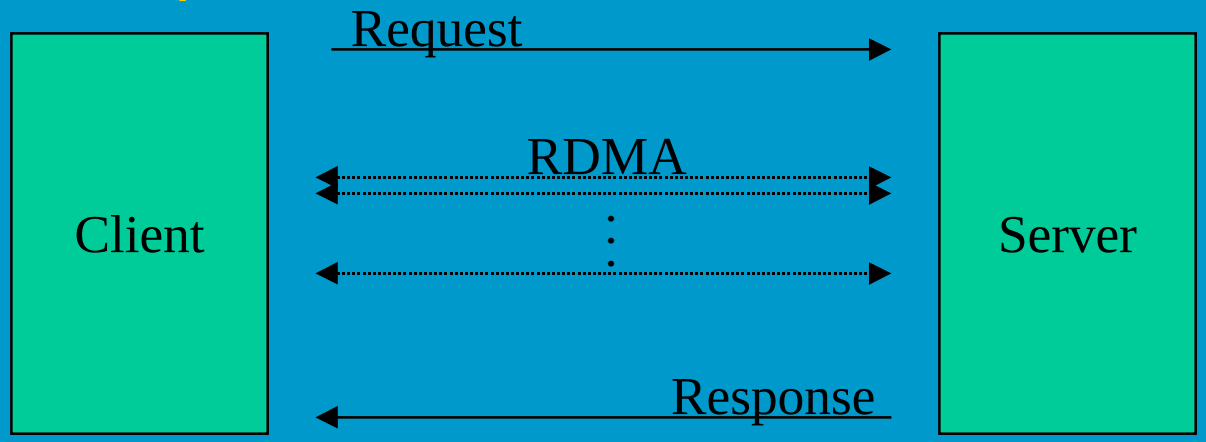
- Direct Read and Write
  - 3-part transfer



  - Server initiates RDMA operation
  - Receive buffer sized & placed at runtime
  - Used for large messages
  - Zero-copy, low CPU cost

DAFS Batch I/O Operations

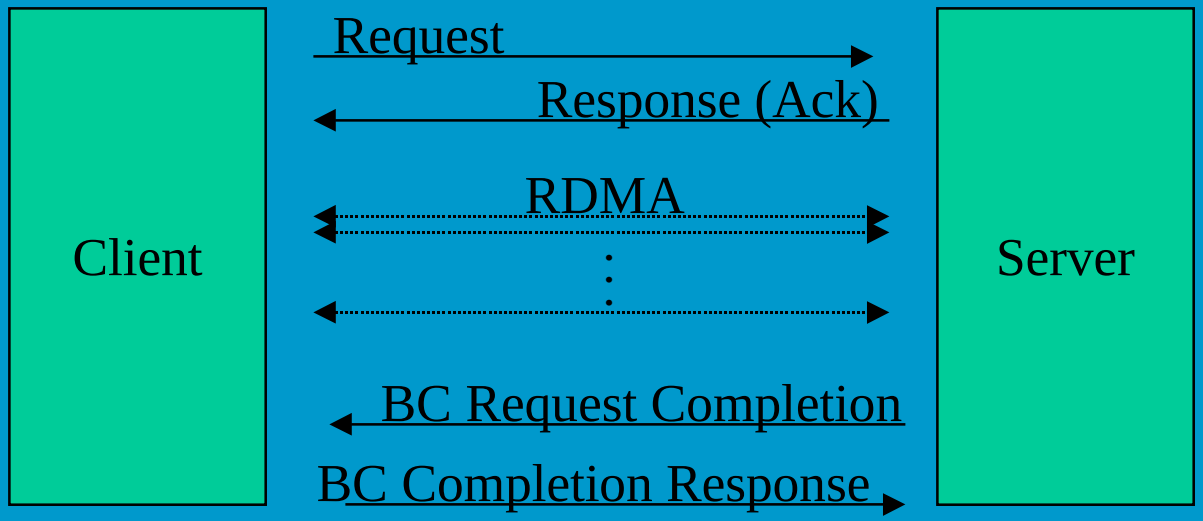
- Synchronous Batch I/O
  - Multi-part transfer



  - Server initiates RDMA operation
  - Supports multiple files, read/write, scatter-gather, sync/async, stable delay hint

DAFS Batch I/O Ops (2)

- Asynchronous Batch I/O
  - Multi-part transfer



  - Server initiates RDMA operation

NFS INDUSTRY CONFERENCE

DAFS Cache Hints

- Similar to POSIX madvise
  - Access prediction to inform caching policy
    - Random, Sequential, Unknown, Will Need, Don't Need
  - Buffer-specific future use hint to inform specific caching decisions
    - Two separate likelihood hints for future Read & Write access
- Cache hint operation and hints included in I/O Operations

NFS INDUSTRY CONFERENCE

Multi-component Lookup

- Multi-component path name lookup
  - An entire pathname is looked up in one operation
  - If an error is encountered partway through the lookup, partial lookup information is returned along with an error status.

NFS INDUSTRY CONFERENCE

DAFS Local File Sharing

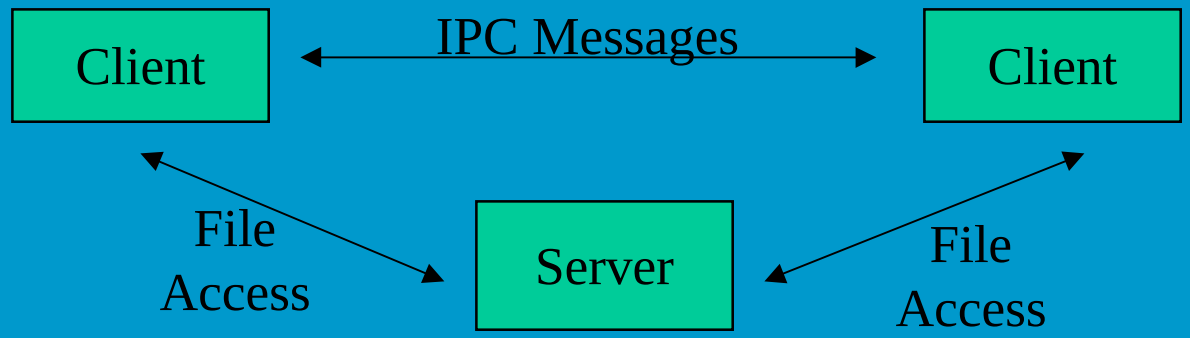
NFS INDUSTRY CONFERENCE

Shared Key Reservations

- Access control for cluster member instances
  - Addresses partial restarts
    - Same client, but different temporal instance
  - Reservation (lock) obtained at open with current "key“
  - Access control for cooperating clients
  - Key distribution left to application

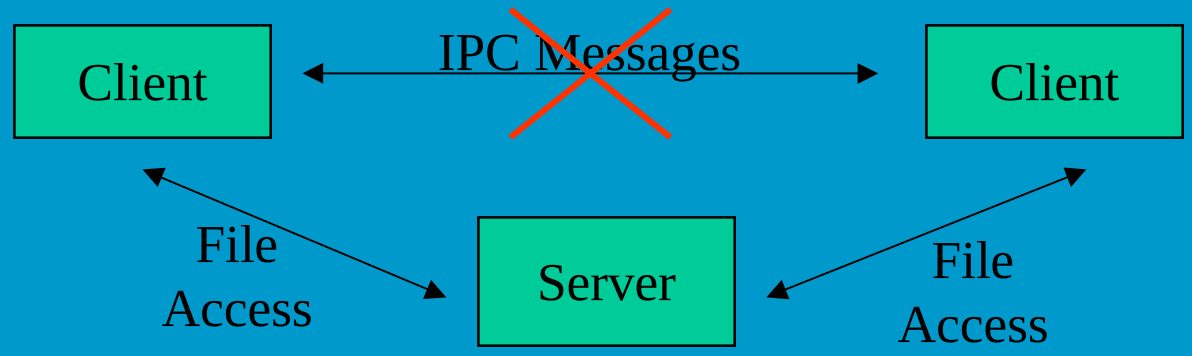
DAFS Fencing

- Access control for cluster failures
  - Stable state



DAFS Fencing (2)

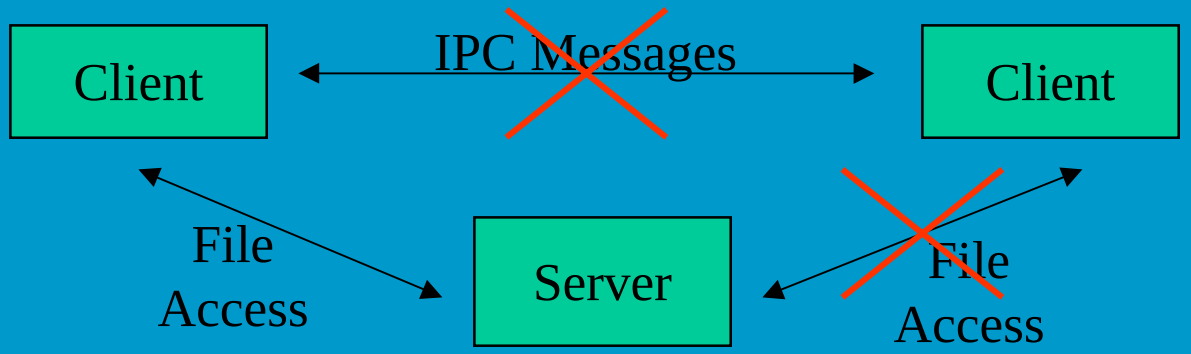
- Access control for cluster failures
  - Partial failure



NFS INDUSTRY CONFERENCE

DAFS Fencing (3)

- Access control for cluster failures



  - Immediate access denial
  - Drain operations on server to establish valid serialization point
- Manage a “fencing access list” of clients: allow, deny

NFS INDUSTRY CONFERENCE

DAFS Extended Locking

- Persist locks
  - take effect on lease expiration or on server reboot
  - remain locked until explicitly repaired
  - allow 3rd party recovery following failure
- Auto-Recovery locks
  - limited undo capability after failure
  - modifications performed under a rollback lock are undone following lease expiration or server reboot

DAFS Atomic Append

- Atomic relative to other writers
  - Closes the gap between GET_ATTR and WRITE
  - Server determines current file size and performs the write operation atomically
- Atomic relative to the data buffer
  - Server performs the I/O operation atomically, up to server file system attribute MAX_APPEND_SZ
- Atomic relative to server failures
  - Server performs synchronous I/O to stable storage

NFS INDUSTRY CONFERENCE

Other DAFS Operations

- Set file attributes during creation
  - Allow file attributes to be specified when a file is created
- Open Unlinked
  - Create a file without having a name in any directory. Subsequently the file can be named by linking it into the directory.
- Delete On Last Close
  - Addresses the .nfs file problem.

NFS INDUSTRY CONFERENCE

DAFS Response Cache

- Session-based non-volatile Response Cache
  - holds recent state modifying operations
  - number of entries bounded: transport message ordering properties and DAFS flow control constraints
  - Client queriable following a failure
- Provides at-most-once semantics for state modifying file operations

NFS INDUSTRY CONFERENCE

Applicability to NFS (Speculation)

NFS INDUSTRY CONFERENCE

Performance Operations

- Inline I/O operations
  - Header padding for data for alignment
- Batch I/O
  - Similar to NFS Compound
  - Stable Hint useful for high throughput applications
- Cache Hints
  - Conduit for application madvise calls
  - Buffer caching hints for 2nd-level cache?

NFS INDUSTRY CONFERENCE

Local File Sharing

- Clustered application support
  - Fencing, Shared key reservations
  - Are cluster applications becoming more prevalent?
  - No POSIX API, but good for user-space NFS
- Atomic append
  - Really a cluster application performance enhancement

NFS INDUSTRY CONFERENCE

Local File Sharing (2)

- Multi-component Lookup
  - Similar to NFS Compound?
- Require persistent server state
  - Open Unlinked (No POSIX API)
  - DOLC
  - Extended Locking (No POSIX API)
  - Persistent Response Cache

NFS INDUSTRY CONFERENCE

Conclusions

- DAFS – NFS structural differences
  - DAFS design (reliable transport assumption, session-orientation, flow-control) enables efficient buffer management, reliable response cache, etc
  - Not trivial to adapt to existing NFS
- DAFS – NFS file sharing features
  - Some features easy to introduce
  - May require more server state
  - POSIX API a limitation for some