SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

Simplified and Consolidated Parallel Media File System Solution

Presenters :

Dr. M. K. Jibbe, Technical Director NetApp ESG

Joey Parnell, Software Architect NetApp, ESG

Agenda

- Abstract
- Technology areas covered in this paper
- Current problem
- Typical and future high-bandwidth filesystem configuration
- Areas of the problem solution / enhancements
- Demo of containerized/integrated metadata controller for parallel filesystem
- Hardware component comparison with and without integrated metadata controller
- Key takeaways and closing thoughts

Abstract

- Simplified architecture for servicing parallel filesystem workload for M&E
  - This presentation describes a simplified architecture for high-bandwidth workloads on parallel filesystems such as StorNext, created by **integrating the filesystem metadata controller (MDC) within a Linux container** running directly on a NetApp E-Series E5700 storage array. This enables **consolidation of hardware** (and potentially vendors) and simplifies configuration complexity for use-cases such as media and entertainment content generation and distribution. One unique attribute of this solution is the **data path** from the filesystem clients directly attaches to the NetApp E5700 via **SCSI/FC** while access to the **MDC** from the filesystem client is via the E5700's **on-board Ethernet** link. Furthermore, access between the integrated MDC in the container and the storage is via an **NVMe coupling driver**. This greatly simplifies complexity of the architecture through component consolidation while using specialization to optimize use of the E5700's multiple I/O interfaces.

Technology areas

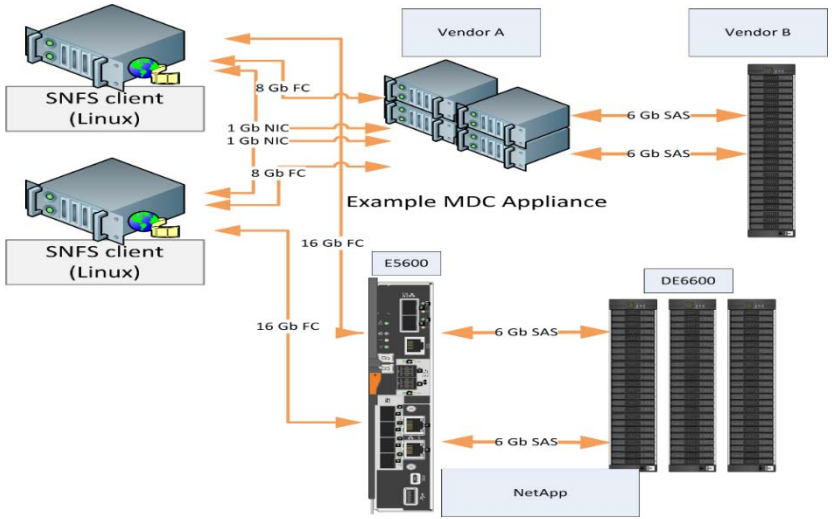
- ✓Containers
- ✓SCSI + Ethernet + NVMe Coupling Driver
- ✓Address known use cases
- ✓Simple solution and support

Current Problem

- Media and entertainment applications deploy parallel filesystems such as StorNext (SNFS) to support high-bandwidth / low-latency requirements
- Current solutions available consist of multiple components often provided by multiple vendors
  - Application servers accessing filesystem
  - Storage array / FC SAN for user data storage
  - Servers for filesystem metadata controllers
    - May include additional storage arrays solely for metadata storage
- Results in a complex deployment and support chain
- Opportunity for consolidation and simplification

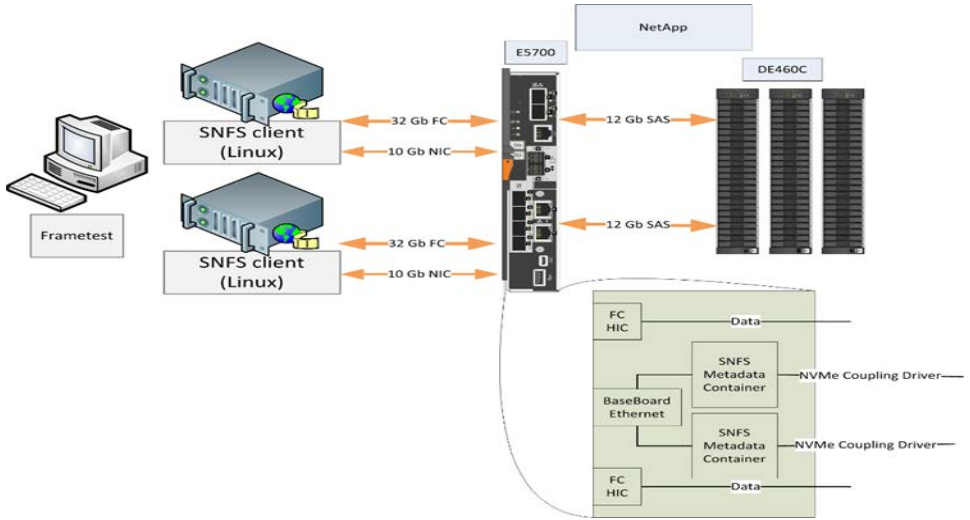
Current/Proposed Solution

Current Media and Entertainment Configuration



- **Heterogeneous solution**
- **Complex support system**
- **Multiple storage arrays**
- **Expensive solution & more rack space (12U)**

NetApp ESG Media and Entertainment Configuration



- **Homogeneous Solution**
- **Simple Support System**
- **Metadata and Data storage consolidated**
- **Metadata server eliminated / integrated into storage**
- **Cheaper solution & less rack space (6U)**

Proposed solution

- **Move SNFS-MDC into the E5700 using the container Infrastructure and a Linux VM**
  - **Consolidates hardware**
  - **Integrated MDC has low-latency NVMe path to metadata storage**
- Provide SNFS client Ethernet access to containerized MDC
  - E5700 baseboard 10Gb Ethernet port bridged to the container/VM
- Provide intermix of host I/Os and container I/Os (Intermix of SCSI and NVMe)
  - Protect NVMe connections from SCSI Task Management functions
- Coupled with functional improvements to storage array firmware to simplify configuration

Move SNFS-MDC into the E5700 using the container Infrastructure and a Linux VM

- Use E-Series container infrastructure to install KVM and run filesystem metadata controller software.
  - Container infrastructure new for E5700 storage array.
  - Could create a Docker container instead but KVM was the quickest path to install and satisfy environmental dependencies for proof-of-concept
    - Docker would almost certainly provide even better performance, but even with KVM we were able to prove solution satisfies requirement of use-case.
    - KVM virtio scsi interfaces add ~250us to reads to E-Series LUNs, so there is room to improve with more effort.
  - Deployment flexibility: KVM environment or Docker container

Provide SNFS Client access to containerized MDC

- Mount an E-Series LUN to install media and provide local storage for VM.
  - SNFS GUI supports SCSI block devices when creating file systems.
  - E-Series coupling driver presents LUNs as NVMe block devices.
  - Used NVMe block devices as storage for virtio SCSI LUNs, so SNFS accepted them as SCSI devices.
- Client accesses MDC via ethernet bridge.
  - Update iptables and deal with local networking/firewall to allow virtual install and subsequent access for configuration.

Intermix of SCSI and NVMeoF

- Provide new use case because of SCSI host access all the way through for data LUNs and MDC traffic through Ethernet with NVMe I/Os for Metadata namespaces.
- Intermix of host I/O and container I/O in E-Series:

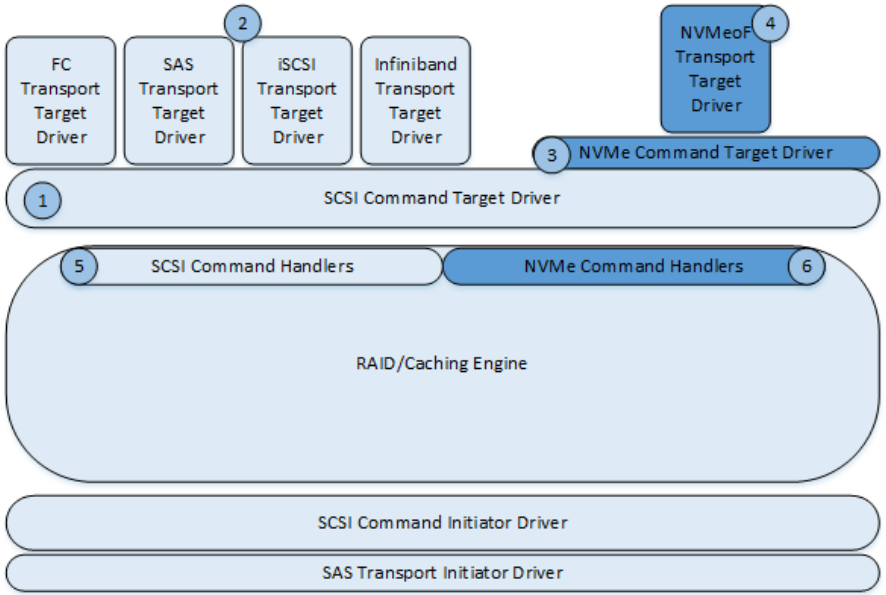
1. Translates I/O commands between NVMeoF and SCSI as the commands traverse the software stack, allowing for maximum re-use of the existing redundancy, performance, and availability capabilities of the software,
2. Handles the different mechanisms by which NVMeoF and SCSI attached hosts deal with error conditions and asynchronous events with a largely common implementation,
3. Provides separate access domains for NVMeoF and SCSI attached hosts to eliminate potential incompatible protocol interactions by not allowing protocol inter-mix within a single storage partition, protect NVMeoF commands from SCSI Task Management requests, and
4. Implements a reservation mechanism for NVMe namespaces and SCSI logical units to provide host clustering capabilities with a common framework for logical host identifier-based NVMe reservations and physical port-based SCSI persistent reservations.

Intermix of SCSI and NVMeoF (2)

Potential inter-mix matrix:

1) Entire storage array in either SCSI or NVMeoF mode, toggled by one switch
2) Hosts or host groups defined as containing either SCSI or NVMeoF connections, with storage array handling abstraction layers
3) Disk pools limited to a single command set or connection type agnostic
4) Data volumes being command-set agnostic basic block storage—can be discovered as SCSI LUNs or NVMeoF namespaces from one discovery to the next*

* dependent on command-set features used



Functional improvements to storage array firmware to simplify configuration (not part of initial POC)

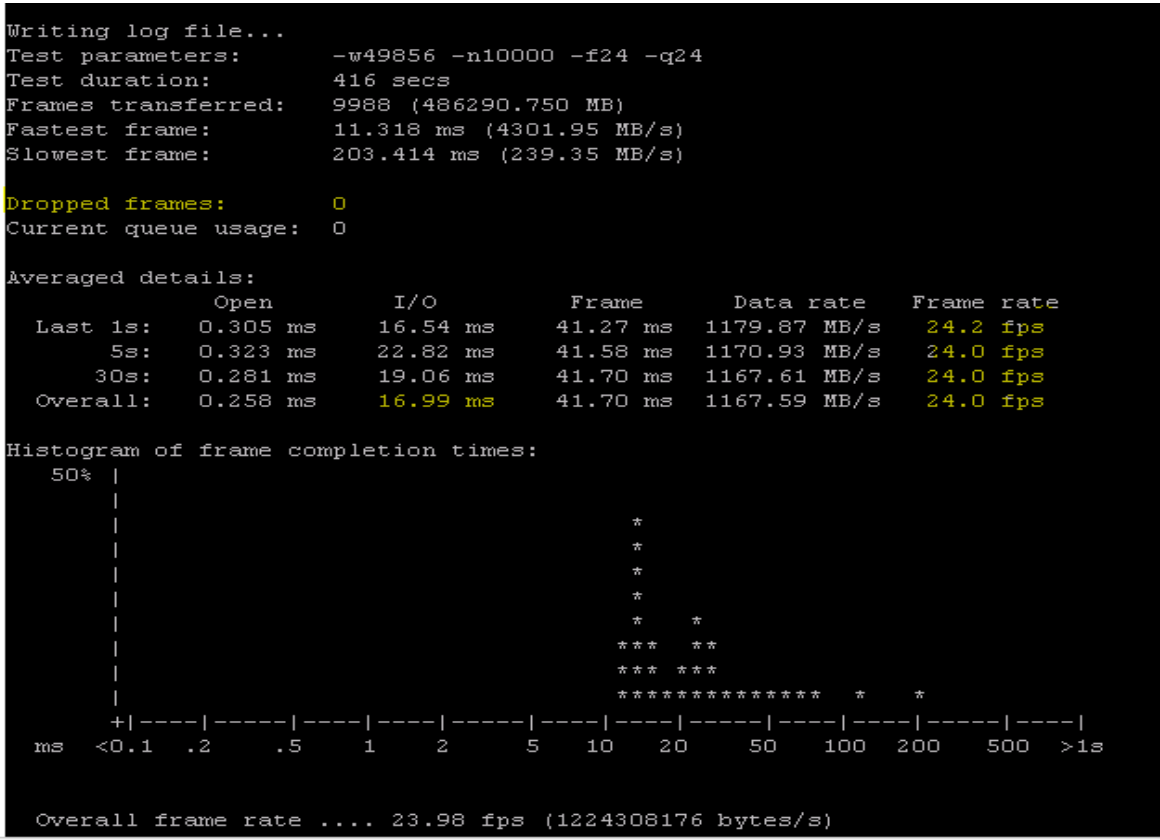
- Host-driver awareness features (currently available for homogenous host deployments in E-Series) for heterogenous host OS deployments
  - Such deployments commonly seen with cross-platform parallel filesystems like SNFS
  - “Default” host access to LUNs without requiring configuration steps to be taken by the administrator
  - Connectivity tracking awareness and target driven LUN failback after recovery from a connectivity fault in the SAN
  - Especially difficult with older Linux multipath solutions that are not ALUA-aware
- Provide failover capability for the Metadata controller running in the container
  - Add failover feature (active/passive or active/active) within the container infrastructure for more robustness

Demo of Containerized Metadata Controller SNFS

- Initial proof-of-concept results without tuning are as such:
- Support for 8 simultaneous users
  - 2 write and 6 read streams
- 4K video streams at 24 fps
- Zero Dropped Frames
- Overall latency: 16.99 ms

Existing implementation with external metadata server can deliver at best:

- Support for 4 simultaneous users
  - 1 write and 3 read streams
- 4K video streams at 24 fps
- ~ 40 msec overall latency



Comparison between Current & NetApp ESG M&E configurations

- Eliminated MDC appliance & disk shelf for appliance
  - **36%** reduction in **cost**
  - **50%** reduction in **rack space**
  - **Simplified support chain** by reducing hardware vendors from 3 to 1
- Twice the throughput of current M&E configuration with double the number of users
  - Without tuning
  - With storage array headroom available

Key Takeaways and Closing Thoughts

- Proof of Concept demonstrates consolidated storage for user data and filesystem metadata in a single appliance.
  - Server consolidation using Linux container infrastructure to better utilize hardware
  - Simplified supply chain and support from reduced hardware footprint
  - Offer more value (twice the throughput with more headroom for tuning, 36% cost reduction, ½ the rack space “6U vs 12U”, less power)
- Specialized use of multiple I/O interfaces on storage array
  - Intermix of host I/O and container I/O utilizing SCSI and NVMe simultaneously
  - Provides for low-latency I/O where needed using NVMe yet preserves existing infrastructure investment by providing for client I/O via SCSI
  - Filesystem metadata requests directly served from the storage array over Ethernet

Team Recognition

- Thanks to NetApp ESG engineers for their contributions to this project:
  - Dean Lang
  - Anthony Gitchell
  - Amine Bennani


NetApp
Thank You