PowerScale OneFS: NFS over RDMA for Media

October 2021

H18987

White Paper

Abstract

OneFS 9.2 now supports NFS over RDMA. The increased throughput and reduced latency of RDMA positions Dell EMC PowerScale for modern high-performance media workflows.

Dell Technologies

Copyright

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Dell Inc. believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Contents

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email (subject line: Feedback for document: H18987).

Author: Gregory Shiff

Note: For links to additional documentation for this solution, see the PowerScale Info Hub.

Using NFS over RDMA for media creation workflows

High resolution and high frame rate video are widely adopted for modern content creation. Adoption of Remote Direct Memory Access (RDMA) on Dell EMC PowerScale ensures that storytellers can seamlessly use these formats. Streaming providers have turbo charged the pace of new content. PowerScale OneFS provides the performance and peace of mind that creators can keep up with the opportunities ahead.

PowerScale OneFS's new NFS over RDMA implementation results in dramatic performance increases in storage performance and reductions in workstation CPU utilization. RDMA allows for high frame-rate and high-resolution playback on PowerScale that was previously the domain of Fibre Channel SAN storage. Playing and processing this content is challenging even for modern CPUs. RDMA reduces the CPU cycles required to use network storage, reserving more cycles for displaying and processing video.

RDMA is a technology that has been around for some time. However, its adoption has mostly been in the high-performance computing industry. With the release of OneFS 9.2, the PowerScale family of storage makes adoption of NFS over RDMA a reality for content producers.

I have recently completed NFS over RDMA testing with media applications at the Dell Media lab in Seattle, WA and the Dell Solutions lab in Round Rock, TX. The tests show excellent results in low-latency and high throughput workflows. Simply mounting the OneFS file system using RDMA (as opposed to TCP) made a huge difference in terms of storage performance and workstation CPU utilization.

This white paper provides details about RDMA, high-resolution professional video formats, and how all these technologies come together on PowerScale OneFS.

OneFS uses the RDMA over Converged Ethernet (RoCEv2) protocol for doing RDMA traffic. This is an effective combination. In the professional video world, there is a lot of interest in high frame rates and 8K resolution. Understanding how these standards impact storage utilization and workstation performance is crucial to success. Storage is a big part of the equation, but it is not the whole equation. Workstation CPU load can be just as consequential, as the following sections describe.

What is NFS over RDMA?

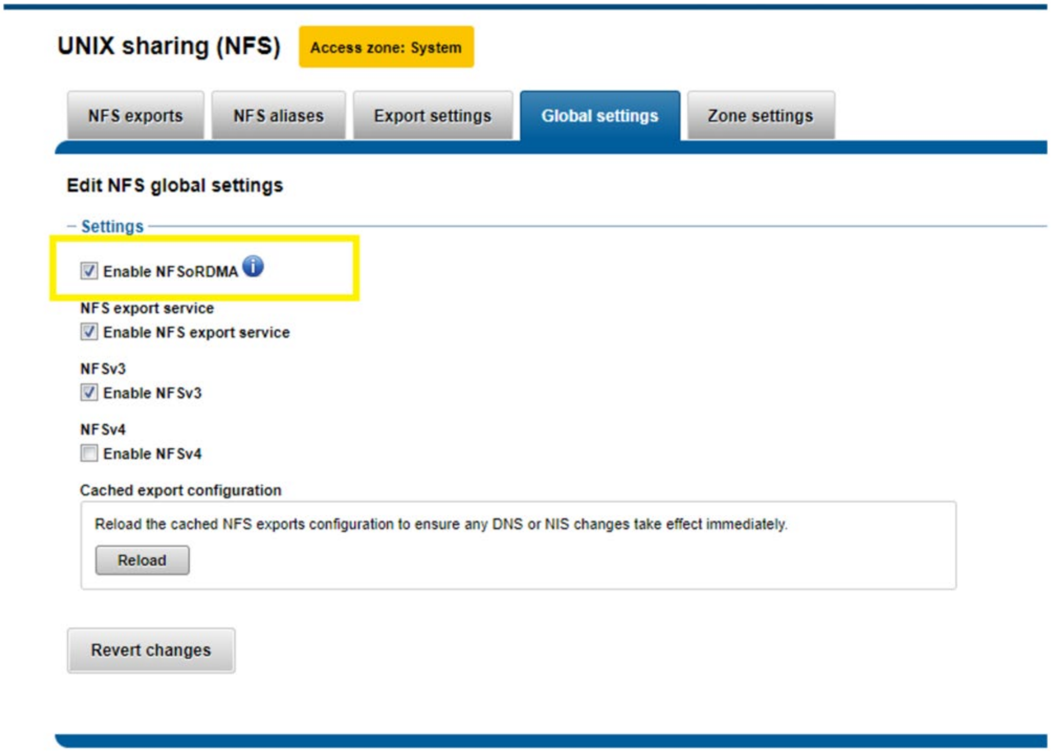
Remote Direct Memory Access (RDMA) is a protocol that allows for a client system to copy data from a storage server's memory directly into that client's own memory. This allows the client system to bypass many of the buffering layers inherent in TCP. This direct communication improves storage throughput and reduces latency in moving data between server and client. It also reduces CPU load on both the client and storage server.

RDMA was developed in the late 1990s to support high performance compute workloads running over InfiniBand networks. In the 2000s, two methods of running RDMA over Ethernet networks were developed: iWARP and RoCE. There are various benefits and drawbacks in these two approaches:

- iWARP uses TCP for RDMA communications. iWARP's reliance on TCP allows for greater flexibility in network design but it suffers from many of the same performance drawbacks of native TCP communications.
- RoCE uses UDP for RDMA communications. RoCE has reduced CPU overhead as compared to iWARP, but requires a lossless network (though the Mellanox implementation of "Resilient RoCE" enables RoCE to gracefully handle some packet loss).

OneFS's RDMA implementation uses the RoCE protocol. Or more specifically, OneFS uses the RoCEv2 protocol (also known as "Routable RoCE"' or "RRoCE"). Given that RDMA in OneFS is meant for maximum storage performance and lowest CPU utilization, the choice of RoCEv2 as the RDMA implementation in OneFS makes sense.

OneFS 9.2+ is required to run NFS traffic over RDMA. NFS over RDMA is enabled in OneFS with a simple checkbox in the global UNIX Sharing settings.



Edit NFS global settings

Settings

☑ Enable NFSoRDMA

NFS export service
☑ Enable NFS export service

NFSv3
☑ Enable NFSv3

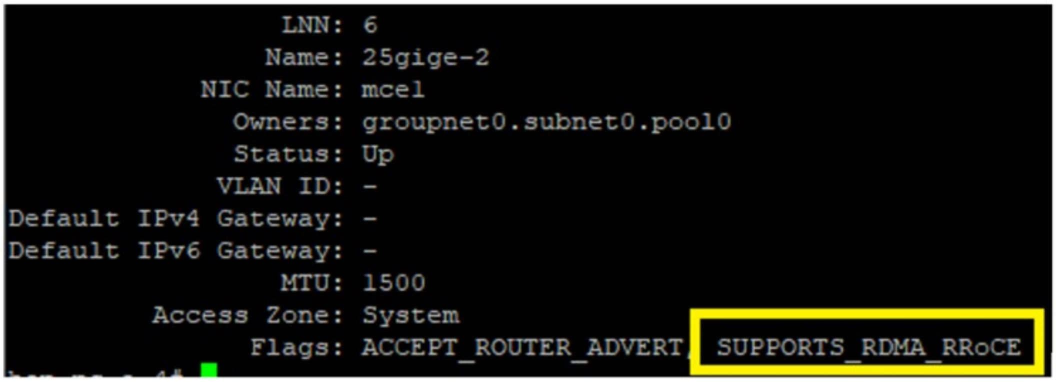
NFSv4
☐ Enable NFSv4

Cached export configuration
Reload the cached NFS exports configuration to ensure any DNS or NIS changes take effect immediately.

The OneFS hardware must have Mellanox ConnectX-3 Pro, ConnectX-4, or ConnectX-5 front-end NICs. To determine NIC type, run the following commandfrom the OneFS command line:

```
isi network interfaces list -v
```

The text “SUPPORTS_RDMA_RROCE” will be present in the flags section on the interfaces that support RDMA.



```
                 LNN: 6
                Name: 25gige-2
            NIC Name: mce1
              Owners: groupnet0.subnet0.pool0
              Status: Up
             VLAN ID: -
Default IPv4 Gateway: -
Default IPv6 Gateway: -
                 MTU: 1500
         Access Zone: System
               Flags: ACCEPT_ROUTER_ADVERT, SUPPORTS_RDMA_RRoCE
```

The client systems must be running a Linux distribution with NIC hardware and drivers that support NFS over RDMA, such as Mellanox ConnectX-3 (and higher) or ATTO FastFrame3 NICs. For details about mounting a OneFS file system using RDMA, see How to Configure NFS over RDMA.

For considerations when preparing an environment for NFS over RDMA traffic, see the section “NFS over RDMA” in Dell EMC PowerScale OneFS Best Practices.

ATTO Technology has also developed OneFS specific profiles that simplify configuring NFS over RDMA with their FastFrame3 family of NICs. For details, see Maximize Ethernet Bandwidth on Dell OneFS Storage with ATTO SmartNICs using NFS over RDMA.

Professional video formats, resolutions, and frame rates

To understand the tools used and the test results, it is worth reviewing some vital background information about various frame rates, image sizes, and types of video files.

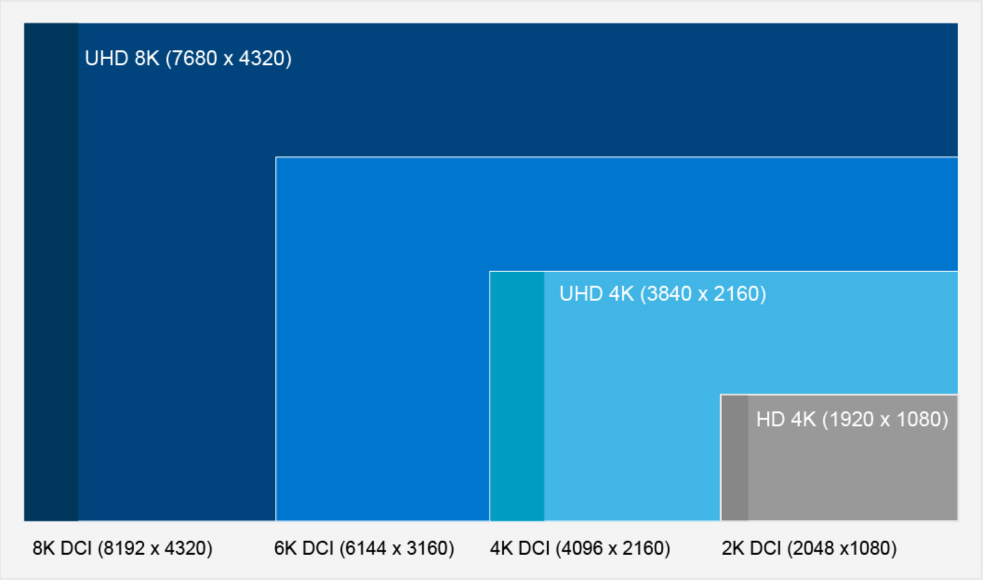
Video resolution

Some of the most common video resolutions used today are: High Definition (HD), 4K and 8K. Some studios use another resolution, 6K, as an intermediary format, though that resolution is rarely used for delivery. There are other resolutions, but HD, 4K, 6K, and 8K provide a good overview of what is found in real world productions.

Unfortunately, there is some confusion about the actual resolution of 4K and 8K video. When referencing 4K and 8K resolutions, two sets of numbers are used. Consumer TV manufacturers came up with 4K resolutions by doubling the height and width of HD resolution (and then doubling those numbers again for 8K resolution). The cinematic world used the Digital Cinema Initiatives (DCI) specifications which determined 4K and 8K resolutions by multiplying the height and width of 2K resolution.

To separate out these competing numbers, TV 4K and 8K images are referred to as Ultra-HD (UHD) 4K and UHD 8K. Cinema resolutions are referred to as 4K DCI or 8K DCI. Confusion arises when video is referred to as 4K or 8K with no mention of UHD or DCI, leaving the actual resolution in question.

The chart below illustrates various resolutions, both UHD and DCI. In terms of absolute resolution, 4K resolution has four times the resolution of HD (or 2K DCI) and 8K has four times the resolution of 4K. If the height and width of an image are doubled, the resulting image has four times greater resolution.



Video frame rate

The next major driver of video performance is frame rate (how often the image is changing on the screen). As the old saying goes, “The nice thing about standards is that you have so many to choose from.” The same holds true for frame rates.

In the United States, TV is commonly 29.97 frames per second. The reason for this weird number has roots in the introduction of color TV in the 1950s. In Europe (and in most of the world), 25 frames per second is the standard. Cinema is typically at 24 frames per second. Adoption of higher frames rates, such as 60 frames per second or higher, is being driven (largely) by the video game industry, where these frame rates are typical.

Most moving images that people are used to seeing are somewhere between 24 to 30 frames per second. Watching video at these frame rates feels “natural”. This “natural” feeling is so ingrained in us that for some, watching video at 60 frames per second feels “artificial”.

There is vigorous artistic debate about the merits of high frame rate video (>30 frames per second) with good arguments on either side. For content such as live sports, there are real benefits to be gained from high frame rate video and being able to view the nuances of an athlete’s performance. In some immersive environments, higher frame rates also seem to reduce the incidence of dizziness.

Video files and image sequences

Professional content creation workflows deal with both video files and image sequences. Video files and image sequences can be compressed or uncompressed. This compression is referred to as the video codec (compressor/decompressor). Video is compressed to save space or bandwidth and then decompressed while being read. Video codecs can either be lossy (where some of the image data is discarded), or lossless (where no data is lost). Examples of codecs are H.264 (a common consumer delivery codec), ProRes (Apple's codec designed for video editing), and Sony XAVC (used by some Sony cameras to capture video). PIZ compression is a lossless codec used with EXR image sequences, as described below. Technical and artistic requirements drive choice of codec, compression, and file type.

Video files can contain multiple tracks of audio along with video and are stored in a "container" format such as .mov or .mxf. The file format is referred to as a "container" because the file extension does not define which codec the video (and audio) stored within uses. For instance, a .mov file could contain video that is encoded with ProRes, H.264, or some other codec.

Image sequences store each frame of video in a discrete file. Some common image sequence formats are DPX and OpenEXR (EXR). DPX is an uncompressed format. Reading back an uncompressed DPX image sequence puts less load on the workstation because the machine does not have to expend CPU cycles decompressing the video frames. However, this benefit comes at the expense of more storage space and bandwidth consumed while playing back these large files. EXR image sequences can be compressed or uncompressed. As mentioned above, PIZ compression is a lossless compression type that can optionally be used in EXR image sequences. PIZ compression saves storage space, because each frame is compressed, and all image data is retained on playback because the compression is lossless. Playing back PIZ compressed EXR image sequences puts a heavy load on the computer reading those image sequences because the system has to decompress each frame in real time.

Image sequence playback also tends to be challenging for storage performance. To address this challenge, PowerScale OneFS has specific file-format aware techniques for dealing with image sequences, as outlined in this white paper: Isilon Filename Based Prefetch.

So why use a video file or an image sequence? As with codec choice, there are technical and artistic reasons to use each type. For example, some visual effects and animation software function better with image sequences because the intraframe compression of a video file does not play nicely with those applications. However, image sequences, with their file-per-frame format, can be unwieldy to store and transfer, making video files a better choice in many circumstances.

Putting it all together

Having a high-level view of how video is stored and displayed in professional environments is necessary for interpreting the benefits of NFS over RDMA. There are other factors not covered here, such as color bit-depth and image dynamic range, but those topics are beyond the scope of this white paper.

Understanding video resolution, frame rate, codec, and whether the material is an image sequence or video file is important. Video files may contain video and audio. These files may be compressed or uncompressed. The file extension ("container") does not specify how video data within is encoded. Compressed video saves storage space and bandwidth at the expense of increased workstation CPU cycles. Lossy compression does not retain all the original image data. Lossless compression retains all the original image data. Image sequences store video with each frame as a separate file. Some applications work better with image sequences. Image sequences may be compressed or uncompressed. With both types of video data, the higher the resolution and higher the frame rate, the greater burden is put on storage and client CPU.

Testing architecture and applications

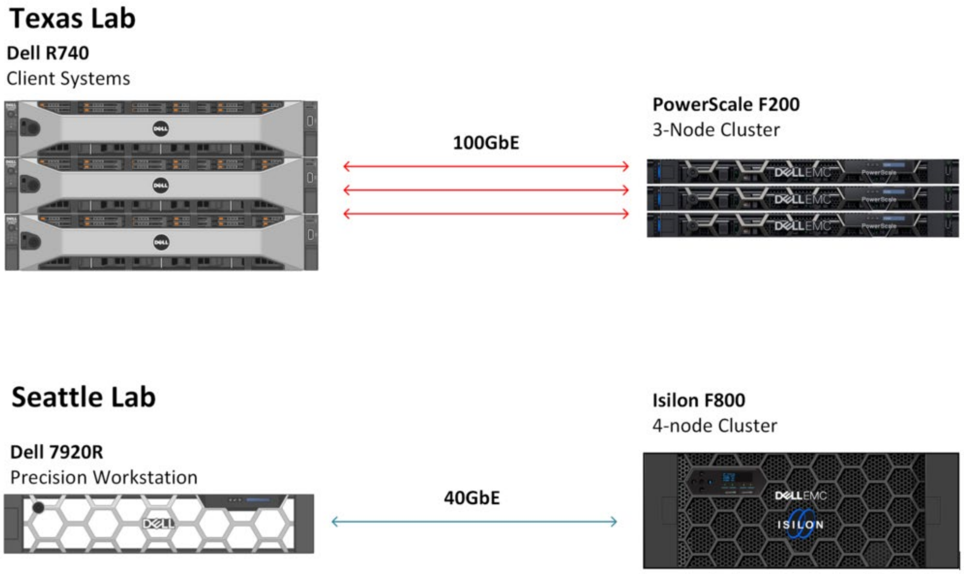
Testing was performed in Dell labs with various applications. These applications include:

- FileStream - An internal tool for simulating image sequence workloads
- Autodesk Flame 2022 - A workstation intensive visual effects and compositing application
- DaVinci Resolve 16 - A color correction application optimized for high-resolution playback

At the Dell Solutions lab in Round Rock, TX, a 3-node F600 cluster running OneFS 9.2.2 was connected using 100 GbE to three GPU equipped Dell R740 servers. The network ports were configured at 9000MTU and flow control enabled on switch ports.

Two sets of tests were run in this environment: FileStream and DaVinci Resolve. The client systems ran a base install of CentOS 7.9. The only performance tweaks performed on the client side were configuring the system BIOS for performance mode and disabling SE Linux in /etc/selinux/config. On the OneFS storage side, streaming mode was enabled for directories with video files and filename-based prefetch was enabled for directories with image sequences. OneFS inline deduplication and compression were disabled.

In the Seattle media lab, a 4-node F800 cluster running OneFS 9.2.2 was connected to a Dell Precision 7920 workstation. The network ports were configured at 9000MTU and flow control enabled on switch ports. FileStream and Autodesk Flame 2022 were tested against this cluster. The workstation ran the Autodesk provided Flame ISO with no tweaks beyond configuring the workstation BIOS for performance mode. An AJA Kona4 card provided broadcast output on the Flame workstation. The F800 had streaming mode enabled for directories with media files and filename-based prefetch configured for directories with image sequences.



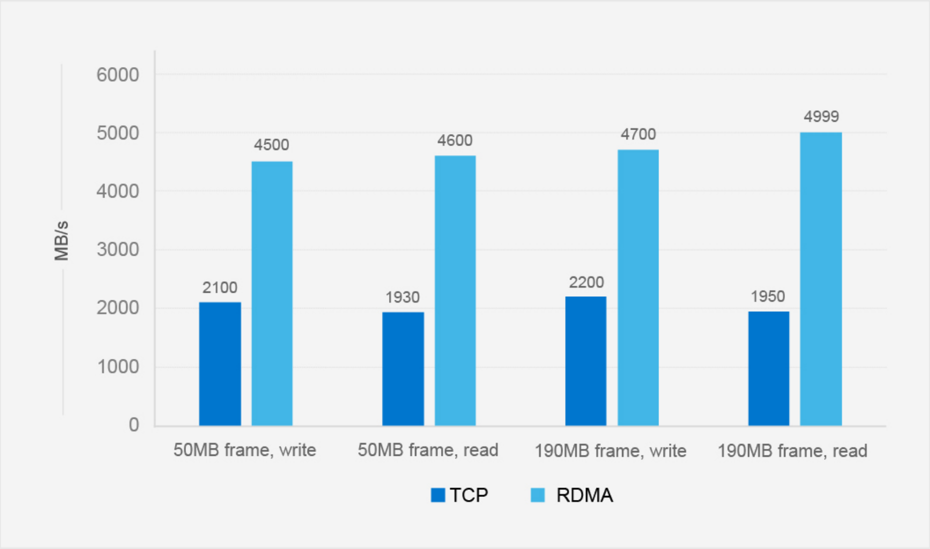
FileStream testing results

FileStream – F600 results

The below chart shows results of FileStream writing and reading to an F600 cluster from a single client over 100 GbE. Filestream was configured with four threads and set to read or write an image sequence to the F600 as fast as it could. Two different frame sizes were tested:

- 50 MB frame - Simulating a 4K DCI resolution image sequence
- 190 MB frame - Simulating an 8K DCI resolution image sequence

In the following figure, the Y-Axis is in MB per second. A higher bar indicates a better result.



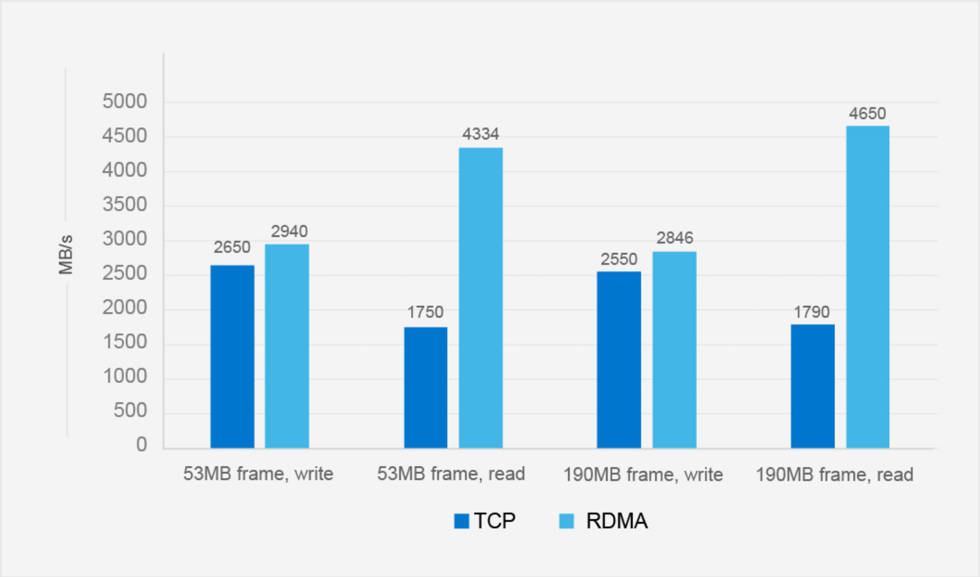
Notable takeaways from this testing are read results for the 4K frame size (50MB) and 8K frame size (190MB). For the 4K frames, a frame rate of 92 frames per second was achieved. This frame rate is more than enough to support 60-frames per second uncompressed 4K playback. On the 8K resolution frames, 26 frames per second was achieved, supporting the cinematic 24 frames per second standard.

FileStream – F800 results

The following figure shows the results of FileStream writing and reading to an F800 cluster from a single client over 40 GbE. As with the F600 testing, FileStream was configured with four threads and set to read or write an image sequence to the F800 as fast as it could. 50 MB and 190 MB frames were tested.

Both OneFS and Linux caches were cleared between runs.

In this figure, the Y-Axis is in MB per second. A higher bar indicates a better result.



Notable takeaways from this testing are the read results for the 4K frame size (50MB) and 8K frame size (190MB). For the 4K frames, a frame rate of 86 frames per second was achieved. This frame rate is more than enough to support 60-frames per second uncompressed 4K playback. On the 8K resolution frames, 24 frames per second was achieved.

Interestingly, the write performance of F600 with RDMA is far greater than that for the F800. This is likely due to the new architecture and drive technologies in the F600 PowerScale nodes. The numbers are impressive.

FileStream Caveat

FileStream is a good indicator of how hard a single client can push an NFS connection with a limited number of threads. It shows what is possible and how large of a delta exists between TCP and RDMA storage performance.

However, FileStream is a storage load generator and has little to no impact on workstation performance. Presumably, in the real world, video is being displayed and processed in some form or another. With demanding media such as 60 frames per second 4K or 24 frames per second 8K, the workstation's ability to both read data from the storage and do something with it needs to be factored into the equation.

Autodesk Flame 2022 results

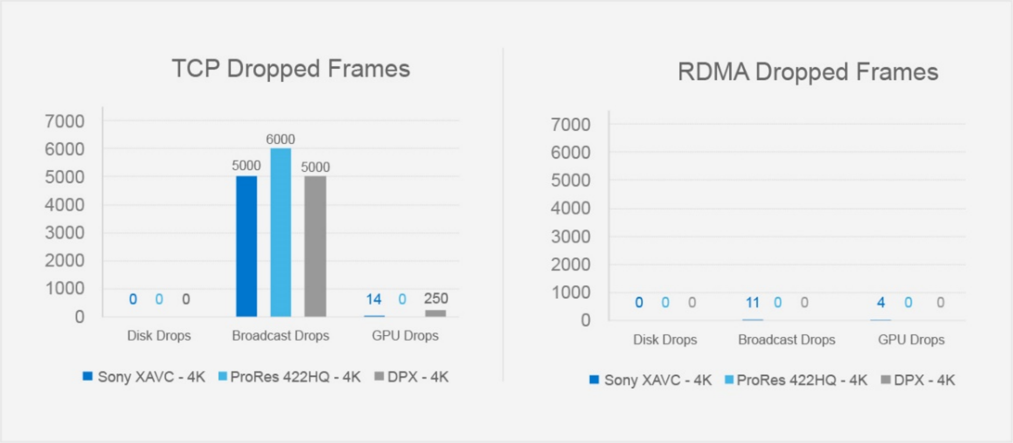
Autodesk Flame is a sophisticated visual effects compositing and finishing application. Flame is demanding both in terms of workstation performance and storage performance. To verify systems can meet Flame's requirements, Autodesk has implemented a regimented qualification process for Flame. The qualification of PowerScale and OneFS 9.2 with Flame 2022 provided a perfect chance to compare TCP and RDMA performance.

Workstation CPU and GPU utilization were more of a limiting factor in these tests than storage performance. Often the differences between TCP and RDMA were minimal because workstation bottlenecks were reached before storage performance. One notable exception was real-time playback of 59.94 frames per second 4K DCI material. With this test, the extra CPU cycles afforded by RDMA resulted in a major improvement in broadcast output.

The full suite of Flame 2022 tests was run against a 4-node PowerScale F800 cluster connected using 40 GbE to a Dell Precision 7920 workstation. On the Dell workstation, system BIOS was configured for performance mode and the Flame 2022 ISO installed. On the PowerScale F800 cluster, streaming mode was enabled for directories with video files and filename-based prefetch configured for directories with image sequences.

The real-time playback test consists of playing back challenging material with debug mode enabled in Flame. Debug mode displays dropped frame counters for video disk (PowerScale), GPU, and broadcast output. While running the tests with both TCP and RDMA, the “video disk” never dropped frames. The video disk, PowerScale F800, delivered enough performance.

By contrast, with the file system mounted using TCP, there were many dropped frames in the broadcast output. By mounting the file system using RDMA, there were few dropped frames from the broadcast output. The impact was dramatic with upwards of 6000 dropped frames for TCP compared with 11 (yes, 11) dropped frames for RDMA. As mentioned, the video disk never dropped frames. It was all about freeing up the workstation to send video to the broadcast output (in this case, an AJA Kona4 card).



DaVinci Resolve play back results

DaVinci Resolve is primarily used in color correction workflows. It is optimized for playback of high-resolution media. For the DaVinci Resolve benchmarks, three client systems were connected to a 3-node PowerScale F600 cluster using 100 GbE. The client systems were Dell R740 servers with NVIDIA GPUs. The BIOS in these systems was configured for performance mode. CentOS 7.9 was installed and SE Linux disabled. On the PowerScale cluster, streaming mode was enabled for directories with video files and filename-based prefetch enabled for directories with image sequences.

8K DCI DPX image sequence playback, 24 frames per second

The first test was playing back an uncompressed DPX image sequence at 8K DCI resolution, 1x image sequence per client. These frames were each approximately 190 MB. Each client would need 4,560 MB/s of performance to sustain 24 frames per second playback.

The 3-node F600 cluster could sustain playback of 2x simultaneous 8K image sequences total using RDMA. Adding the third client caused dropped frames. Rerunning the test with the F600 mounted using TCP, the 8K image sequences would not play at all, showing that RDMA was vital in this workflow.

6K PIZ compressed EXR image sequence playback, 24 frames per second

8K uncompressed image sequences are an unwieldy format that is not widely used. 8K uncompressed uses up a tremendous amount of storage space and bandwidth. By comparison, 6K PIZ compressed EXR image sequences are a more common format in cinematic production.

6K resolution provides enough flexibility to zoom into a portion of the image and maintain sharpness. PIZ compression is a lossless codec, meaning image quality is maintained while still saving storage space and bandwidth. The challenge posed by PIZ compressed EXR is that it is extremely taxing on the workstation. The individual frames are still large, often over 100 MB each, and the workstation must decompress each frame in real time.

The PIZ compressed EXR image sequence used in this testing had frame sizes ranging from 80 MB to 110 MB. The frame size varied depending on how well each individual frame compressed. This frame size meant each client would require 2640 MB/s of performance.

The 3-node F600 could sustain playback of 3x simultaneous PIZ compressed image sequences. It is likely that the cluster could have sustained even more, but the three clients could only handle one of these challenging image sequences each. As with the 8K material above, with the file system mounted using TCP, the images sequences would not play back, RDMA is vital.

4K DCI DPX image sequence playback, 60 frames per second

Sixty frames per second 4K media is probably the most common format of the ones tested with DaVinci Resolve. Particularly in the live sports world, the higher frame rate makes fast moving action clear. This material had frame sizes of approximately 50 MB each. To sustain 60 frames per second playback, each client would be pulling 3000 MB/s.

The 3-node F600 cluster could sustain 3x 60 frames per second image sequences, 1x per client. As with the other two tests above, playback using TCP was not possible. The extra bandwidth and reduced CPU load of RDMA enabled this workflow.

Conclusions

Using NFS over RDMA resulted in impressive outcomes in the content creation scenarios tested. In particular, it enabled workflows that TCP-based NFS is unable to support. For demanding post-production applications running in Linux, use of RDMA with PowerScale OneFS is an obvious choice.

To quickly recap some of the findings:

- FileStream - On PowerScale F600 nodes, RDMA doubled performance compared to TCP.

- Autodesk Flame 2022 - Using RDMA with 59.94 frames per second 4K DCI video reduced dropped frames from the broadcast output from 6000 to 11.
- DaVinci Resolve 16 - RDMA enabled workstations to play uncompressed 8K DCI, PIZ compressed 6K, and 60 frames per second 4K DCI content. None of this media would play using TCP.

In terms of speeds and feeds, RDMA is a clear winner for these workloads. But just as important is the reduction in load placed on the workstation CPU. For instance, playing 60 frames per second 4K DCI is demanding. Even when the storage can support it, freeing up the workstation CPU is critical to smooth playback.

RDMA is not for everything. In circumstances with high thread and connection counts, other bottlenecks may be reached. In these scenarios, there is little difference between RDMA and TCP. But for “hero” seats that place the highest demand on storage and workstation, RDMA is by far the best choice.

References

Dell Technologies documentation

The following Dell Technologies documentation provides additional information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- PowerScale Info Hub
- How to Configure NFS over RDMA, Nick Trimbee, Dell Technologies
- Dell EMC PowerScale OneFS Best Practices
- Isilon Filename Based Prefetch