Overcoming Performance Bottlenecks With a Network File System in Solid State Drives

David Flynn, Hammerspace

Thomas Coughlin, Coughlin Associates, Inc.

Since the introduction of nonvolatile memory express, data-driven applications and workload requirements have changed dramatically. A decentralized storage architecture with elements of file systems embedded in solid state drives can help support these new applications.

The need for speed is nothing new to the data compute and data storage industries. With exponentially larger quantities of data being created, analyzed, and processed, demands for performance are not just growing but accelerating. It has become a business imperative that an organization's data be more accessible and usable to support modern applications and workflows.

INTRODUCTION

Workflows, artificial intelligence (AI) and machine learning (ML) engines, and applications need data generated at the edge, in data centers, and in the cloud to be aggregated and shared over a network or a fabric. This has become critical in today's decentralized workflows with the massive adoption of cloud computing, the increase in edge devices, and the large number of data-driven employees working in remote locations.

As datasets grow and the applications for data-driven insights grow in tandem with them, continual focus is needed to chase down and eliminate those obstacles

Digital Object Identifier 10.1109/MC.2023.3290388
Date of current version: 23 August 2023

0018-9162/23©2023IEEE

EDITOR TOM COUGHLIN
Coughlin Associates; tom@tomcoughlin.com

that prevent putting data to work efficiently. While most innovations are incremental advancements in a single vendor's own technology, on occasion there is a massive leap forward, when multiple technologies converge to address a single problem.

Our industry finds itself now at one of those game-changing moments in the advancement of data path performance. With the rise of AI, ML, and data analytics applications, the timing could not be better. Let us look at how this convergence of communication, storage, and computing technologies will result in a new storage architecture that enables data-intensive, distributed applications.

While most innovations are incremental advancements in a single vendor's own technology, on occasion there is a massive leap forward, when multiple technologies converge to address a single problem.

THE EVOLUTION OF NETWORK-ATTACHED STORAGE PERFORMANCE

Taking a step back, there was a time when a redundant array of independent disks (RAID) controller was a limiting factor in performance.[1] The introduction of nonvolatile memory express (NVMe) and Peripheral Component Interconnect Express (PCIe) allowed direct-attached NVMe-based storage systems to remove the RAID controller from the data path.

For shared storage environments, however, leveraging the native capabilities of the underlying solid state drive (SSD) performance is more challenging. In shared storage use cases, high-performance erasure coding is needed for efficient data protection,[2] and additional physical layers are required within the data path for the file system to map files to blocks and then to map blocks to a flash address. Internal networking is needed in systems to connect and scale-out multiple nodes into a single system to accomplish all of this.

What is important to note is that both the efficiency of the file system software and the physical architecture are important in high-performance environments. Let us delve into the physical architecture further. In scale-out network-attached storage (NAS) solutions using an InfiniBand[3] or Ethernet backplane, there are no fewer than nine data retransmissions with each high-speed serial bus move, going chip to chip or over a cable as shown in Figure 1. Each data retransmission removes a bit of the native performance of the underlying NVMe hardware away from the performance available from the data path, the result being that the client does not see the full native performance of the NVMe.

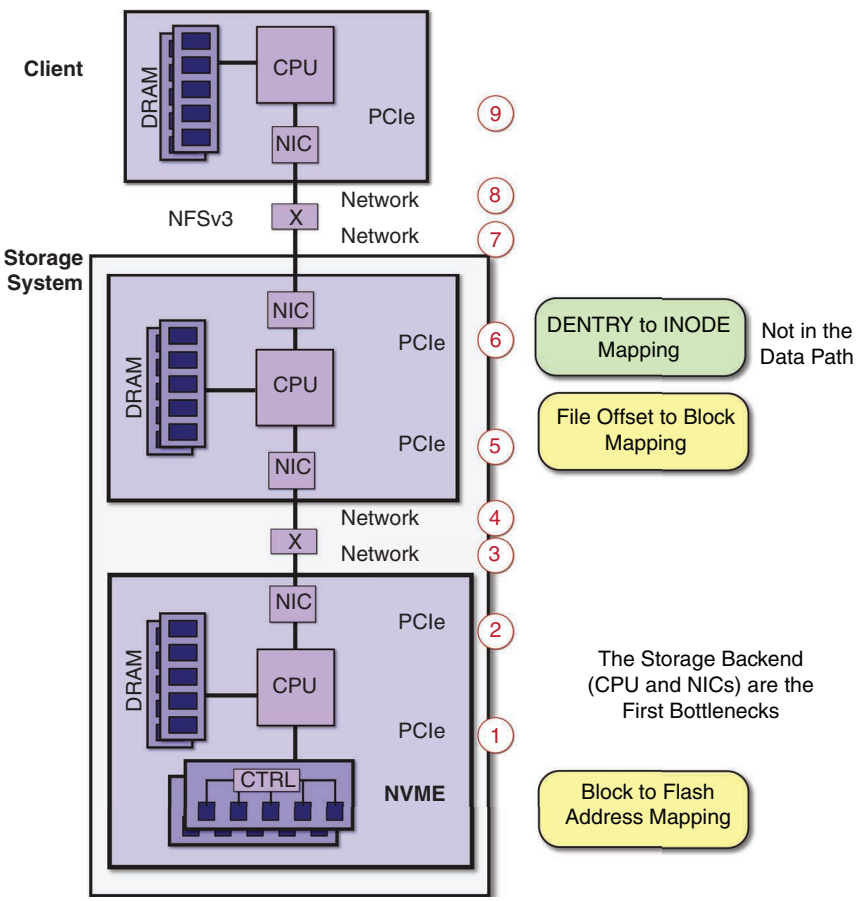

FIGURE 1. Scale-out NAS with NVMe and InfiniBand or Ethernet backplane.

Scale-out NAS solution providers tackling performance workloads saw these challenges and designed their technology upon NVMe-over-Fabric (NVMe-oF).[4] These solutions take the CPU out of the storage server data path and use PCIe for routing to the network adapters. This data path optimization is a big step forward, as shown in Figure 2, but still requires extensive data retransmissions and unnecessary hardware costs. It also causes power consumption inefficiency as a network port is needed for each NVMe device in order to benefit from the full performance of the device.

> The ability to efficiently scale from very small to very large is dramatically simplified with this approach as the storage is directly connected to the in-place Ethernet networks.

With Network File System (NFS) v4.2, the Linux community introduced a standards-based software solution to further drive speed and efficiency.[5] NFSv4.2 allows workloads to remove the file server and directory mapping from the data path, which enables the NFSv3 data path to have uninterrupted connection to the storage.

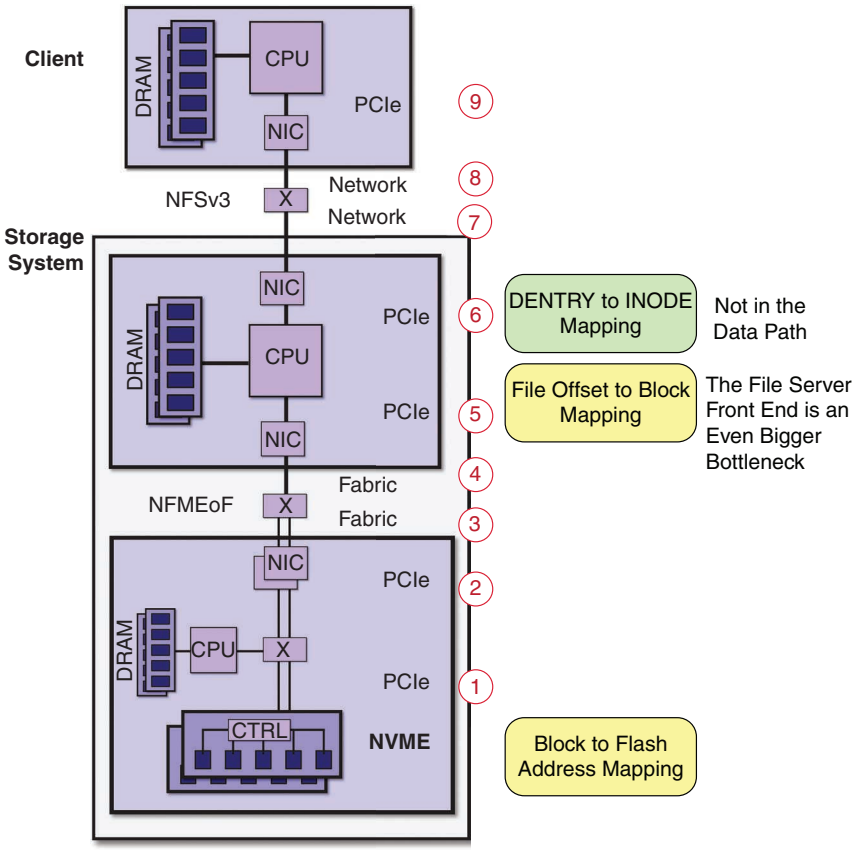

FIGURE 2. NAS using NVMe-oF.

This advancement drops the number of data retransmissions by 44% (from nine retransmissions to five), as shown in Figure 3. With NFSv4.2, NAS environments have the advantage previously only found in parallel file systems, by which the metadata server is out of the data path, removing the added cost and power consumption of an additional internal storage network. The ability to efficiently scale from very small to very large is dramatically simplified with this approach as the storage is directly connected to the in-place Ethernet networks.

Yet another data path efficiency is now available leveraging GPUDirect architectures. This passes a single PCIe hop and copy through the host CPU and memory offering similar data path efficiencies to those of NVMe-oF, as shown in Figure 4.

WHAT IS THE NEXT BIG OPPORTUNITY?

The next big innovation for file storage performance will come when the now-available Ethernet-attached SSDs (eSSD)[6] are able to directly connect to clients with NFS. This would allow clients to achieve nearly the full performance of the underlying NVMe fabric and remove added infrastructure costs.

This innovation finally enables the technology to evolve so that the SSD itself is directly attached to Ethernet (an NFS-eSSD). The SSD would have the ability to communicate with NFSv3 natively, and block-to-flash address mapping could be converted into a single level of mapping. The outcome of this architecture is a further reduction in the number of data retransmissions from nine in traditional NAS and five in parallel file systems to just three when leveraging NFSv4.2 and NFS-eSSD, as shown in Figure 5.

BENEFITS OF NFS IN THE eSSD

The benefits of this new architecture are significant.[7] It provides lower

latency with fewer data retransmissions and lower power consumption since the serial transmission of data across chip-to-chip connections or over wires consumes the vast majority of power in data centers. This approach also has lower operational and capital costs since much extra hardware is eliminated.

For these NFS-eSSDs, write amplification can be reduced since the SSDs become better aware of their free space, avoiding overwrites and improving SSD endurance. Since this reduces the need for overprovisioning, higher capacity densities and higher access densities can be achieved without sacrificing the ability to fully utilize the actual available performance. The overall reliability, availability, and serviceability can be optimized with less hardware and fewer data retransmissions than with conventional NVMe SSDs. As a result, a much wider dynamic scale is possible, allowing storage systems to scale up or down with ease, directly on existing Ethernet networks and using standards-based storage connectivity.

The overall reliability, availability, and serviceability can be optimized with less hardware and fewer data retransmissions than with conventional NVMe SSDs.

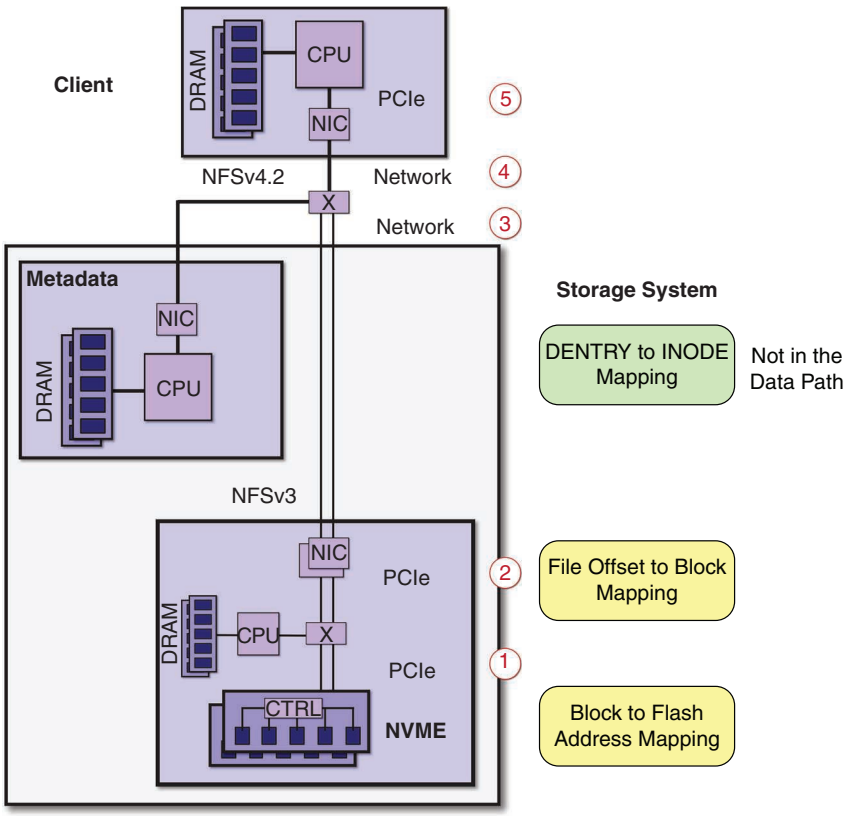

FIGURE 3. NAS using NFSv4.2.

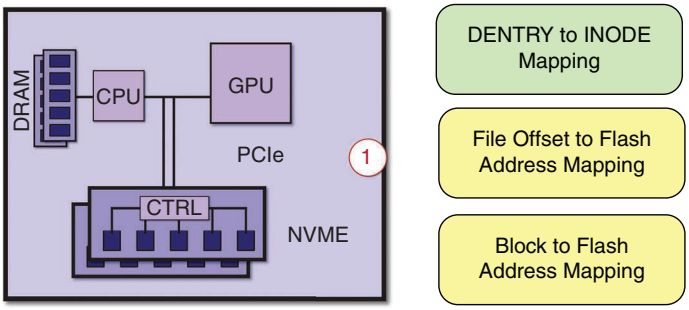

FIGURE 4. Direct-attached storage with GPUDirect.

WHY NOW?

Enormous value can be derived from data with modern data analytics and AI tools, but creating this value requires rapid access to and processing of that data. These data analysis workloads require large datasets, stored on efficient and very high performance storage devices, and the ability to intelligently orchestrate data.

The required high-speed, orchestrated data pipeline must provide a shared view of an organization's data to all systems. It needs to be optimized for small, random I/O patterns and provide high peak system performance and high aggregate file system performance to meet the variety of training workloads an organization may encounter. The data pipeline needs to be driven by a standards-based solution that will be easy to deploy on machines with diverse security and build environments, and the storage systems need to deliver the full performance of the SSD to the workflow in order to maximize opportunity while containing infrastructure costs.

With the broad adoption of high-speed Ethernet, 64-bit processor IP availability, Internet Protocol v6,[8]

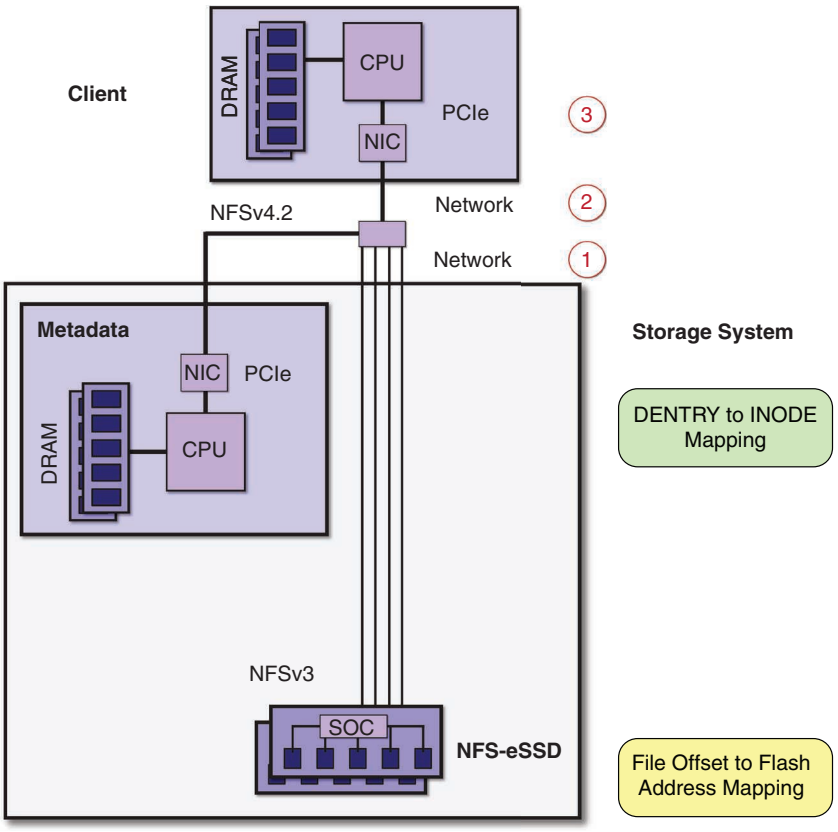

FIGURE 5. NAS using NFSv4.2 and NFS-eSSD.

NFSv4 in Linux, and high-performance lightweight file systems, all of the technology is available to make this possible. The industry just needs to put the pieces together!

ACKNOWLEDGMENT

The corresponding author is Thomas Coughlin.

REFERENCES

1. P. Goodwin, "SSDs and RAID: What's the right strategy, JEDEC at the CES," JEDEC, Arlington, VA, USA, 2011. Accessed: Jun. 22, 2023. [Online]. Available: https://www.jedec.org/sites/default/files/1330-1400%20Paul%20Goodwin%20RAID%20and%20SSDs.pdf
2. J. L. Hafner, V. Deenadhayalan, T. Kanungo, and K. K. Rao, "Performance metrics for erasure codes in storage systems," IBM, San Jose, CA, USA, Tech. Rep. RJ 10321 (A0408-003), Aug. 2004.
3. G. F. Pfister, "An introduction to the infiniband architecture," in *High Performance Mass Storage and Parallel I/O: Technologies and Applications*, H. Jin, T. Cortes, and R. Buyya, Eds. Piscataway, NJ, USA: IEEE, 2001, ch. 42, pp. 617–632.
4. E. Kim and F. Zhang, "Optimizing NVMe® over fabrics (NVMe-oF™)," SNIA, Santa Clara, CA, USA, White Paper, Apr. 2021. [Online]. Available: https://www.snia.org/sites/default/files/education/snia-optimizing-nvme-over-fabrics-nvme-of.pdf
5. S. Dickson. *Linux Support of NFS v4.1 and v4.2*. (Mar. 2017). RedHat. [Online]. Available: https://events.static.linuxfound.org/sites/events/files/slides/Vault2017.pdf
6. T. Vojnovich, M. Carlson, R. Davis, and J. F. Kim, "Ethernet-attached SSDs, brilliant idea or storage silliness?" SNIA, Santa Clara, CA, USA, SNIA Webcast Presentation, Mar. 2020. Accessed: Jun. 22, 2023. [Online]. Available: https://www.snia.org/sites/default/files/ESF/SNIA-Ethernet-SSDs-Final.pdf
7. D. Flynn, "The case for NFS-eSSDs," in *Proc. IEEE Mass Storage Conf.*, May 2023. [Online]. Available: https://storageconference.us/2023/FlynnPresentation2.pdf
8. "Google collects statistics about IPv6 adoption in the internet on an ongoing basis, continuously updated." Google. Accessed: Jun. 22, 2023. [Online]. Available: https://www.google.com/intl/en/ipv6/statistics.html

DAVID FLYNN is the founder and CEO of Hammerspace, San Mateo, CA 94403 USA, and the previous founder and CEO of Fusion-io. Contact him at david.flynn@hammerspace.com.

THOMAS COUGHLIN is the president of Coughlin Associates, Inc., Atascadero, CA 93422 USA. Contact him at thomasmcoughlin@gmail.com.