LA-UR-20-25703
Approved for public release; distribution is unlimited.

Title: Hunting for Bottlenecks in ZFS Failure Recovery using NVMe Drives

Author(s): Bautista, Trevor Scott
Manno, Dominic Anthony
Parga, Alex

Intended for: 2020 HPC Intern Showcase, 2020-08-13 (Los Alamos, New Mexico, United States)

Issued: 2020-07-30 (Draft)

Thanks to currently available storage technology, modern data storage systems are capable of very high bandwidth. Determining which data protection schemes to use for high bandwidth storage systems heavily depends on disk bandwidth and capacity, which determine the amount of time it takes to fully recover data from drive failures. Within HPC, storage systems must balance throughput and data protection according to the goals of each system. With the recent affordability and performance increase of Non-Volatile Memory Express (NVMe) SSD technologies, the significance of disk bandwidth as a potential bottleneck is drastically lowered. However, it is unclear if the overlying filesystem, ZFS, exhibits inherent bottlenecks when not limited by disk bandwidth. Here we hunt for ZFS rebuild and resilver performance bottlenecks when used with NVMe devices. Using various realistic ZFS configurations, we simulate disk failure and measure both the amount of data rebuilt/resilvered and the amount of time this operation takes. These configurations vary from a default production ZFS configuration with no external I/O load to a rebuild-favored ZFS configuration with a heavy user workload. With our results, we expect to provide storage system designers with a better understanding of ZFS rebuild/resilver performance bottlenecks.