RAIT

Rapid Transit for Mass Storage

BY: Fred Moore President
www.horison.com

Abstract

RAIT (Redundant Arrays of Independent Tape) are the tape drive equivalent of RAID (Redundant Arrays of Independent Disks) for disk drives. RAIT is sometimes called "tape RAID". RAIT arrays use the same algorithms that RAID arrays have successfully employed and the motivations for tape arrays are the same as those for disk arrays – much higher bandwidths and increased availability to accompany the parallel movement of large files. RAIT is most beneficial in the HPC, Hyperscale and large enterprise organizations where high-levels of tape throughput and fault tolerance are must-haves, not luxuries.

RAIT first appeared in the 1990s with Data General's CLARiiON Tape Array Subsystem using 4mm DAT tape drives. NCR also announced a tape array software product for NCR uniprocessors and StorageTek had developed a RAIT array implementation for their Redwood helical scan drives. Synchronizing the tape drives in the array was more difficult then as drives had very small drive-level buffers and performance would degrade, as a result none of these early offerings ever gained much visibility or traction from a market perspective. Today's modern tape drives have much larger 1 GB or 2 GB drive buffers for compressed data making drive synchronization routine. The stage is set for the ultra-high data transfer rates of modern tape technology to redefine throughput levels for all storage systems making *rapid transit for mass storage* systems become a reality!

RAID Came First

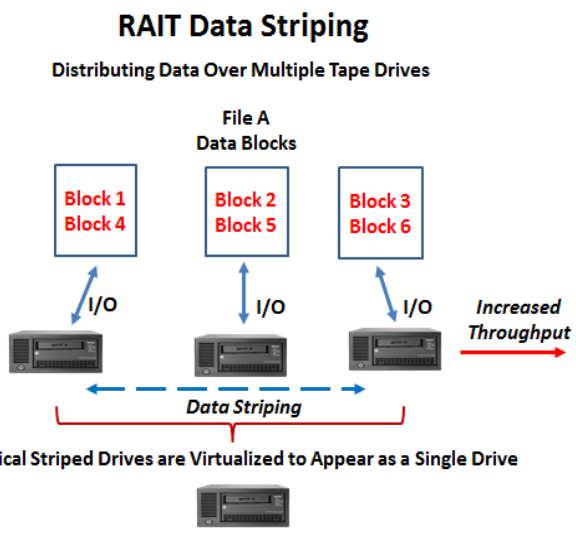
RAID arrays preceded RAIT and were first proposed in the 1980s at the University of California, Berkeley to use parallelism between multiple disks to improve aggregate I/O performance and boost disk drive availability. RAID uses a performance enhancing method of spreading or duplicating data across multiple disk drives called striping. By incorporating parity recording used for error correction, active hard disk drives in an array can be used to provide fault-tolerance if one or more drives fail.

Today RAID is widely used and appears in the product lines of most major disk subsystem suppliers. There are several RAID levels and the one you choose depends on whether you are using RAID for performance or fault tolerance - or both. RAID 2 was rarely used and both RAID 3 and RAID 4 were quickly replaced by RAID 5. It also matters whether you have implemented hardware or software RAID, as software supports fewer levels than hardware-based RAID. In the case of hardware RAID, the type of storage controller used also matters. Different controllers support different levels of RAID and define the kinds of disks you can use in an array such as SAS, SATA or SSDs.

RAIT

RAIT significantly improves the throughput of large sequential files by creating multiple parallel data lanes into the tape subsystem and can also provide various degrees of fault tolerance for higher availability. Fault tolerance means providing a safety net for failed hardware by ensuring that the drive with the failed component can still operate while not impacting availability or increasing the chance of data loss. RAIT levels are implemented in software and depend on how many tape drives you have in an array, how critical drive recovery is, and how important it is to maximize tape performance. The only extra RAIT cost is the amount of space used for parity. The data transfer rate is restricted to the slowest drive in the stripe though ideally all RAIT drives will be the same type. RAIT provides data redundancy without needing to create multiple copies. Striping and parity are the keys to RAID and RAIT implementations.

Data striping segments logically sequential data so that consecutive segments are stored on different physical storage devices. Segments can be either bits, bytes or blocks though block striping is most always used. Striping architectures are useful when a server requests data more quickly than a *single* storage device can provide it. Spreading segments across multiple devices allows data to be processed in parallel and total data throughput is increased. The number of drives in the array is called the stripe width and stripe capacity refers to the amount of data within a stripe. The stripe capacity is limited to the capacity of the lowest capacity drive in the array. The minimum size of a RAIT stripe is 3 and can range to 16 drives depending on the software used.



Parity provides striped RAID and RAIT systems with fault-tolerance capability for higher availability. Parity is an error-detection method used to detect and correct errors in data transmissions by performing specific checks of the data blocks within a stripe. Parity is calculated during write activity and is written to the parity drive(s). In the event of a single drive failure, the information for the missing drive or media can be recreated by examining the remaining data and the parity information. For example, the parity data is computed by XOR'ing (Exclusive OR) a bit from drive 1 with a bit from drive 2 and storing the result on drive 3.

RAID and RAIT Levels

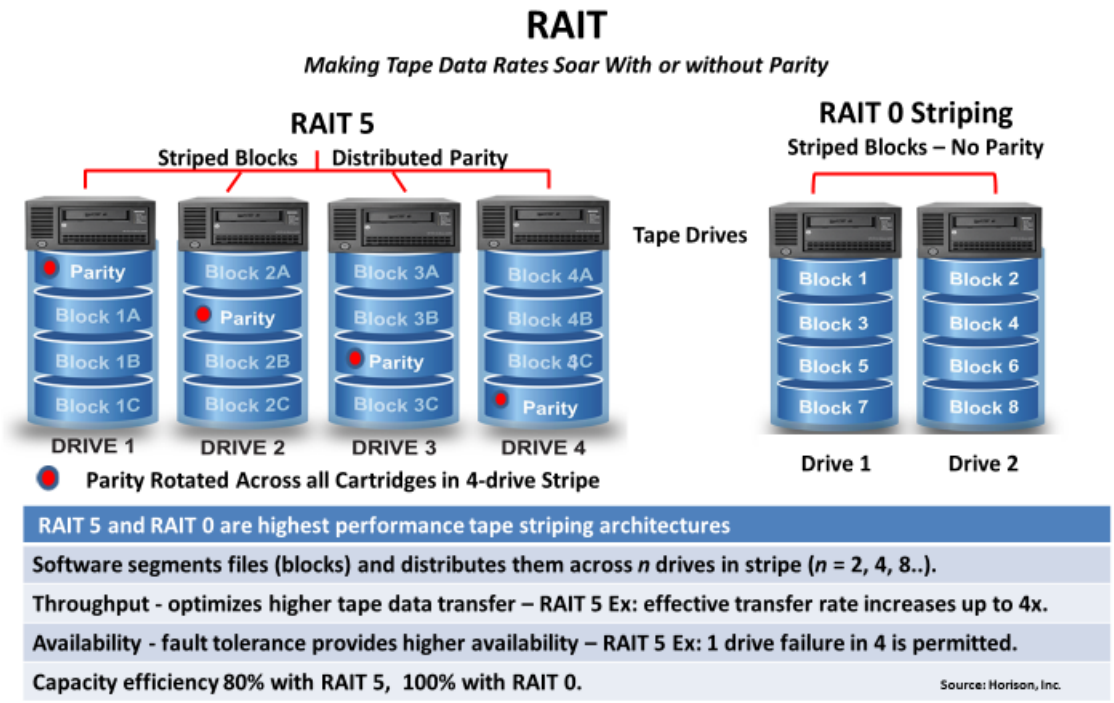
Level	Description	Strengths
RAID 0 **RAIT 0**	Striped array without fault tolerance – min. 2 drives	Provides data striping (spreading out blocks of each file across multiple drives) but with no redundancy. This significantly improves performance but does not deliver fault tolerance. If one drive fails then all data in the array is lost.
RAID 1 **RAIT 1**	Mirroring and duplexing – min. 2 drives	Provides disk or tape mirroring. Level 1 provides twice the read transaction rate of single disks and the same write transaction rate single disks. This is the costliest option as it doubles the number of drives and media required and has a 50% capacity efficiency.
RAID 2	Bit-level striping across multiple drives – min. 3 drives	No practical use and rarely if ever used. RAID 2 stripes data at the bit level rather than the block level. Uses dedicated parity drive. Supplanted by RAID 3 and 4.
RAID 3	Byte-level striping across multiple drives – min. 3 drives	An additional disk is added that calculates and stores parity on dedicated drive. Used for high-performance sequential files. Write requests can suffer from dedicated parity-drive bottleneck.
RAID 4 **RAIT 4**	Block-level striping across multiple drives – min. 3 drives	RAIT 4 stripes data at the block level. It calculates and stores parity on dedicated drive. Good for high-performance sequential files. Write requests can suffer from dedicated parity-drive bottleneck.
RAID 5 **RAIT 5**	Block-interleaved striping with single striped parity-min. 3 drives	Identical to RAID 4 with added rotating parity protection. With a hot spare drive for RAID, data from a failed drive can be rebuilt to protect against a second drive failure. Balances data availability and read/write performance well, the most popular RAIT and RAID level.
RAID 6	Block-interleaved striping dual striped parity	Like RAID 5 with an added level of parity protection for highest availability and data protection.
RAID 10	Striped mirrors – min. 4 drives	Not one of the original RAID levels, multiple RAID 1 mirrors are created, and a RAID 0 stripe is created over these.
RAID 50	Striped RAID 5 – min. 4 drives	RAID 50 stripes RAID 5 groups like RAID 0. Costly option by doubling disks but ads throughput.
JBOD	Just a Bunch of Disks	Each drive functions independently, no redundancy is supported.

RAIT 0 stripes data blocks and is an ideal solution when the highest performance level is required but doesn't offer fault tolerance capabilities as it stripes data on two or more drives without using parity.

RAIT 1 is mirroring and writes data to two (or more) tape drives offering a high level of fault tolerance and uses no parity. If one drive fails, the second keeps running with no interruption of data availability. Mirroring requires two drives and twice as much media capacity making it the costliest RAIT option.

RAIT 4 uses block level striping with a dedicated tape drive to record parity. Data remains fully available if a drive fails. The single parity drive can become a bottleneck for write intensive and performance can be impacted.

RAIT 5 is the most popular RAIT option and uses block-level data striping with distributed parity written across all drives in the array stripe. The virtualized capacity of tape media is increased as at least three drives appear to the system as one virtual drive. RAIT 5 combines good performance, fault tolerance, high capacity, and storage efficiency.



RAIT Software Support

Amanda (Advanced Maryland Automatic Network Disk Archiver) is a widely-used Open Source Backup and Archiving software product. Amanda supports RAIT 4 (block striping, dedicated parity drive) and allows the system administrator of to set up a single master backup server to back up multiple hosts running in parallel to either a single large capacity tape or disk drive. Amanda can back up many workstations running multiple versions of Unix, Mac OS X, Linux and Windows operating systems.

BrightStor ARCserve Backup Tape RAID option is a widely-used commercial software product providing RAIT support. With the RAIT option operations are performed using multiple tape drives to simultaneously process data during backup and restore operations. Data is written across all the drives in the stripe in a predefined order. Each stripe appears on the ARCserve management console as a single virtual tape drive. When configuring a backup job, you select the group of tape drives as your RAIT destination and this group contains all the drives configured in that RAIT set. The ARCserve option supports RAIT levels 0, 1, and 5.

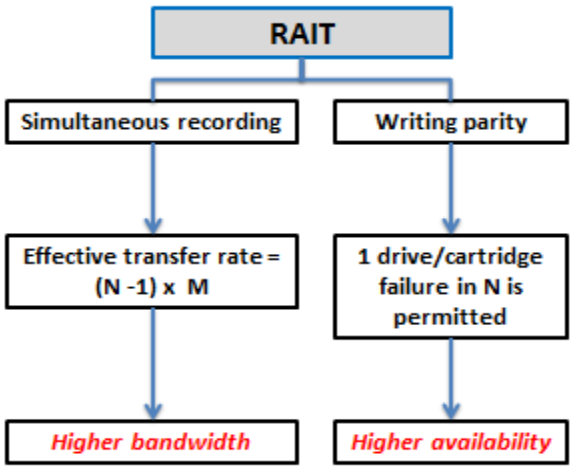
HPSS (High Performance Storage System) is widely used by HPC and supercomputer organizations. HPSS was developed as the result of collaboration among five Department of Energy (DOE) laboratories and IBM in the USA, and with significant contributions by universities and other laboratories worldwide. HPSS provides scalable hierarchical storage management (HSM), archive, and file system services to move large files between storage devices and parallel or clustered computers much faster than today's commercial software products. The HPSS system currently implements a multiple RAIT stripe solution (up to 16 tapes wide) that includes write recovery by mounting a new volume group, read recovery using multiple copies, and the ability to repack a volume group to reclaim empty space on a volume. Excellent case studies using tape and RAIT are available from the NCSA (National Center for Supercomputing Applications) Blue Waters Study and the Oak Ridge National Laboratory ORNL study. Since RAIT was implemented in January 2015, the ORNL center has placed more than 15 PB and 42,585,959 files on RAIT tapes.

RAIT Performance

Significantly higher throughput (data rate) and higher availability for large files are the key benefits of using a RAIT subsystem. RAIT can be used with or without parity and different drive types can be intermixed, however the RAIT subsystem performs at the speed of the slowest drive. RAIT stripes can range to 16 tape drives but presently 3-5 drives are most common implementation. RAIT performance, just like the performance of any other type of storage system, can be impacted by slow server speed, too few or too small buffers, blocksize, channel contention from other devices, or various software bottlenecks. The RAIT file is not inter-changeable if the external company doesn't have a RAIT system.

RAIT Provides Higher Bandwidth and Higher Availability

FUJIFILM



RAIT Effective Transfer Rate MB/sec.

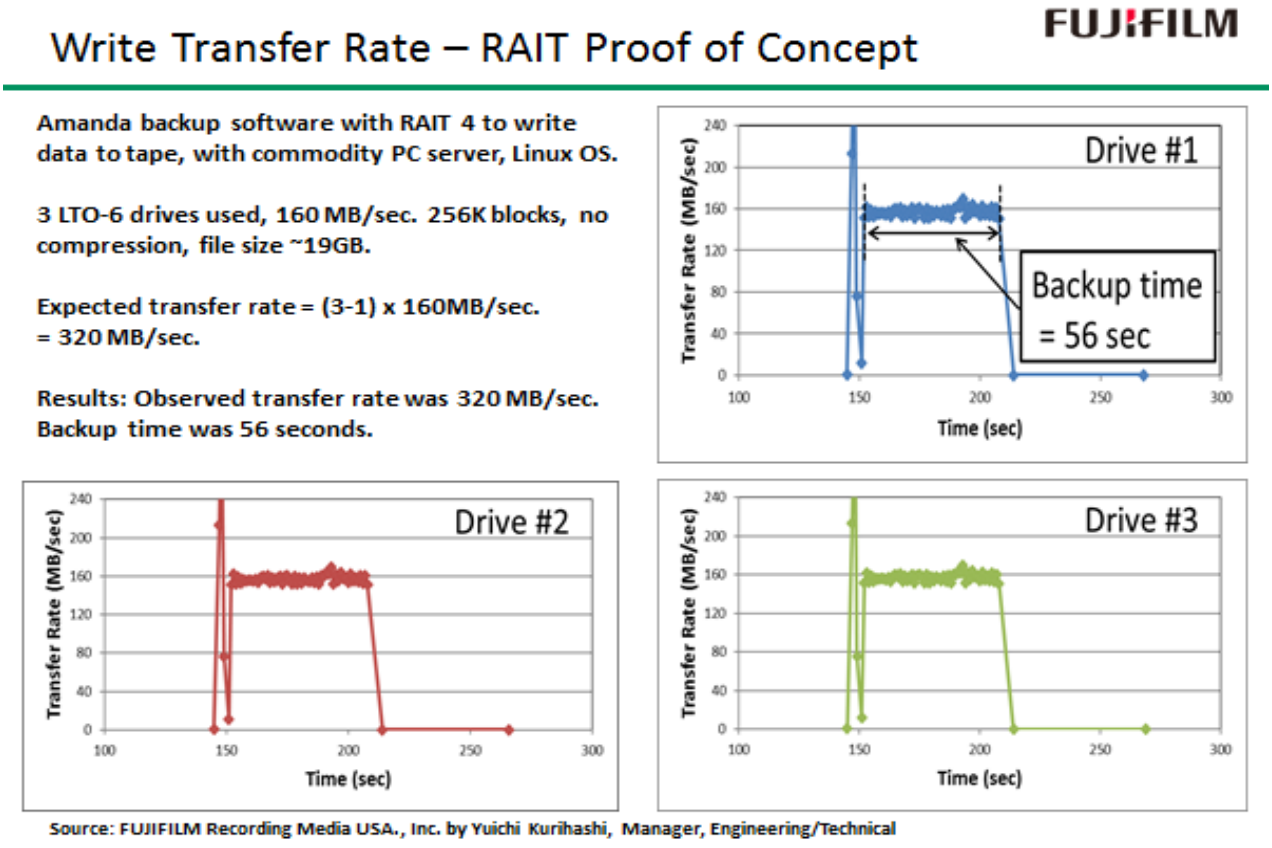
Drive Type	M	N=3 E=67%	N=4 E=75%	N=5 E=80%
LTO-6	160	320	480	640
LTO-7	300	600	900	1,200
TS1155	360	720	1,080	1,440

Effective Transfer Rate = (N-1) x M
E: Efficiency factor for RAIT group. E = (N-1)/N
M: Transfer rate of single drive
N: The number of drives in a RAIT group. Min. N =3, typically N <= 5

Laboratory Proof of Concept by Fujifilm

To verify that tape drives using a RAIT implementation can transfer at maximum drive data rates, proof of concept testing was conducted at FUJIFILM Recording Media USA., Inc. by Yuichi Kurihashi, Manager Engineering/Technical to measure the RAIT subsystem throughput. The steps used in the proof concept were:

1) Using the Amanda open source backup software which uses RAIT 4, the ~19 GB user data file was first backed-up (copied) from disk to a buffer-disk before writing to tape to prevent the update of the data during backup.
2) Then the file was backed up to a RAIT 4 configuration using three LTO-6 drives with a data rate of 160 MB/sec. taking 56 seconds.
3) The observed and expected data rate from the table above was 320 MB/sec. for the three drive RAIT configuration as maximum drive data rates were achieved.
4) Next the data was intentionally deleted from disk.
5) Then data was restored to disk from two tape cartridges, assuming one cartridge was not functional, to validate the RAIT data restore functionality using the parity drive.

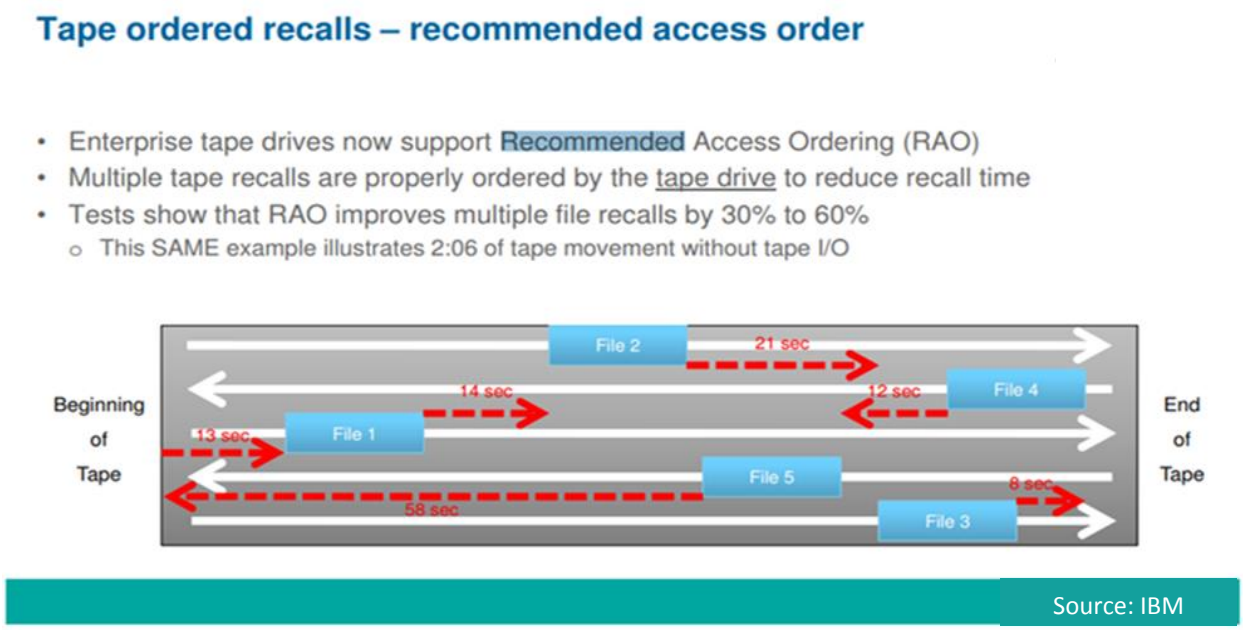


Note: Tests using four and five LTO-6 drives (charts not shown) were conducted with observed maximum write data rates of 480 MB/sec. and 640 MB/sec. respectively, further validating the RAIT throughput model while attaining maximum data rates.

Recommended Access Order (RAO) Improves Tape File Access Times

In addition to the increased throughput that RAIT provides for large tape files, another new capability is available for improving tape access times for smaller files called RAO (Recommended Access Order). RAO generates the best access order for randomly stored files on enterprise tape. Presently, tape files have been accessed (reading data) in the order that the data was written on the tape which is typically random. This can become inefficient as the files may be in different physical locations, in different wraps, in different servo bands, and on an opposite side of the tape centerline or in a different head travel direction. This inefficiency has been tolerated in the past, but as tape capacities and therefore the number files on a cartridge continue to increase, file access times can be expected to increase also.

The RAO determination method is adapted for sorting the list of files or data sets provided by the application or user for best or significantly improved performance using a relatively small amount of time to produce the reordered list of files. This optimized list is called "best access order" and provides the least amount of time that is needed to locate and read all files or data sets on a serpentine recorded tape. Native tape capacities have reached 15 TB and several laboratory tests have demonstrated areal densities capable of storing 100s of TBs on a single cartridge. In the future RAO will provide a highly valuable function for much faster file access as the number of files on a tape cartridge increase.

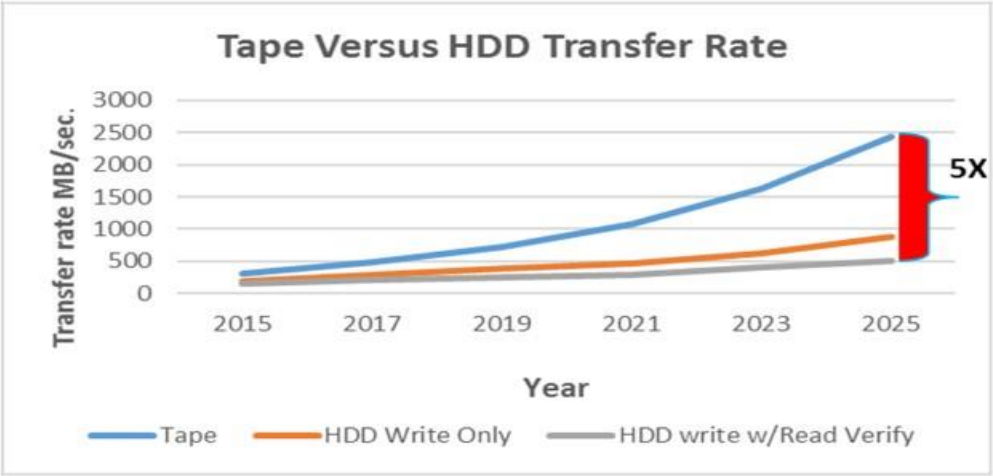


Tape Data Rates to Soar

Tape drive data rates are expected to be 5x faster than HDD data rates by 2025 (See INSIC data rate projection chart below). The dramatic transfer rate multiplier that RAIT provides, coupled with the much faster tape drive data rates uniquely positions RAIT for the HPC, hyperscale, cloud and the enterprise market. These businesses constantly push for higher throughput capability to move massive amounts of data between locations. For example, assume a RAIT array has 4 drives and each has a data rate 5x faster than a HDD, the RAIT array will have an effective transfer rate = (4-1) x 5 = 15 times greater than the HDD. This means it will take 15 times longer to transfer the same file from HDD as from the RAIT array.

Tape Data Rates Surpass HDD

Tape Data Rates to Exceed HDD by 5X



Source: TSC State of the Tape Industry Memo 2016

Summary

Tape performance is quickly improving. With RAIT, the Active Archive, RAO, LTFS and much higher data rate capabilities on the way, the tape industry is making significant strides in delivering much faster initial access times and throughput levels. For all the amazing technological progress made in the traditional data centers and with cloud computing, the fundamental challenges of reliably transferring large files and volumes of data at high speeds to locations around the world are rapidly increasing. This trend presents an enormous mass transit problem for digital data and a potential impediment to future growth across many segments of the IT industry pushing the storage industry to continually innovate.

RAIT is designed for those applications that require significantly higher bandwidth and availability including HPC, hyperscale, and cloud provider ingests and migrations. Many large backup and recovery, video surveillance, media and entertainment, hi-definition video, the IoT, Big Data and scientific research applications can directly benefit from RAIT. With autonomous cars now emerging, the need to transfer petabytes of real-world video, weather information and high-definition map data collected by vehicles in a few hours to remote development teams working on next-gen software for self-driving cars is quickly approaching. High bandwidth capability is also a critical factor for disaster recovery situations where massive amounts of data, if not the entire computing environment, need to be moved in a short period of time to bring a business to operational status again. Tomorrow's *mass storage* requirements will demand a *rapid transit* solution signaling that the stage is set for RAIT to address these issues.

End of report.