

*Technology Brief*

# NERSC: Proving Tape as Cost-Effective and Reliable Primary Data Storage

**Date:** December 2010 **Author:** Mark Peters, Senior Analyst

**Abstract:** *When is tape a direct access storage device? And when is a very active tape archive really just another tier of “regular” storage? At the NERSC Center in California, tape is simply seen as extremely efficient, cost-effective, scalable—and reliable—storage. With over 13 PB of data on tape (growing at around 60% p.a.) and decades of history, NERSC has the facts to prove tape’s capabilities; so much so that it does not employ additional copies.*

## Executive Summary

While many IT organizations eschew tape as either a necessary evil or something to be gradually wheeled out of their data centers, the truth is that tape is still a very viable element in data infrastructures worldwide. At NERSC, there is over 13 PB of data on tape: 30-40% of its tape activity is reads, it has a measured and proven reliability of 99.945%, and its \$/GB cost is around 5% of that of its disk storage.<sup>1</sup> NERSC’s story not only validates that general truth but enhances our understanding of the organizational and business value tape can deliver by using it aggressively and assiduously—not because other media isn’t affordable, certainly *not* in an unthinking “this is how we’ve always done things” manner, and definitely *not* just for backup or what might be called “stagnant archive.” Indeed, nothing could be further from the truth. NERSC has chosen its tape usage carefully, with an automated tape library as a very active archive; which by definition means that it is used for primary data. This is *primary* data, with no secondary copy and so much activity that it begs the semantic question of when does an active archive stop being an archive and simply become a part of everyday online storage infrastructure.

More than just a pragmatic decision, NERSC has definite proof that its use of tape is viable not only in terms of economy and performance, but also in terms of reliability; its tape media (whether fast access or high capacity variants) has been proven to be readable after 12+ years with 99.945% of its tapes being 100% readable. Jason Hick, Storage Systems Group Lead at NERSC, uses a wonderful analogy to put this level of reliability into perspective: NERSC’s data migration (from June 2009 to March 2010) involved reading 14,805,823 meters of tape, which is the distance between San Francisco and Perth; unreadable data resided in at least one block of 14 files. Those 14 files representing 108 meters of tape, approximately the length of the Boeing 777 one might use to fly between those two cities!

Hick feels that NERSC’s tape infrastructure is more of an investment than a cost, largely because of its extensive usable lifetime, minimal power and cooling needs, and very competitive cost per gigabyte (up to 20X less expensive than its disk alternatives). NERSC’s ability to use tape so heavily for primary data is partly a function of the scientific world it inhabits, but is at least as much a function of careful planning and proven reliability which combine with tape’s economic advantage to provide the most efficient approach for its mass storage—and IO—needs. The bottom line is that at NERSC, tape is a relevant and operational IO infrastructure, *not* a repository for dead and dying data.

## NERSC

NERSC stands for the National Energy Research Scientific Computing Center, which is managed by Lawrence Berkeley National Laboratory. It is the Department of Energy’s Office of Science national user facility. As such, it focuses its open computing on a broad range of projects: investigating climate change, renewable energy sources, environmental

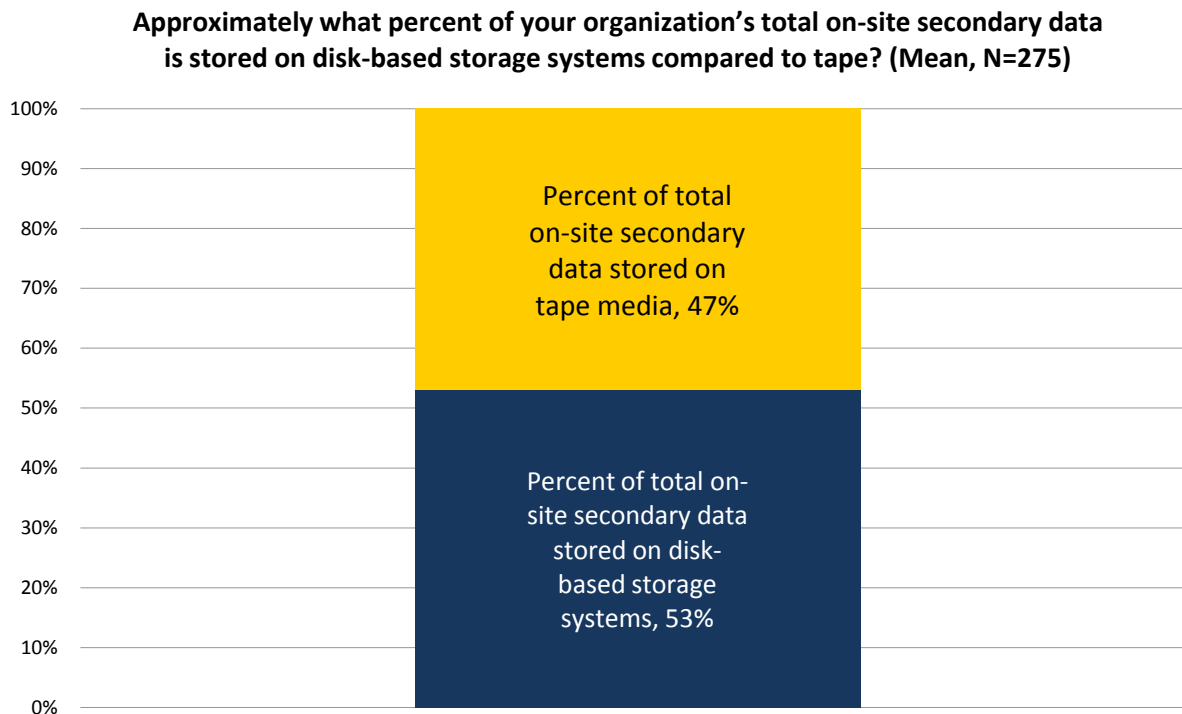
<sup>1</sup> SOURCE NOTE: This paper was developed following direct interviews in late 2010 with Jason Hick, the Storage Systems Group Lead at NERSC. His group supports a “desktop to petaflop” environment that needs to be both highly efficient and able to cope with exponential growth. Thanks to Jason for his involvement, insights, and willingness to share his approach and thoughts with others both via this case study itself and also via the quotes that appear throughout.

modeling, chemistry, biology, physics, nanotechnologies, fusion, astronomy (with direct feeds from NASA satellites), and other scientific areas. Funded by the Department of Energy’s Advanced Scientific Computing Research Office, NERSC is connected to some of the most powerful experimental facilities in the world through the “ESnet” network, also managed by Berkeley Lab and tailored to moving large scientific data sets around so they can be analyzed and stored at centers like NERSC. NERSC has a direct mandate to be hyper-efficient and mature to be “the best run HPC center in the world.” To that end, its staff of just 60 handles roughly 4,000 users working on 400 distinct projects. About half of these projects are active in any given week with around 400 users daily. The user community at NERSC is one of the largest in the center business and very broad in its scientific mission, with uses ranging from some of the largest computational problems to data-intensive science applications. From an IT perspective, this creates a “real mixed bag” with NERSC needing to remain flexible and efficient. For Hick, the challenges were obvious (“For ages there was the disk fan club and the tape fan club, with the former seeing tape as just a necessary evil”), but so was the objective: “To us, it’s just about delivering cycles and storage to all our users in the best and most efficient way possible.”

### Tape in IT

While it has been common for decades to talk about tape being dead, the truth is, regardless of its popularity and relative difficulty to use well, its demise is no more true than the death of paper in the office or the music of Mozart and the Beatles. Whereas tape for decades was synonymous with backup, it is now being used more for DR (with a good proportion of data and applications being backed up to disk initially and then to tape) and is growing in importance for archive applications especially as the need to archive expanding volumes of data increases exponentially. Overall, for these types of “secondary” data (although archive is, by definition, paradoxically a primary copy!) where tape’s economics are typically attractive, it continues to command a large market share. Recent ESG research<sup>2</sup> (see Figure 1) shows that tape is pretty much equal to disk (47% vs. 53%) in this respect. While this may be indeed be a relative decline for tape over the last decade or two, it remains absolutely a significant media and a significant market, much to the chagrin of those vendors that would like to write its obituary.

Figure 1. Secondary Data Stored on Disk vs. Tape

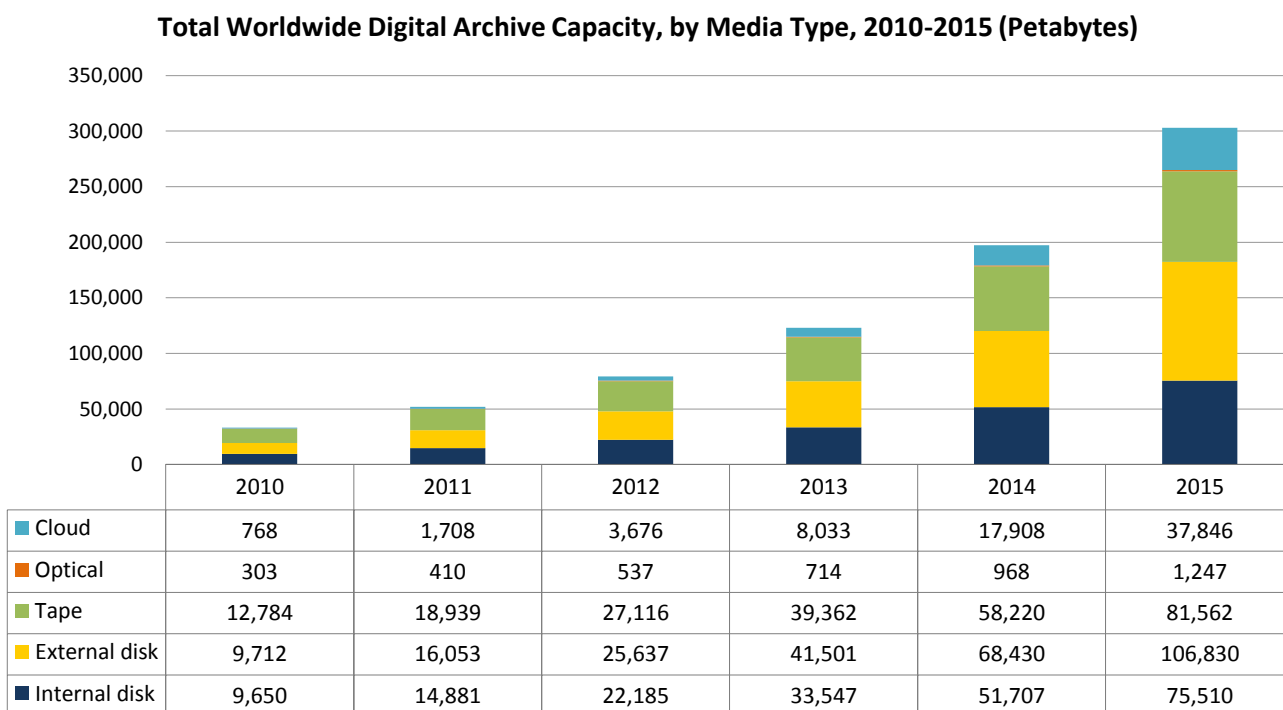


Source: Enterprise Strategy Group, 2010.

<sup>2</sup> Source: ESG Research Report, [2010 Data Protection Trends](#), April 2010.

As opposed to secondary data (which will typically be data that is of lower value and/or use) or backup data (which is a copy of data in case of need), *archive data* is the primary copy of data. Think of the “stacks” in a library basement or annex as compared to the “live” shelves of the main lending area. The operational delicacy and care, whether it’s physical data or electronic volumes, requires great balance as the material is important yet the access is (typically) low. [NERSC has a particular challenge inasmuch as its drive to efficiency has led it to implement earlier movement to an archive tier than might usually be the case, combined with a scientific focus that demands far more frequent and sizeable access requests to that archive than usual]. Whatever the level of activity of the archived data, this much is clear: the volumes of digital archives to be stored are already enormous, are forecast to grow significantly, and tape will remain a significant media for such digital archives. ESG’s research<sup>3</sup> (shown in Figure 2) gives an idea of the scale of things; while tape’s share of the overall digital archive capacity is expected to drop from 38% this year to 27% by 2015, it still represents a more than six-fold increase in petabytes stored over that period.

Figure 2. *Worldwide Digital Archive Capacity, by Media Type, 2010-2015*



Source: Enterprise Strategy Group, 2010.

What constitutes an archive is an interesting semantic question. At what percentage of reference activity should it be called something else? As a rule of thumb, most archives would only ever read 5% of the data contained therein per year; beyond some read percentage point (10%? 15%?), it would make sense to use a different name, perhaps a “distribution archive” or, more prosaically (given that NERSC reads 30%-40% of its “archive data”), perhaps it should just be seen as a regular IO tier.

### The NERSC Tape Environment

With petaflops of processing power (NERSC has both the 5<sup>th</sup> and 17<sup>th</sup> fastest computers in the world as measured by the “top 500” list) and some 15 PB of data (including some that dates to the 1970s) NERSC is clearly a significant operation. It operates Lustre for local scratch and IBM’s GPFS (2-3 PB each) as its own global file system that enables all of its users to get their data from any of the systems in the center. Tape accounts for around 90% of storage capacity, with 30% of tape IO being reads! Where most tape archives would be 95% or more write operations, NERSC’s tape environment is optimized for its high read percentage. To make matters more challenging, the read activities are pretty random, driving

<sup>3</sup> Source: ESG Research Report, [Digital Archive Market Forecast 2010-2015](#), July 2010.

a need for fast tape exchanges and plenty of horsepower in terms of EPH (exchanges per hour). A standard “time/recall likelihood” curve doesn’t apply to NERSC’s user community as scientists “just need stuff” (for example, perhaps there’s a new supernova signature which will require looking back at all their data in case they missed something).

To summarize some of the main statistics:

- NERSC’s archive handles 30-40% reads, so it needs a mix of access- and capacity-oriented drives.
- 50-70% data growth per year (with resources of all sorts constrained, NERSC relies on technology improvements—especially drive capacities and media reuse to address this challenge).
- Average tape exchange is 60 per hour (10,000 per week), average concurrent transfers is 15.
- A strong correlation exists between the capacity of main system memory and archive capacity growth per year. As of 2009, each 1 TB of main system memory was yielding 35 TB of archive data per year.
- On average, newly ingested data remains on disk cache for five days before it is purged.
- On average, NERSC grows by 1 million new files per month; while 70% of new files are less than 2 MB, 50% of the total new data created is from files greater than 6 MB in size.
- NERSC operates an Oracle (StorageTek brand) tape environment of 128 tape drives. Thirty-five of these are T9840Ds (a fast access device) and 93 are T10000Bs (high capacity).
  - The T9840D fast access tape drive averages 30 seconds to first byte and handles 83% of NERSC’s *files* (<110 MB each).
  - The T10000B capacity tape drive averages 1.5 minutes to first byte and handles 94% of NERSC’s *data* (>110 MB each).
- The tape drives are operated within four SL8500 automated tape libraries with pass-thru capabilities (which are important to effectively make the tape one large pool. The pass-thru doesn’t help on time-to-first-byte of data, but ensures data can still be served and thereby manages “tape hot spots.”).

## Planning & Use

Hick spends a lot of time marketing and explaining the value of the tape environment to his users: “Our goal is to make tape work ... almost like a file system.” The main provisioning management tool is called an SRU (Storage Resource Unit), a theoretical request that allows Hick and his team to track and account needs and lets users see their burn rate: “quotas set up both sides for success.”

Every aspect of the system is fanatically measured, monitored, and managed to squeeze economic value and media longevity. It is this in-depth planning and analysis to which Hick attributes the success with primary, active data on tape.

## Results

### Reliability

Although Hick’s understated comment (“in the end, it’s surprisingly reliable”) sounds anecdotal, it is based on really thorough measured experience during which all the data from an entire decade was read. To migrate to the new tape technologies (T9840D and T10000B) in 2009-2010, NERSC read back nearly 24,000 cartridges, as follows:

- 6,859 T10000A (up to two years old)
- 9,155 T9940B (up to eight years old)
- 7,806 T9840A (up to twelve years old)

There were only 13 tapes containing some data that could not be read: in other words, 99.945% of tapes were 100% readable. On those 13 tapes, much of the cartridge *was* read successfully: only 14 files had real issues and these represented just under 100 GB of data. Indeed, the unreadable data within those files was normally in one or two blocks (250-500MB) of data with the remainder of the file readable. This level of reliability is more than adequate at NERSC, so it will remain with its single copy of the majority of its data on tape. Hick said: “Even the tape vendors say we should consider dual copy ... they say we’re crazy, but our stats don’t show that.”

## Financial Value

Given all the focus on efficiency at NERSC, it is no surprise to discover that the financial value of its primary data tape archive is very compelling. In a comparison of absolute costs between its tape and disk infrastructures, NERSC's tape file system (HPSS) has a \$/GB that is almost 20X less than that of the disk file system (GPFS). And in a relative cost perspective, NERSC's significant capital investment in 2008 (for three SL8500s and 88 new tape drives) actually managed to drive its 2009 \$/GB costs below those of 2007. There are some pragmatic approaches that Hick recommends in order to optimize the financial aspects:

- Early adoption of new larger tape capacities provides immediate operational savings.
- Media reuse also provides savings in terms of avoiding the purchase of new media and because rewrites are usually at least twice the capacity of the previous records.

Overall, Hick believes that for NERSC, *tape drives and libraries are an investment rather than a cost*, especially with the libraries and drives typically lasting as long as vendors support them (20 years and 10 years, respectively).

## Other Considerations

While this case study is, not surprisingly, a largely positive account, there are still things that Hick would like to see. First and foremost, he would welcome a more competitive tape marketplace: at the enterprise and HPC end of the market, there are only a handful of vendor choices for tape libraries and even less when it comes to the media and drives. When committing for such a long period of time to any direction, Hick would like more choice. As a corollary to this, he would also like to see a more organized and concerted effort by the tape industry as a whole to both grow the market and share awareness and opportunities for tape. Simply put, a rising tide lifts all boats and might serve to dispel some of the myths that surround tape. For instance, he would like to see less reliance on disk head technology in the tape world and more specialist monitoring and statistical analysis tools for tape. To be fair, Oracle has not given NERSC any reason to change suppliers, but Hick is always monitoring his options; he is committed to tape, but not to a particular vendor.

## The Bigger Truth

Some users might think this case study has been derived from a "Ripley's Believe it or Not" story. After all, received opinion in many quarters is that tape is something to be avoided and here is a user embracing it. This is diametrically opposed to most users' views of tape: most organizations are set on mitigating the negative attributes of tape whereas NERSC is accentuating the positives. It sees tape as an investment and not just a cost and uses its tape as a dynamic part of its tired IO infrastructure rather than as a collection bin. If this were all just a *matter of opinion*, it would not represent a truly useful case study; what makes it interesting is that NERSC's use of, and commitment to, tape is a *matter of fact*. It has measured and verified its approach from both a business and a financial perspective as well as in terms of operations and reliability.

Of course, there are some things that make NERSC a little different from other users: deliberately avoiding spending massive amounts of money to preclude data loss. That said, NERSC isn't *that* different inasmuch as it has made a straightforward cost-benefit business decision based on providing optimum efficiency to its users. And in so doing, it has debunked the ideas that tape is inherently unreliable and that disk and tape share similar long term TCOs. Perhaps most telling of all is that NERSC's approach is by design and not by default, whereas many users see tape as just something they have to have. As Jason Hick said in summary, "I couldn't afford to put all this on disk, but neither would I want to."