

## Spectra Goes Big (Data)

Addison Snell

November 2011

*White paper*

### MARKET DYNAMICS

#### How Big Is Your Data?

The term “Big Data” sounds simple enough. We know what each of those words means separately, and putting them together isn’t that difficult. What is it? Data. What sort of data? Big. At the title-slide level, it’s easy to understand. The tricky parts are in the details – not only how to learn from the data, but how to store and access it as well.

There are many types of data that comprise the Big Data challenge. Traditionally, computing dealt with four types of data: integer (whole numbers), floating-point (numbers with fractions), Boolean (binary variables like zero/one, true/false, on/off, yes/no), and strings (text or other non-numerical values). Now an increasing amount of data does not fit neatly into such categories, including images, audio, and video, which are difficult to analyze for content. (Imagine you are given a task to filter years of college football footage to find plays in which the offense passed for a touchdown of over twenty yards against a safety blitz. Unless you are lucky enough to find the search terms as unlikely metadata tags, you have a lot of viewing in front of you.)

Furthermore data is created in more ways by more types of devices than before. Scanners, sensors, satellites, medical devices, and ticker feeds provide constant streams of digital information far beyond what had ever been available before. Some data analysis is done in real time – the sort of analysis for which there is no such thing as a late correct answer; timeliness is part of the solution. Some involve sharing among broadly distributed teams; some are intensely private; some are a mix of both, shared among a privileged few.

Intersect360 Research looks at different categories of Big Data applications, including real-time analytics, enterprise analytics, research analytics, complex event processing, data mining, and visualization. The difficulties in managing the growing data requirements of these applications – the “big” in Big Data – demand many of the same technologies that serve High Performance Computing (HPC) applications. Indeed, there is a great deal of overlap between Big Data and HPC, in areas such as financial services, genomics, entertainment, health care, medical research, and national defense.

And while much of the discussion of Big Data is rightfully focused on the analytics algorithms that are applied to data in order to derive insights, there are practical aspects involved in the bigness itself. All of this data needs to be stored somewhere, and often for a long time, with a high degree of confidence that the data can be retrieved when needed. Gigabytes seem increasingly quaint, and previously obscure prefixes (Tera, Peta, Exa, ...) each become mainstream in turn. (Wait a few years. Exabytes will become boring as we move on to Zetta and Yotta.)

Here again we see an overlap between HPC and Big Data, in the importance of digital tape storage media to these environments. Rumors of tape’s demise as a technology aren’t merely premature; they are simply incorrect. Tape generally offers not only better capacity than disk, with more bytes per dollar, per watt, and per

floor tile, but also longer data life, a critical metric of data archives. Tape has also been boosted by recent technology enhancements, including LTFS (Linear Tape File System)<sup>1</sup> and the emergence of “Active Archives,” in which storage administrators can view and search their data on tape through traditional file system interfaces as network attached storage. Approximately two-thirds of organizations that run HPC applications use tape for at least part of their data storage, a figure that is highly correlated to organizational size. That is, once an organization’s storage requirements grow large enough, tape tends to enter the storage mix.

### New Announcements from Spectra Logic

In a series of announcements this fall, Spectra Logic reaffirmed its commitment to continue to deliver scalable tape archive solutions for HPC and Big Data applications. This is significant in an industry experiencing dwindling tape competition, most notably with the decreasing role of StorageTek tape libraries. While Oracle (which purchased Sun Microsystems, which purchased StorageTek) does not prominently feature its tape solutions, Spectra is clear on the opportunities tape continues to offer, and the company is continuing to advance.

If there is a consistent theme in the Spectra announcements, it is scalability. Reliability, performance, and density all play their roles, but each of these can be viewed in the context of how they support exascale systems. To store exabytes of data efficiently, an archive must be reliable enough not to lose data, fast enough to retrieve data within usable timeframes, and dense enough to be contained in the datacenter.

Highlights of the Spectra announcements include:

- CarbideClean and Certified Media Solutions:** Tape media itself has a long lifespan. How long depends on access and usage patterns, but typically 15 to 30 years, which can be much longer than disk, which can experience magnetic degradation in less than ten years. In exascale storage systems, the tape drive read/write heads are subject to wear over extended periods of time from excessive cleaning or overly abrasive tape. Spectra’s trademarked CarbideClean technology is a sort of “high octane” option to provide burnished tapes, with less microscopic debris, thereby extending the life of the drive by up to 2x, according to Spectra. Targeting Big Data arenas in which archived data need to last decades, CarbideClean extends the life of tape heads in exascale libraries. Spectra is now offering this new feature for its certified media to enterprise data centers, incorporating CarbideClean and other RAS (reliability, availability, and serviceability) features.
- Media Lifecycle Management (MLM):** With data meant to last so long in an archive (many users would express the desire for their data to be permanent), media lifecycle management is an important feature for enterprise reliability. With red, yellow, and green icons indicating tapes’ health, MLM gives users a simple view of any tapes that may have been overused, giving them a chance to migrate data off these tapes to retire them, without loss of data.



Colored icons indicate tape health with Spectra’s Media Lifecycle Management.

Picture source: Spectra Logic, 2011

<sup>1</sup> <http://www.lto.org/technology/lfts.html>

Available on its LTO-based libraries since 2008, Spectra is now offering MLM on its TS1140 drive-based T-Series line of tape libraries.

- **Spectra Service PriceLock:** To help enterprises plan for long-life archives, Spectra has introduced guaranteed future pricing for service contracts on all T-Series tape libraries in the U.S. Service contracts for new customers are locked in at 2008 MSRP levels, subject only to an annual adjustment for inflation.
- **BlueScale 12 management software:** Spectra announced performance, management, and integration enhancements to its BlueScale environment, which runs across all Spectra T-Series tape libraries.
- **3.6 Exabyte T-Finity:** With improvements in density, Spectra has extended its highest-end scalability, entering the Exabyte realm for the first time. The T-Finity line is now capable of scaling to 3.6 Exabytes, with up to 400,000 slots in each library complex (eight libraries per complex; 40 frames in each library).

## INTERSECT360 RESEARCH ANALYSIS

Big Data and HPC have several areas of overlap, both in the magnitude of their computational and data management challenges and in the technology choices organizations must make in order to address them. For archiving (and still accessing) petabytes to exabytes of data, for long periods of time, tape will continue to be an important datacenter component.

With this flight of announcements, Spectra has established that it will continue to innovate in large-scale tape libraries. What is particularly noteworthy is the increasing role of reliability and manageability features. These would be necessary in any case for achieving customer-usable exascale storage, even in research environments, but the inclusion of enterprise RAS features will allow Spectra to address Big Data application areas in commercial accounts beyond its traditional HPC and media markets.

As Big Data areas such as enterprise analytics and data mining take enterprises into new levels of scale and data management, these organizations will need to evaluate how they will manage archives with near-term planning horizons in the exabytes. Tape was already a leading technology for high-capacity archiving. Spectra's enhancements aim to bring tape to exascale storage and Big Data by making tape more user-friendly as well.