SNIA Activities in Big Data and Big Data Benchmarking

Alan Yoder
May 2012

The Storage Networking Industry Association (SNIA) is a non-profit trade association (501 (c) 6) representing all the major players in the global enterprise storage market. It has both marketing and technical concentrations. Given that it is an industry association under US antitrust constraints, it cannot easily lead R&D in IP-sensitive areas. This has caused it to focus on standardization efforts in areas where multiple companies already have products. Recently, however, it has been able to make contributions in emerging technology areas such as cloud storage, solid state benchmarking and green storage. Its CDMI specification (Cloud Data Management Interface) is in process to become an ISO standard and its Operational Power spec is a key input into the US EPA's effort to develop an ENERGY STAR program for business- and enterprise-class storage. The Solid State Storage Initiative's Performance Test Specification has helped the emerging solid state storage industry settle on standard ways of characterizing the performance of SSDs and other flash storage products.

The SNIA charges fees commensurate with company size, type and product range. Universities, non-profit organizations and individuals can participate for nominal fees ($300 per annum). All members are eligible to participate in all technical activities and symposia free of further charge, subject to SNIA IP policy constraints. The organization maintains a technology center, currently in Colorado Springs, which is able to host and maintain equipment necessary for testing and experimentation for ongoing projects. The IOTTA group, for example, headed up by Geoff Kuening of Harvey Mudd University, uses it as their staging point for a large repository of I/O traces.

The SNIA has recently established an Analytics and Big Data Committee. Committees in SNIA are no-fee activity areas in which member companies can do collaborative work. The committee has a primarily marketing-oriented focus at present, but welcomes input on potential collaborative technical efforts.

The SNIA is particularly interested in the power and storage aspects of big data computing. It is common for big data systems such as Hadoop to use triple replication as a resiliency measure and possible contributor to increased throughput. The power implications of this are impressive. Storage manufacturing companies in the enterprise storage space are using every device in their arsenal to reduce the amount of space and power that is required to store, protect, backup and archive data while maintaining compliance and disaster preparedness. Given the power implications of the architecture of Hadoop and cousins, we are somewhat mystified by claims the these systems "don't need RAID." The Lustre-based Sequoia system being built at LLNL, on the other hand, uses RAID 6 storage on enterprise-class JBODs. This seems more likely to have acceptable data storage power consumption.

During the course of its researches, the SNIA Green TWG (Technical Working Group) found to its surprise that for enterprise class storage gear, idle power approaches and can even exceed active power, and rarely is less than 85% of active power. This is in part because a disk rotor uses much more power than the arm, but also because that class of storage detects periods of idle activity and uses it for disk housekeeping activities like RAID scrubbing, which is read intensive.

Among the questions SNIA would find interesting are:

1. What is a normal ratio of server, networking and storage power in big data deployments? It would seem that with current architectures it will remain relatively constant under quite different "workloads". Is this true?
2. What ratio of disk space to processor power is optimal for various loads?
3. Does the concept of a "working set" apply to big data workloads? If so, can it be characterized with respect to various workloads?
4. Does the concept of storage tiering apply to big datasets? What percentage of the data can be stored on less expensive, slower storage?
5. Are tape technologies such as LTFS applicable to big data problems? If so, how, and under what circumstances? (related to question 4).
6. Does big data have the same availability, resiliency, compliance and disaster preparedness requirements that pertain to enterprise class data? Most research to date seems predicated on the assumption that it does not. Is this a viable assumption going forward, as big data hits analysis hits the mainstream and becomes mission critical?
7. Depending on who is quoting it, between 50% and 90% of all enterprise and business data is never read again once a week passes after its initial creation. This makes idle power management a critical feature of any data center power strategy. Does big data participate in this dynamic at all? Has any work been done on big data lifecycle analysis?

The SNIA recognizes that storage is simultaneously complex and not currently the highest consumer of power in a big data deployment. It may be the case that these questions do not get the highest priority ranking for researchers. We nevertheless look forward to activities aimed at answering them, both internally and in the wider research community.

Alan G. Yoder, Ph.D.

SNIA Technical Council

Governing Board, SNIA Green Storage Inititative