Establishing a Robust Data Collaborative to Enable Distributed Research

David Fellinger

DataDirect Networks, Inc, Chatsworth California, USA, dfellinger@ddn.com

Establishing a Robust Data Collaborative

The traditional method for storing research data has been to establish a high reliability file system with rotating media backed by a tape archive in a data center. Techniques such as parity protection are generally used to guard against data loss in case of a component failure. While these architectures can be constructed with a very high probability of data retention (99.999%) there is always the possibility of a disaster affecting the site yielding catastrophic data loss. The method for mitigating disaster risk is generally some sort of remote mirror where a second site is configured to contain the identical data of the first. This often leads to concerns and decisions regarding synchronous versus asynchronous replication to the mirror site which opens time domain concerns regarding locking and the stability of data. As data sets grow, management of the entire infrastructure becomes more complex and assurances must be put into place to guarantee the consistency of both the data and metadata in both geographies. The file systems that are generally used for this purpose are extents based and POSIX compliant utilizing file allocation tables to facilitate I-node lookup. These traditional block based file systems have been designed for data modification and have complex mechanisms for gathering unused blocks to be placed into an I-node if a file is extended or modified. The mirror is maintained by servers that read data from one location and write it to the mirror site.

A potentially unintended consequence of a mirror is that the second site may be electrically closer to a set of users in that geography. This offers load balancing as well since interesting data may be accessed in two rather than one site at increased bandwidth. While this makes the reading of data more efficient it can make the writing of data very inefficient since a mirror infers identical data and all system writes must utilize locking mechanisms to ensure that the same files are not being modified in both locations at the same time. These locks must be in place even though there may never be a circumstance where two users are attempting to modify the same mirrored file.

A new approach to an old problem

A very high percentage of research data is immutable and, in fact, is “write once read many” (WORM). The traditional file systems, however, are complex structures designed with locking mechanisms, file allocation tables, and block reclamation systems designed for the purpose of modifying files. It is much more efficient from a viewpoint of protocol layers and data placement to build a file system specifically designed for WORM data. Files can be incorporated with relevant metadata and the combination becomes an object. An object based data placement system can simplify the placement of immutable objects on rotating media. Objects are placed in containers with other objects of the same size and are stored in physically contiguous structures.

An object based system can incorporate the notion of zones such that a storage cluster can be comprised of multiple nodes in multiple geographic zones coupled by a WAN infrastructure. Objects have associated policies which dictate data placement in zones such that multiple instances of objects can be automatically replicated and retained. This replication is node to node eliminating the need for any server outside of the cluster. The end result is a data placement system that is not encumbered by locking mechanisms and file allocation tables. Objects are assigned object identifiers (OID) which define the placement of each object regardless of zone. These OIDs are retained as attributes in an external data base completely eliminating the need for I-nodes.

Objects placed in any zone can be replicated by policy to other zones for the purpose of disaster recovery, load balancing, and access proximity.

A data storage system that relies solely on replication for redundancy requires multiple copies of the same data. In other words, an object may have to be replicated in three zones for high redundancy requiring three times the storage that would normally be used to store the object in a single location. An object based system allows efficient use of storage space if high speed network connections exist between zones. A function has been developed which calculates a redundant erasure code for an object then separates the object into pieces and parity to be distributed over the zones of a storage cluster. No single zone contains the complete object, however, the redundancy of the system ensures that in a three zone system, as an example, the loss of a single zone does not impact the accessibility of an object. Further redundancy is provided within zones such that the loss of a node or rotating media can be recovered locally. A Markov chain model simulation of a three zone system has been developed based on this function which

proves that probability of data loss is less than 5.9x10^-12 only using 1.87 times the space which would be normally required to store an object. This low probability coupled with the geographic distribution of the data ensures unprecedented reliability for storing critical data.
An accompanying data grid allows policies which can maintain replication of heavily utilized objects for collaboration over geographies but which can be modified to reduce space utilization based on need.
Data grids can be constructed that assure the distribution of data regardless of the initial zone of placement with an extremely high degree of reliability.

ABOUT THE AUTHOR

Dave Fellinger is Chief Scientist, Strategy & Technology at DataDirect Networks. Dave has over three decades of engineering experience, including film systems, ASIC design and development, GaAs semiconductor manufacture, RAID and storage systems, and video processing devices, and has architected high-performance storage systems for the world's fastest supercomputers.
He attended Carnegie-Mellon University and holds patents in optics, motion control, video processing, file system technology, and pattern recognition.

In his role as chief scientist of DDN, Dave guides the company's product and market strategy to resolve key customer challenges at acute levels of scalability and has been instrumental in establishing DDN as a leader of the Big Data and Cloud eras. He serves on the board of the iRODS Consortium and the External Advisory Board of the DataNet Federation Consortium.