Adapting RAID Methods for Use in Object Storage Systems

David Bigelow, Scott A. Brandt, Carlos Maltzahn, Sage Weil
Computer Science Department, University of California, Santa Cruz
{dbigelow, scott, carlosm, sage}@cs.ucsc.edu

Object-based storage systems are currently being investigated as promising possibilities for future large-scale storage needs. The 'Ceph' storage system [2] has been designed with the goal of scalable petabyte-level storage, but currently uses mirroring as its primary means of reliability. While the mirroring of objects has a certain elegance for simplicity of design and understanding, it is extremely inefficient in terms of hardware and economic overhead required to achieve even a basic degree of protection. Work is therefore progressing on adapting RAID-like methods for use in an object-based environment using Ceph as a testbed.

There is no direct mapping of the standard parity-based error correction codes into an object-based environment. Parity-based file striping has been previously implemented in network file systems, such as Zebra [1], but are unfortunately not applicable in thes circumstances. While the same mathematical principles apply, data striping no longer applies on a sequential file basis, only on an object basis. As a natural consequence of this design shift, we need no longer concern ourselves with using RAID-like methods to increase performance and throughput; there are far more efficient ways of manipulating those values by tweaking the parameters of object size and construction. This allows us to focus entirely upon using parity-based error correction codes for data reliability only. Initial efforts are focused upon RAID-4 like behavior (one parity device per n data devices), but these same principles can be extended to other error-correcting codes, such as the RAID-6, which uses two parity devices, or even the Reed-Solomon methods, which have traditionally been thought of as too slow for use in a high-performance storage system, but which may find new opportunities in the realm of object-based storage.

Analysis and simulation work have yielded several interesting results, including one with significant implications for the design of any RAID-like system. According to our simulations, there is a significant performance decrease when the system is forced to operate in degraded more (one device of a group has failed, thus requiring parity calculations to reconstruct data) – a difference which is greater in magnitude than the corresponding difference in a simpler single-system based RAID scheme. This difference, likely arising from a combination of network delay, object size, and system load, is significant when applied to a system involving thousands of commodity disks, and when multiple device failures can be expected each week. This implies that the performance of the system in degraded mode operation is of particular importance, since the system can be expected to be operating partially in this mode for significant portions of time.

We have developed an initial implementation model which takes steps to minimize the performance penalty of degraded mode operation, albeit at the expense of slightly lowered peak performance. It is our thought that being able to guarantee a near-constant performance level is of more use than an irregular performance curve with better peak performance but periods of extremely poor performance which can neither be predicted nor easily mitigated. We also note that the object-based storage system paradigm allows us to improve throughput and response times through other means, and thus performance and reliability can be mostly decoupled from each other.

It is expected that this work will yield a system which is much more economical than the current mirroring system. While we must admit that better performance can be had by the simple expedient of throwing excess money and hardware at the problem for mirroring, we also believe that parity-based reliability schemes can provide equal and better dependability at a fraction of the hardware overhead of current schemes. It is expected that investigation along current lines will lead directly into being able to make certain guarantees for quality of service with high levels of confidence.

References

[1] Hartman, J. H., and Ousterhout, J. K. The zebra striped network file system. Tech. rep., Berkeley, CA, USA, 1993.

[2] Weil, S., Brandt, S. A., Miller, E. L., Long, D. D. E., and Maltzahn, C. Ceph: A scalable, high-performance distributed file system. In *OSDI '06: Proceedings of the 7th Conference on Operating Systems Design and Implementation* (2006).