Campaign Storage: Erasure Coding With GPUs

Walker Haddock
University of Alabama at Birmingham
Computer Science
Birmingham, AL, USA

Matthew L. Curry
Center for Computing Research
Sandia National Laboratories
Albuquerque, NM, USA

Purushotham Bangalore
University of Alabama at Birmingham
Computer Science
Birmingham, AL, USA

Anthony Skjellum
University of Tennessee Chattanooga
SimCenter
Chattanooga, TN, USA

ACM Reference Format:
Walker Haddock, Matthew L. Curry, Purushotham Bangalore, and Anthony Skjellum. 2017. Campaign Storage: Erasure Coding With GPUs. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

Exascale computing not only provides the capability to solve new problems but to perform finer grained analysis on current problems. These new capabilities mean more data products to be shared and stored. Magnetic tape will likely remain the lowest cost media for long term archive storage for High Performance Computing (HPC) for at least the next decade but the bandwidth and latency requirements for research are still too great for this media to be useful in reducing the demand for low latency disk storage for Exascale computing.

Cloud computing has developed high capacity, reliable and economical storage systems based on object technology. Los Alamos National Labs (LANL) has designed the storage systems for Trinity to include a cloud type object storage system as a layer between the Parallel File System (PFS) and the Archive. This pre-archive system has been dubbed "Campaign Storage" because the purpose is to store data products to be quickly accessible during the life of a research project. Data stored on the Campaign storage can be pulled back in to the PFS and staged for compute jobs or moved on out to Archive once the data has been curated. Campaign storage can reduce the capacity requirements for PFS storage by providing large capacity for storing data products from PFS and reducing the time that data must remain in the higher cost storage layer.

Our research makes the following contributions to the pre-archive storage layer: GPU assisted erasure coding, demonstrating erasure coding on File Transfer Agents, reducing erasure recovery costs with "Lazy Erasure Repair", and enabling larger erasure coded disk pools.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 BACKGROUND

In previous work, we designed and reduced to practice a library that performs erasure coding on GPU hardware, Gibraltar [3, 4]; this approach can further lower the price/performance for storage systems and provide opportunities for performing compute close to the data. As common disk drive storage capacities have increased from 750 GB in 2006 [8] to 10 TB in 2016 [15], HPC datacenters no longer get bandwidth for free when buying disks for capacity [9]. The use of these high capacity disks are best suited for low cost near-archive applications such as the Campaign Storage system used in the Trinity supercomputer at Los Alamos.

LANL has published the requirements for the Campaign Storage of Trinity [12]. The Campaign Storage should have about 25 PB capacity with future expansion capability. The bandwidth should be between 20 to 25 GB/s, which should increase with capacity. The files stored in the campaign storage system will not be updated in place. The system should use archive-grade hard disk drives, and gain performance through large scale parallel access.

The current production configuration of the campaign storage system in the Trinity supercomputer has 19.2 PB of available storage using a 20+4 erasure coded pool on 8 TB disks. The storage system is capable of sustaining 5–10 GB/s per user or job compared to the bandwidth of the current PFS bandwidth of 400–800 GB/s [13]. Where the Trinity campaign storage system uses 20 data shards and four erasure coding shards for data protection, our research shows that this can be scaled by a factor of six to a 120+24 erasure pool configuration. Merging 6 20+4 pools into a single 120+24 pool will provide much larger stripes for storing objects and will reduce the risk of data loss by increasing the population of the disk pool. The policy for repairing erasures immediately can be relaxed so that erasures are repaired when the data are read, avoiding the cost of transfering data to compute erasures for repair. With a Lazy Erasure Repair policy, erasures are repaired and written back to disk on a per object basis which futher distributes the workload of repairing erasures.

As the storage technology has evolved and continues to change in the future, the architecture of High Performance Computing (HPC) is changing. Where, in the past, IO bandwidth was free when buying enough disk to meet capacity requirments, that is no longer the case [9]. As the cost of solid state storage technologies have continued to drop, it is now more economical to design in high speed storage between the computer memory and the PFS constructed with solid state devices (SSD) [1]. Because the volume

of data is also increasing, another layer, campaign storage, based on erasure coded cloud technology, can be used which meets the pre-archive requirements at lower cost [9]. As Storage Class Memory (SCM) begins to appear, it may overtake the role of the Burst Buffers (BB) and even the PFS [2]. This would leave the HPC architecture with three storage layers: SCM, Campaign, and Archive. To this end, it is important to study the erasure coded cloud storage technology for HPC to improve the bandwidth, capacity, reliability, and cost reduction.

We have integrated the Gibraltar library as an erasure coding plugin in the Ceph Object Storage System. We have shown that the NVIDIA® K40 GPU [14] can perform erasure coding and erasure reconstruction while exceeding the bandwidth requirements for Trinity Campaign Storage. We have also shown that the Gibraltar library can maintain this performance at high degrees of sharding [10]. These performance facts were determined by comparing the erasure coding and reconstruction of data stripes using the Ceph [17] plugin architecture [7]. We compared the results using Gibraltar [3–6] against the Intel ISA-L library [16] and the Jerasure library [11]. We are now working to show that these capabilities can reduce the cost of campaign storage without reducing the performance, capacity or reliability with respect to other approaches.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grants Nos. OAC-1541310, CNS-0821497 and CNS-1229282. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This material is based upon work supported by Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] John Bent, Gary Grider, Brett Kettering, Adam Manzanares, Meghan McClelland, Aaron Torres, and Alfred Torrez. 2012. Storage challenges at los alamos national lab. In *Mass Storage Systems and Technologies (MSST), 2012 IEEE 28th Symposium on*. IEEE, 1–5.

[2] Lei Cao, Bradley W. Settlemyer, and John Bent. 2017. To Share or Not to Share: Comparing Burst Buffer Architectures. In *Proceedings of the 25th High Performance Computing Symposium (HPC '17)*. Society for Computer Simulation International, San Diego, CA, USA, 4:1–4:10. http://dl.acm.org/citation.cfm?id=3108096.3108100

[3] Matthew L. Curry, Anthony Skjellum, H. Lee Ward, and Ron Brightwell. 2008. Accelerating Reed-Solomon coding in RAID systems with GPUs. In *Proceedings of the 2008 IEEE international parallel & distributed processing symposium*. IEEE, Miami, FL, USA, 1–6. https://doi.org/10.1109/IPDPS.2008.4536322

[4] Matthew L. Curry, Anthony Skjellum, H. Lee Ward, and Ron Brightwell. 2011. Gibraltar: A Reed-Solomon coding library for storage applications on programmable graphics processors. *Concurrency and Computation: Practice and Experience* 23, 18 (Dec. 2011), 2477–2495. https://doi.org/10.1002/cpe.1810

[5] Matthew L. Curry, H. Lee Ward, Anthony Skjellum, and Ron Brightwell. 2010. A Lightweight, GPU-Based Software RAID System. In *2010 39th International Conference on Parallel Processing*. IEEE, San Diego, CA, USA, 565–572. https://doi.org/10.1109/ICPP.2010.64

[6] Curry, Matthew L. 2010. *A highly reliable GPU-based RAID system*. Ph.D. Dissertation. University of Alabama at Birmingham. http://contentdm.mhsl.uab.edu/cdm/ref/collection/etd/id/854

[7] Loic Dachary and Samuel Just. 2013. Erasure Code. (Aug. 2013). https://github.com/dachary/ceph/blob/wip-4929/doc/dev/osd_internals/erasure-code.rst

[8] Rex Farrance. 2006. Timeline: 50 Years of Hard Drives. (Sept. 2006). http://www.pcworld.com/article/127105/article.html

[9] Gary Grider. 2016. HPC Storage and IO Trends and Workflows. (April 2016). http://salishan.ahsc-nm.org/program.html

[10] Haddock, Walker, Curry, Matthew L., Bangalore, Purushotham V., and Skjellum, Anthony. 2017. GPU Erasure Coding for Campaign Storage. In *HPC-IODC ISC'17*. Springer, Frankfurt, Germany.

[11] James S. Plank, Scott Simmerman, and Catherine D. Schuman. 2008. *Jerasure: A Library in C/C++ Facilitating Erasure Coding for Storage Applications*. Technical Report Technical Report CS-08-627. University of Tennessee, Knoxville, TN 37996. 59 pages. http://www.cs.utk.edu/~plank/plank/papers/CS-08-627.html

[12] Kyle Lamb. 2015. Trinity Campaign Storage and Usage Model. (Aug. 2015). https://www.lanl.gov/projects/trinity/_assets/docs/trinity-usage-model-presentation.pdf

[13] David Morton. 2017. Trinity: DataManagement Scheme and Performance. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2017-0812

[14] NVIDIA. 2013. Tesla K40 GPU Active Accelerator. (Nov. 2013). https://www.nvidia.com/content/PDF/kepler/Tesla-K40-Active-Board-Spec-BD-06949-001_v03.pdf

[15] Anton Shilov. 2016. Seagate Unveils 10 TB Helium filled Hard Disk Drive. (Jan. 2016). http://www.anandtech.com/show/9955/seagate-unveils-10-tb-heliumfilled-hard-disk-drive

[16] Greg Tucker. 2014. ISA-L open source v2.14 API doc. (April 2014). https://01.org/sites/default/files/documentation/isa-l_open_src_2.10.pdf

[17] Sage A Weil, Scott A Brandt, Ethan L Miller, Darrell DE Long, and Carlos Maltzahn. 2006. Ceph: A scalable, high-performance distributed file system. In *Proceedings of the 7th symposium on Operating systems design and implementation*. USENIX Association, 307–320.