GPU Erasure Coding for Campaign Storage

2017 HPC-IODC SAND2017-6284C

22 June 2017

Walker Haddock[1] Matthew L. Curry[2]
Purushotham V. Bangalore[1] Anthony Skjellum[3]

Dept. of Computer and Information Sciences
University of Alabama at Birmingham

Center for Computing Research
Sandia National Laboratories

Dept. of Computer Science and Engineering &
McCrary Institute for Critical Infrastructure Protection and Cyber Systems
Auburn University

Agenda

Introduction

- Research in storage for Exascale
  - Velocity
  - Volume
  - Capital Expense
  - Operating Expense
  - Reduce Energy Consuption
  - High Availability and low data loss
- Can GPU Accelerators increase erasure coding performance?
- What is the impact on reads with many erasures?
- Can GPU erasure coding meet 1 GB/s bandwidth per File Transfer Appliance (FTA)?

Trinity Storage Stack
Data Life Times

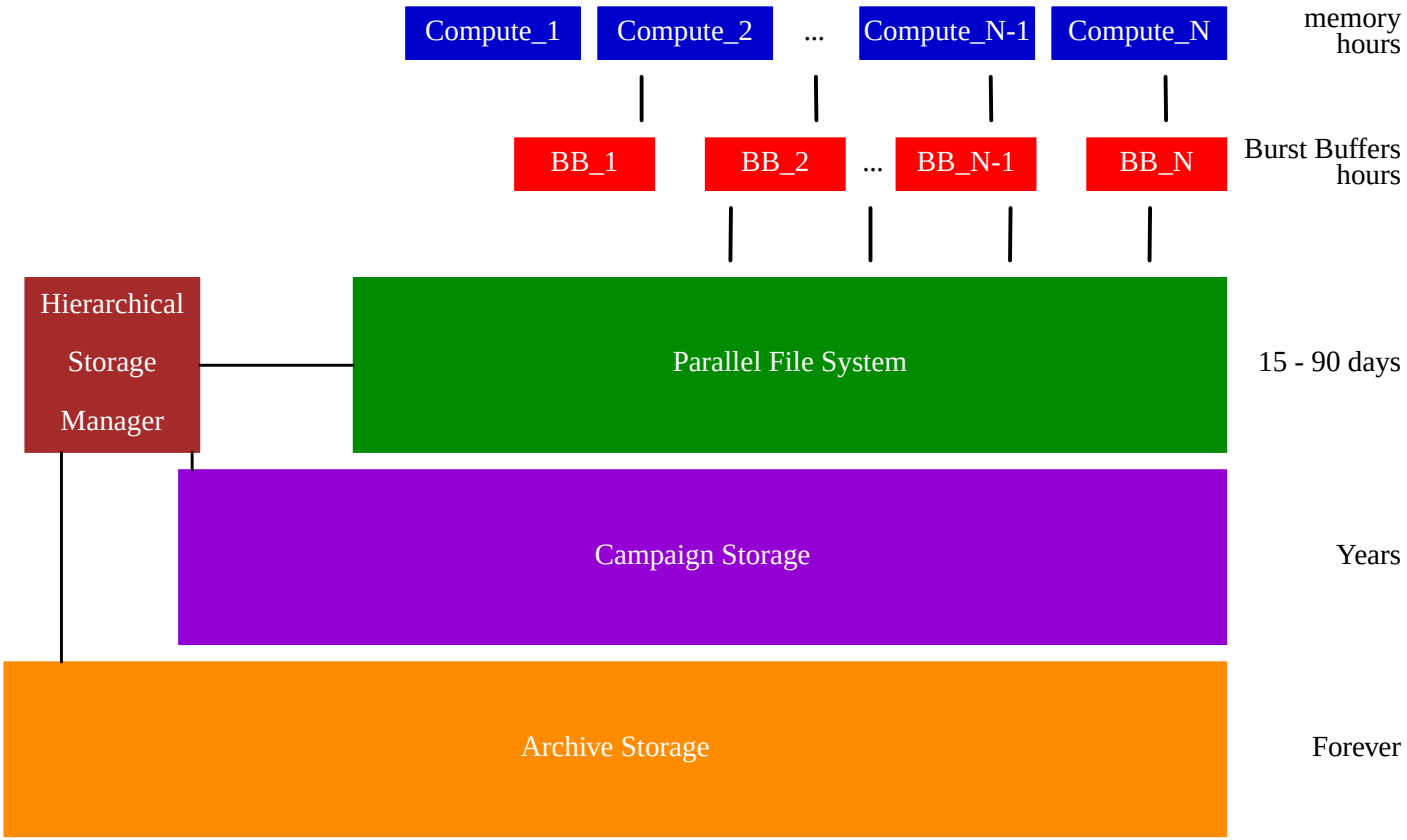

Figure: Hierarchy of HPC storage stack used by Trinity. Data velocity increases towards the top of the stack while the life time is lessened.

Ceph Interface to Gibraltar

GPU Erasure Coding Module

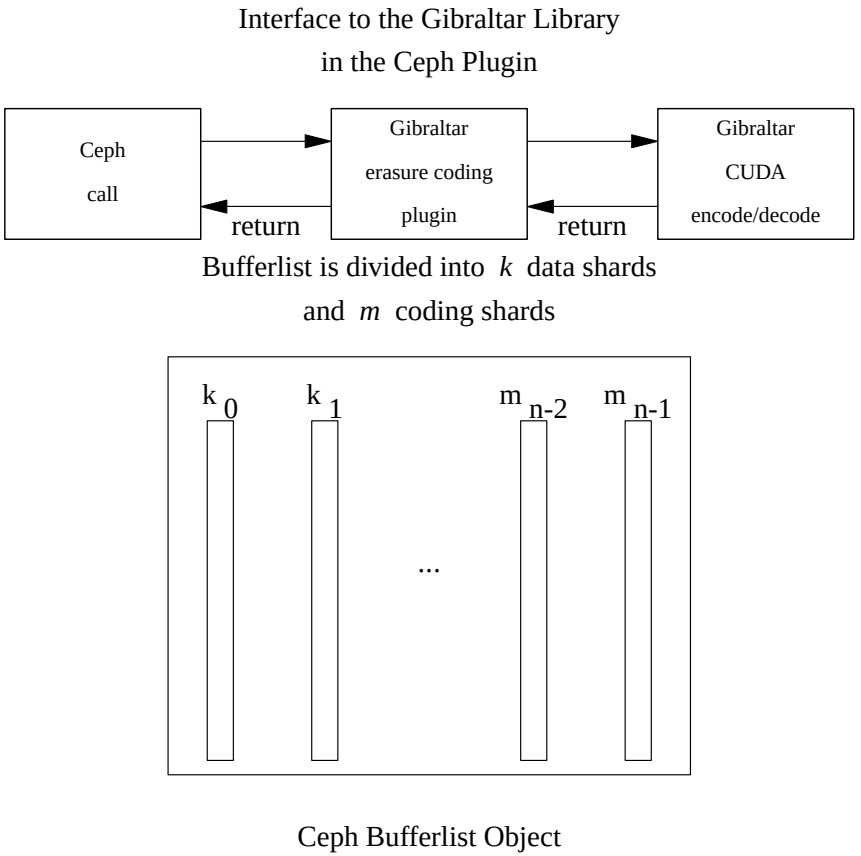

Figure: Ceph calls the erasure coding module with a Bufferlist object containing the stripe to be written to the object. The Plugin divides the Bufferlist into *k* data shards and adds *m* coding shards. Gibraltar is called to perform the coding or recovery.

Experiment Configuration
Compute Platform for Measurements

Table: Dell R730 with GPU Configuration

CPUs	2 Xeon E5-2650v3 @ 2.3 GHZ (HT-enabled: 40 threads)
RAM	128 GB 2133 MT/s RDIMM
Network	2 port Mellanox ConnectX-3 MCX354A-FCBS
	Intel X520 DP 10Gb DA/SFP+, I350 DP 1Gb Ethernet
GPU	NVIDIA® K40m GPU
System Drives	2 300 GB 10K SAS 2
	2 200 GB INTEL SSDSC2BG20 SATA
	2 400 GB TOSHIBA PX02SMF040 SAS 3

Encoding Benchmark

Compare Gibraltar with ISA-L

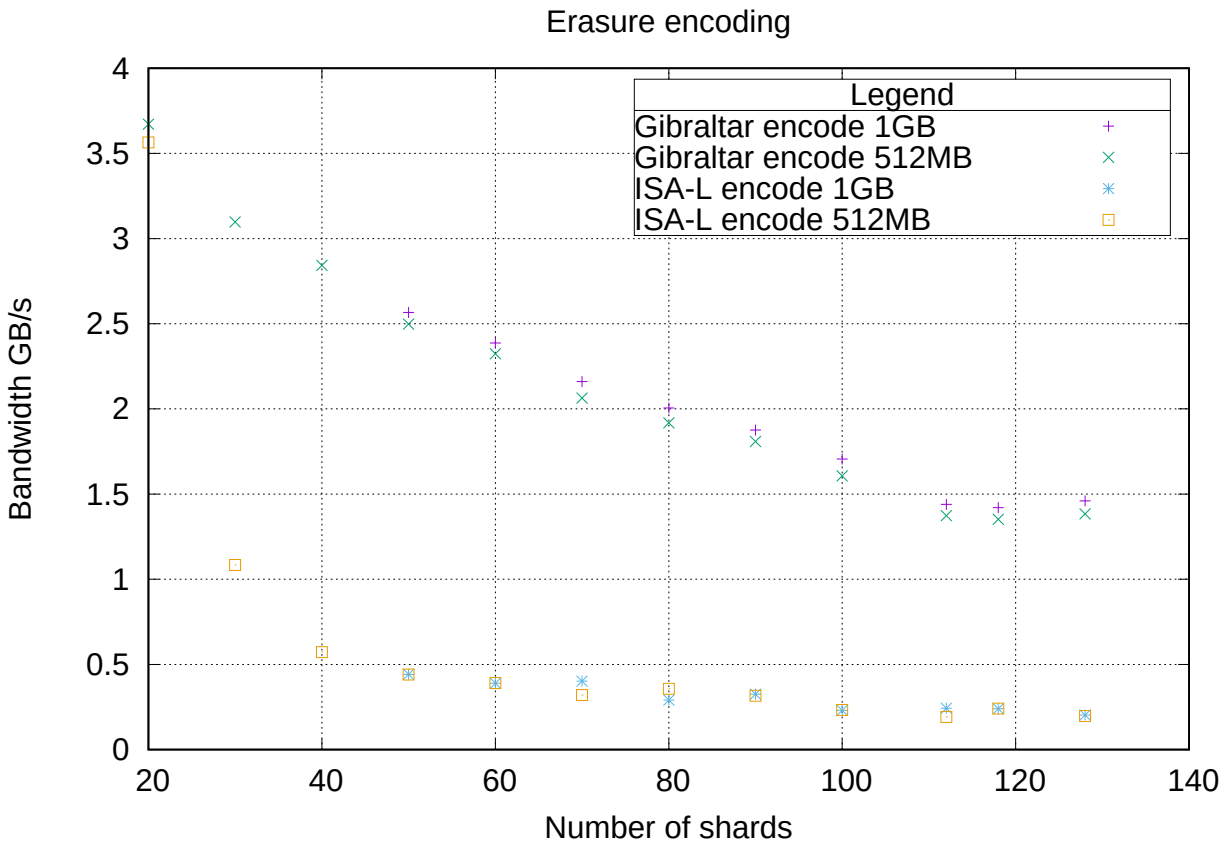

Figure: Erasure coding bandwidth results with increasing number of shards. Coding shards are held to a ratio of 1 coding shard to 5 data shards.

Decoding Benchmark (1)
Compare Gibraltar with ISA-L with 1 Erasure

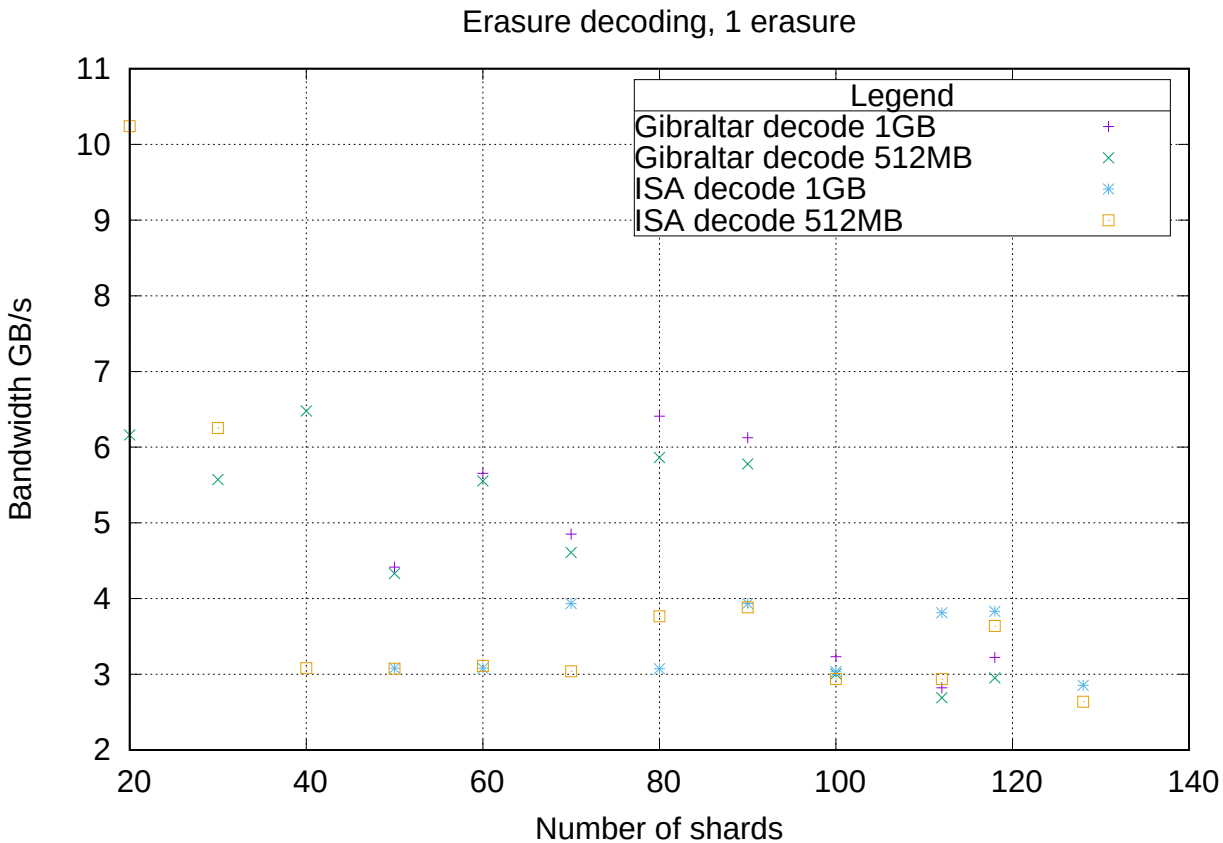

Figure: Erasure recovery bandwidth results with increasing number of shards and 1 erasure. Coding shards are held to a ratio of 1 coding shard to 5 data shards.

Decoding Benchmark (2)

Compare Gibraltar with ISA-L with 4 Erasures

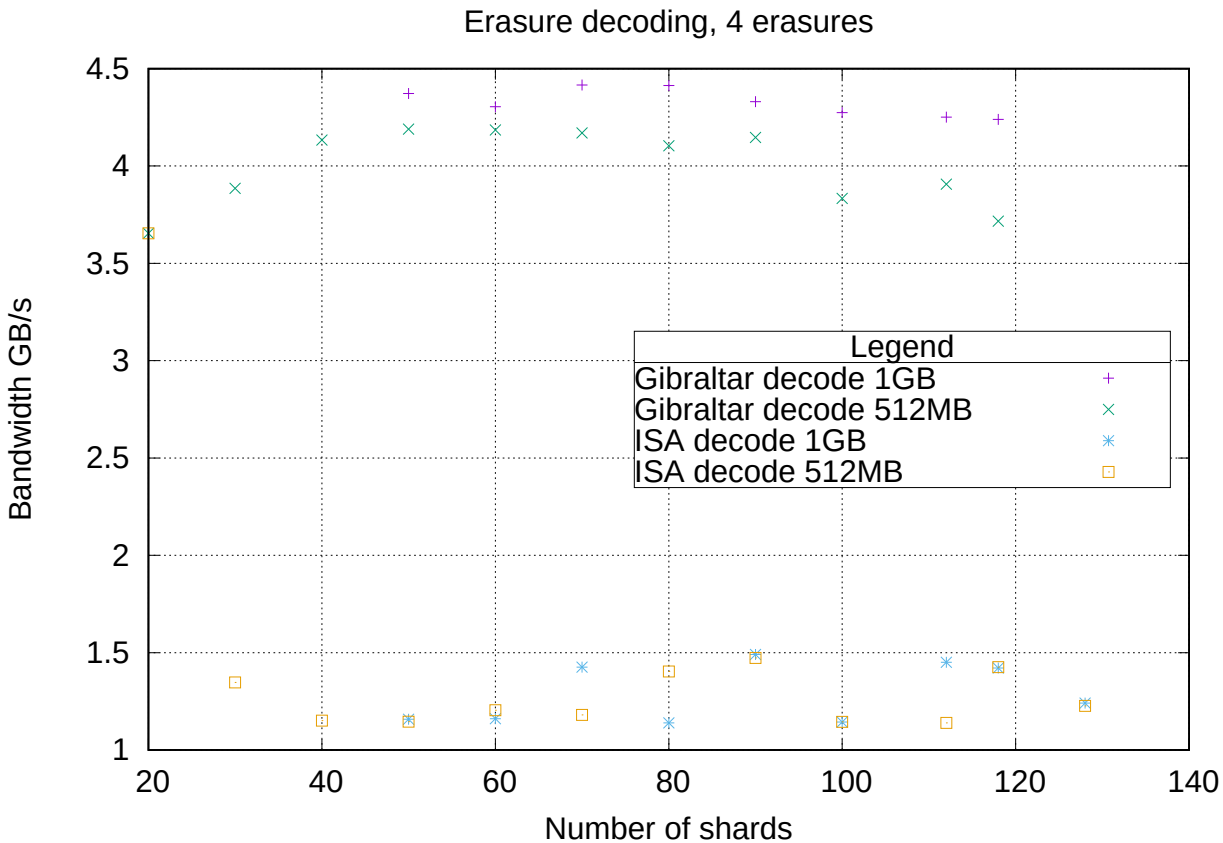

Figure: Erasure recovery bandwidth results with increasing number of shards and 4 erasures. Coding shards are held to a ratio of 1 coding shard to 5 data shards.

Shard Size vs. Sharding Degree

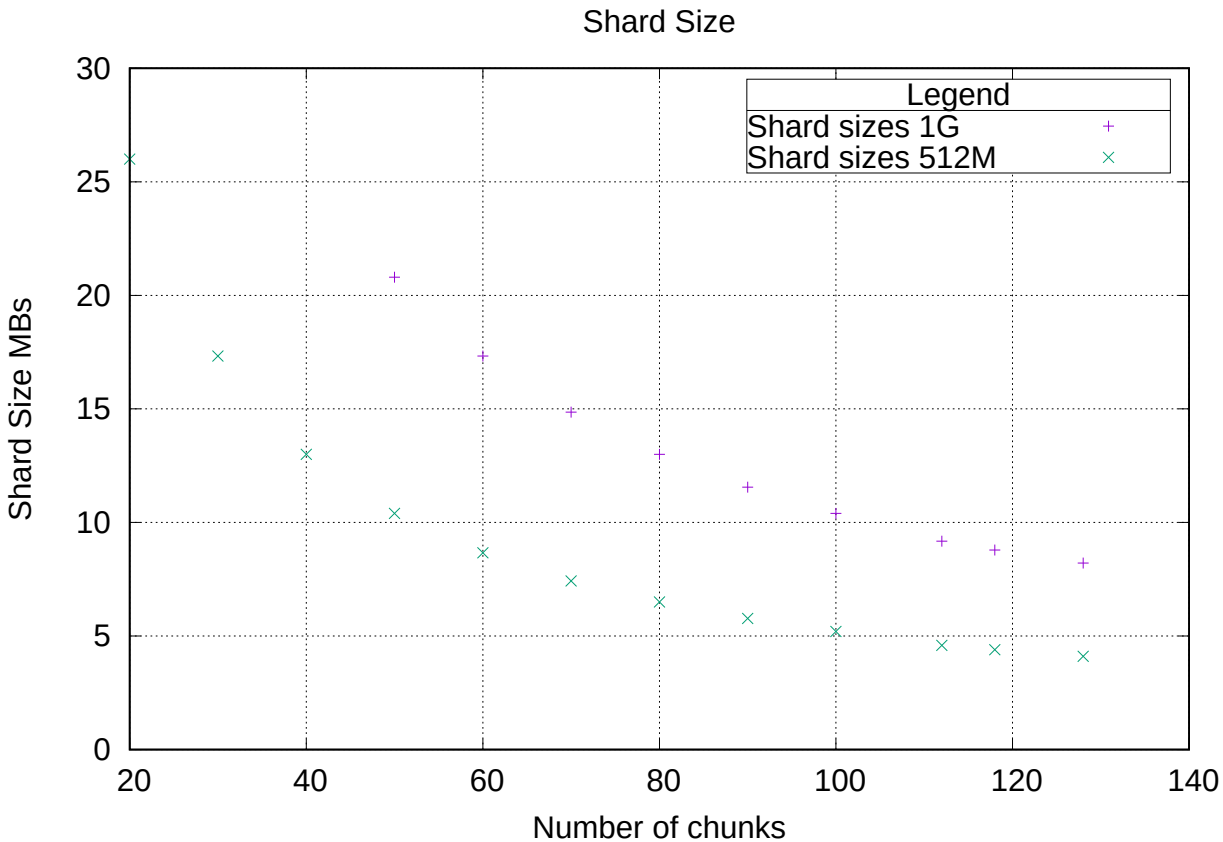

Figure: Shard sizes used in the erasure coding and decoding measurements. Higher degrees of sharding produce smaller shards.

Summary

- The Gibraltar plugin exceeded the 1 GB/s Trinity bandwidth requirement with all degrees of sharding measured while ISA-L falls behind after 20 shards.
- The Gibraltar plugin out performed the ISA-L library while encoding for all sharding degrees.
- The Gibraltar plugin showed no performance degredation with greater erasures while the ISA-L performance slowed down.
- Based on these measurements, the Gibraltar plugin would be capable of providing 1 GB/s full duplex performance per File Transfer Appliance with multiple erasures on read.
- Smaller shard sizes mean lower bandwidth requirements to OSDs. Fan out data over more disks.

Questions?

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants Nos. ACI-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 work has also been supported by Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.