SAND2010-7820P

A GPU-Based Storage System

Matthew L. Curry

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

UAB

Sandia National Laboratories

A RAID Overview

- RAID has served us well
- Several levels introduced initially
  - Some have been added, some have been removed
- Today, there are five major RAID types in use in reliable production systems

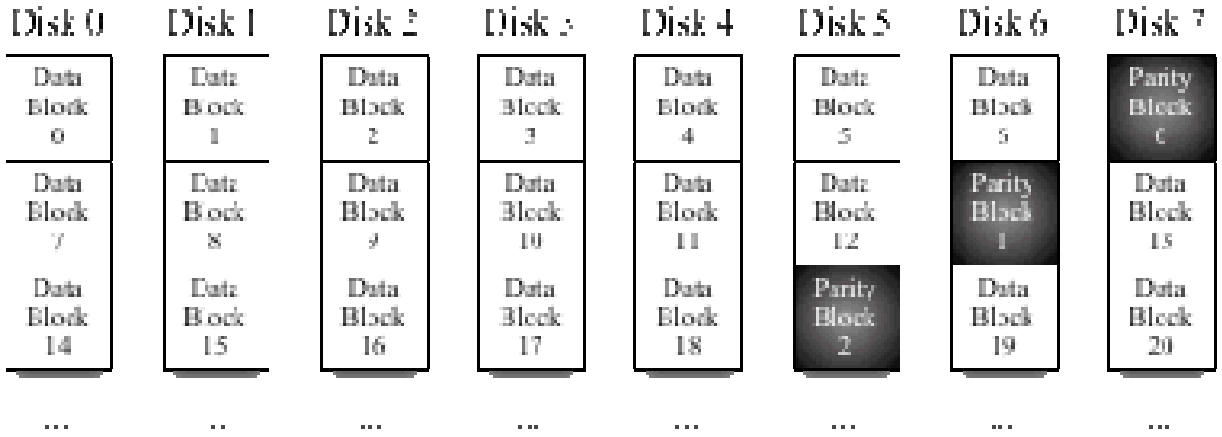
- Going forward, they will not be reliable enough.

UAB

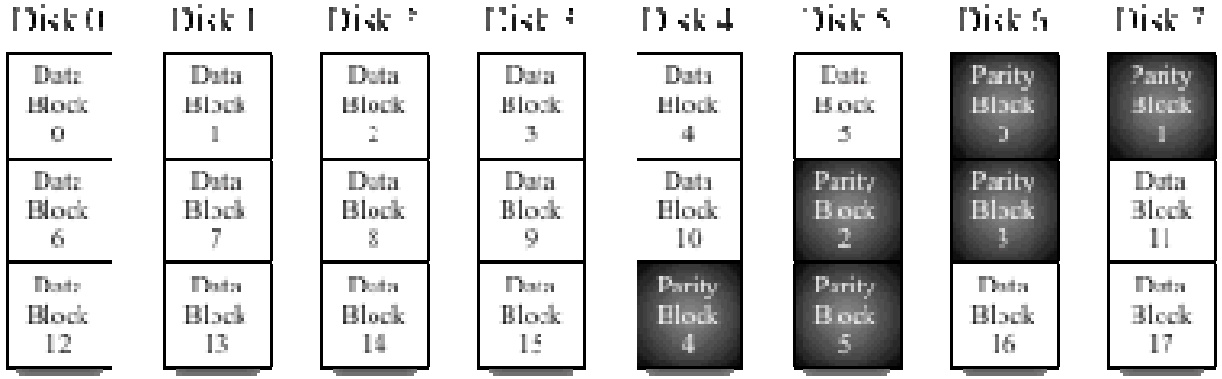
Sandia National Laboratories

Current Parity-Based RAID

RAID 5



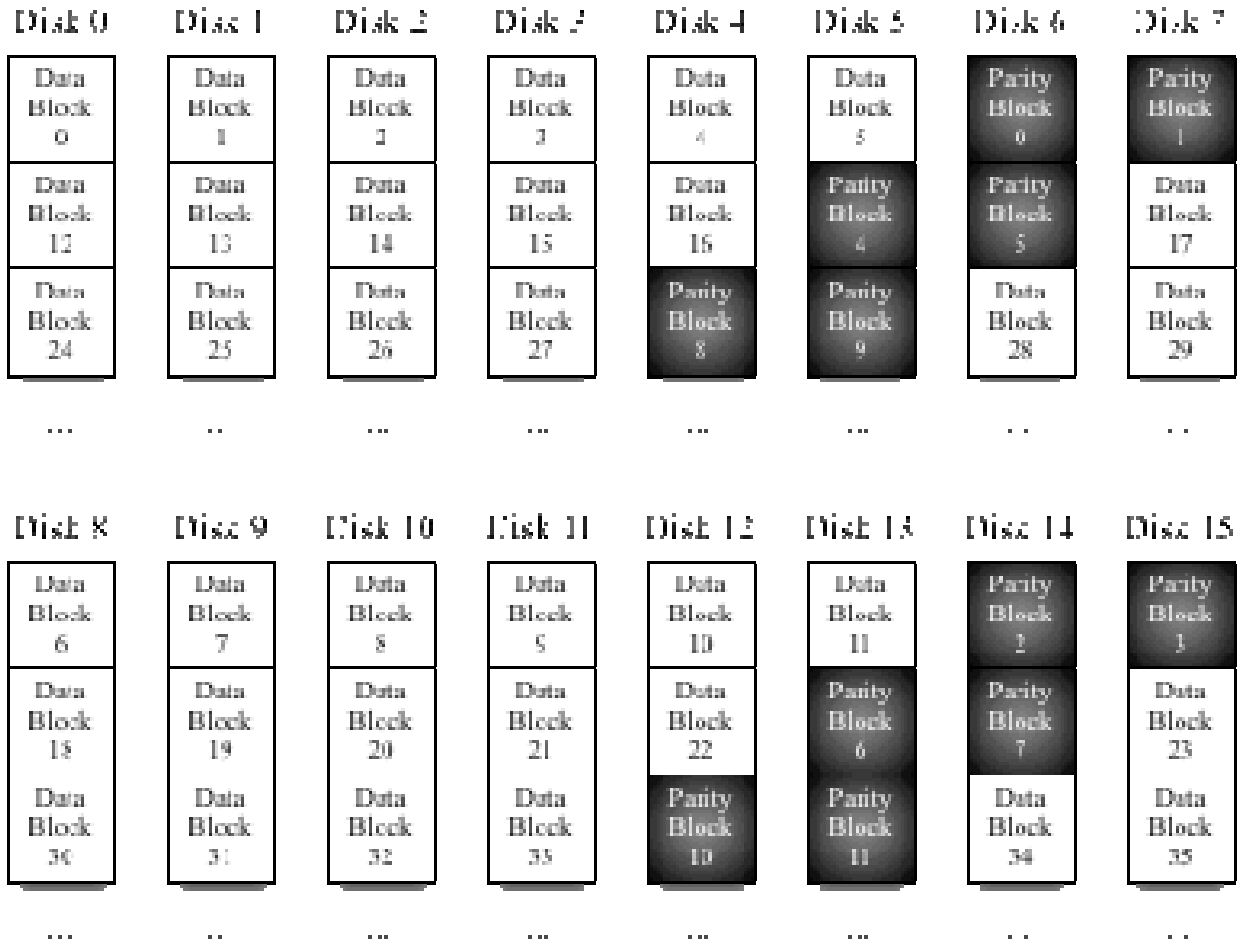
RAID 6



Sandia National Laboratories

Hierarchical Parity-Based RAID

RAID 6+0

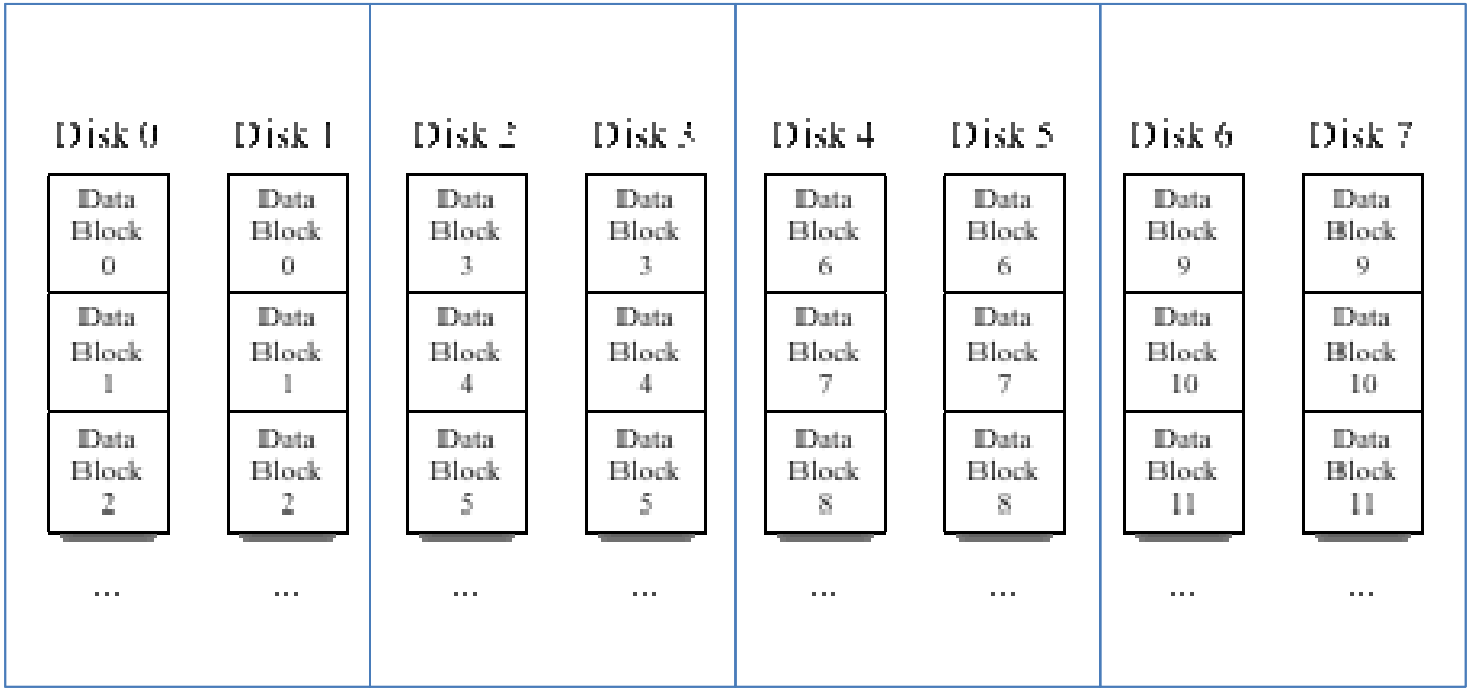


UAB

Sandia National Laboratories

Hierarchical Mirror-Based RAID

RAID 1+0



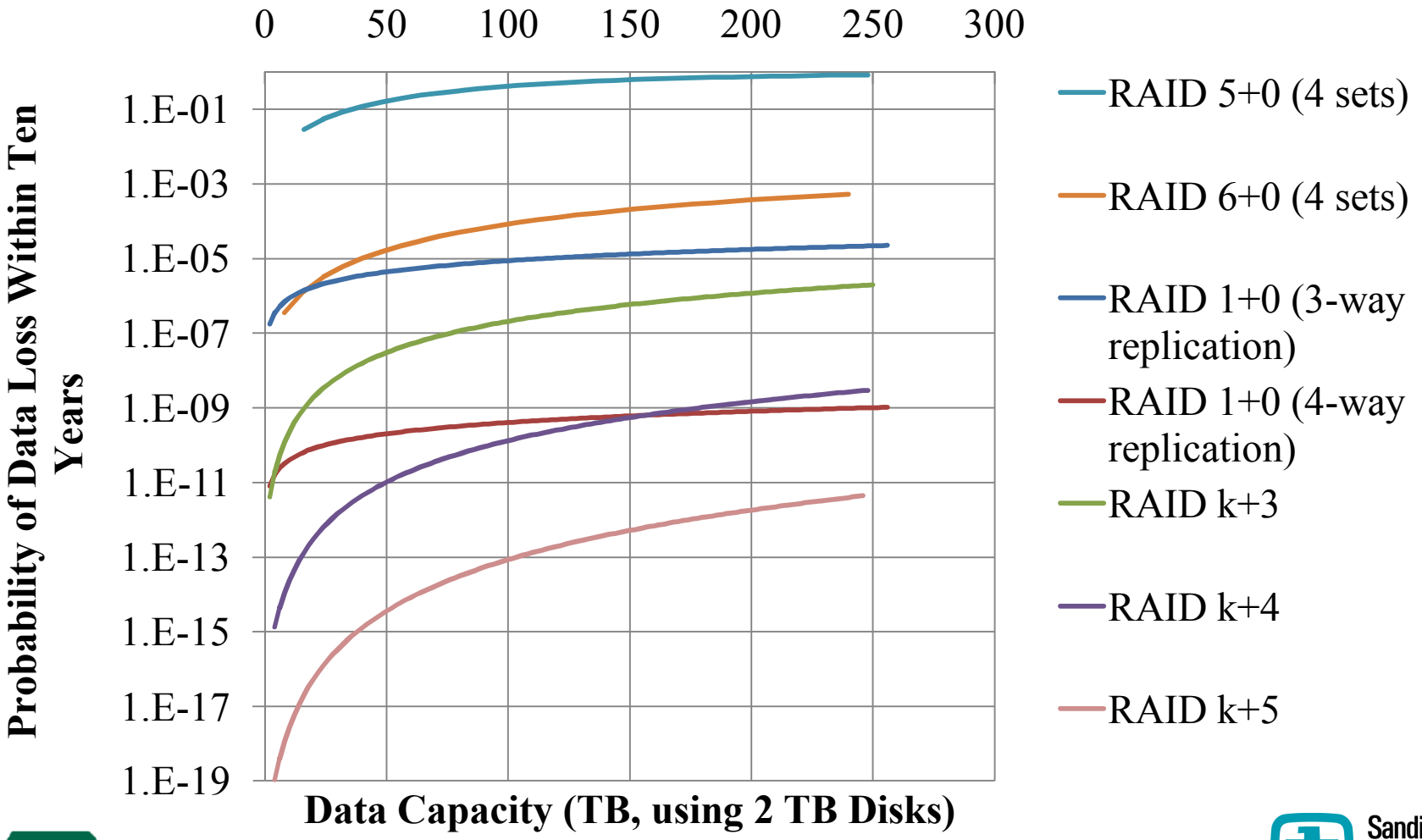
k+m RAID, an Alternative

- The user can choose the amount of parity required for a RAID set
  - *k+1* RAID = RAID 5
  - *k+2* RAID = RAID 6
  - *k+3* RAID = RAID TP (Esoteric)
  - *k+4* RAID or better does not exist anywhere. (RAID 6+0 does not satisfy the requirements.)

UAB

Sandia National Laboratories

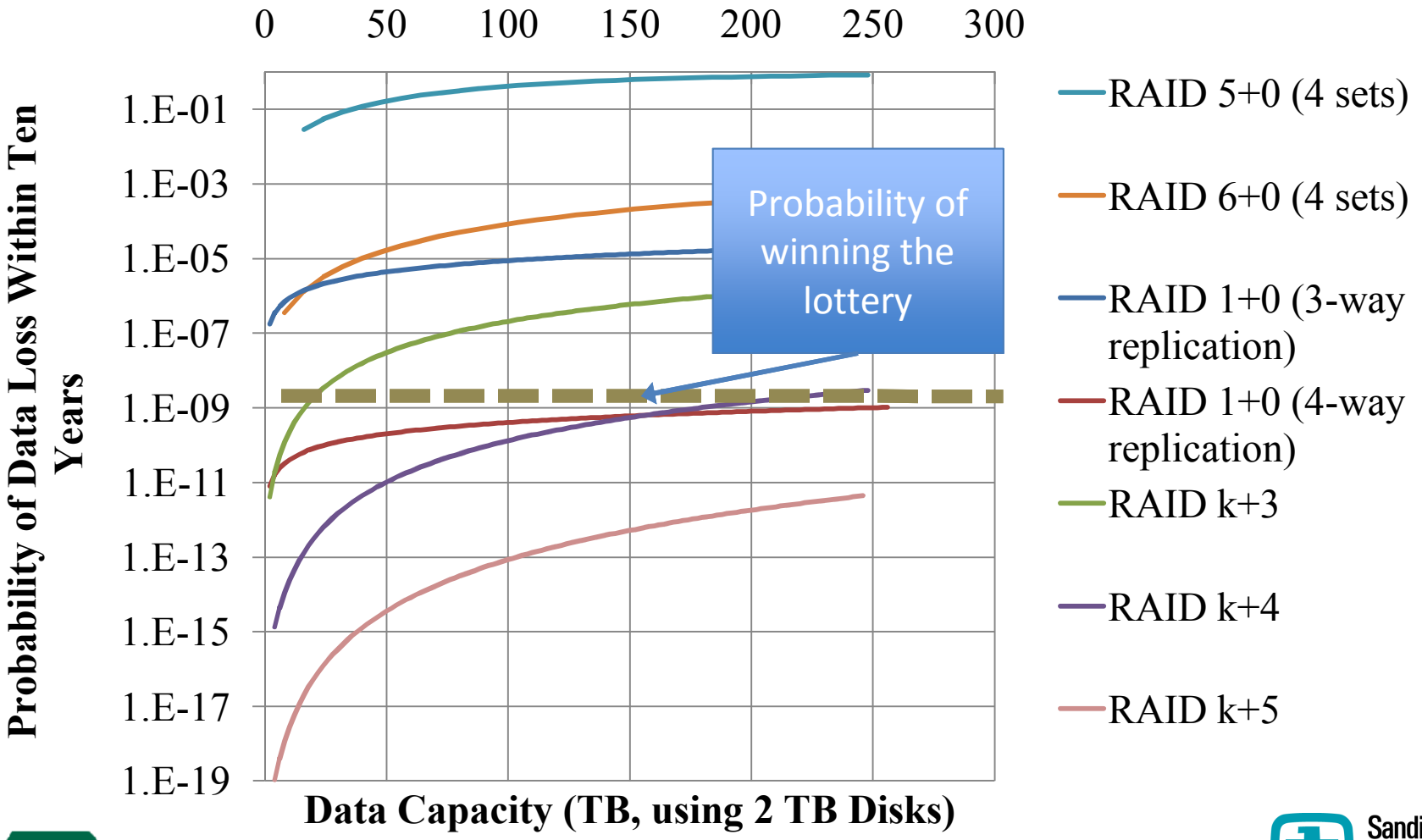
k+m RAID Reliability, Standard Manufacturer Statistics



UAB

Sandia National Laboratories

k+m RAID Reliability, Standard Manufacturer Statistics



UAB

Sandia National Laboratories

Why is *k+m* RAID Necessary?

- Correlated failures
  - Manufacturer defects (batch correlated failures)
  - Environment and workload as factors
- Calculation methods of Mean Time To Failure
  - Disk age/failure correlation
  - Manufacturer-reported disk failure rates are more than 10x lower than observed
- Rebuild times are skyrocketing
  - Disk speeds are not increasing proportionally to disk capacity, and flash cannot save us yet

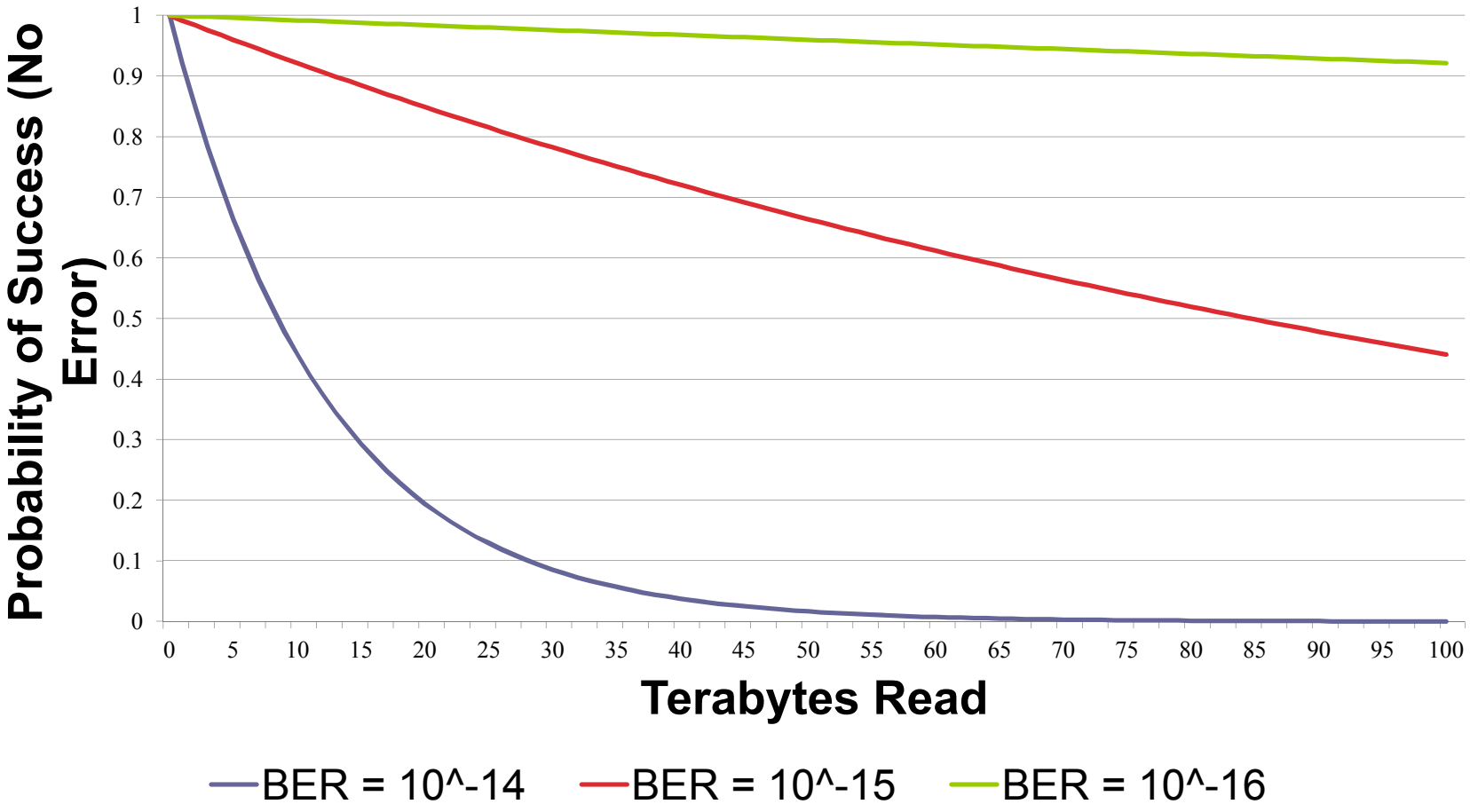
UAB

Sandia National Laboratories

Why is *k+m* RAID Necessary? (2)

- Data corruption/loss
  - Disk firmware bugs
  - Unrecoverable read errors
    - One sector in 100-1000 TB will be unreadable
  - Cabling, misdirected/torn writes
  - Also points to need for read verification
- Parity needs to be available to recover from these errors
  - RAID 6 double disk failures *do* happen

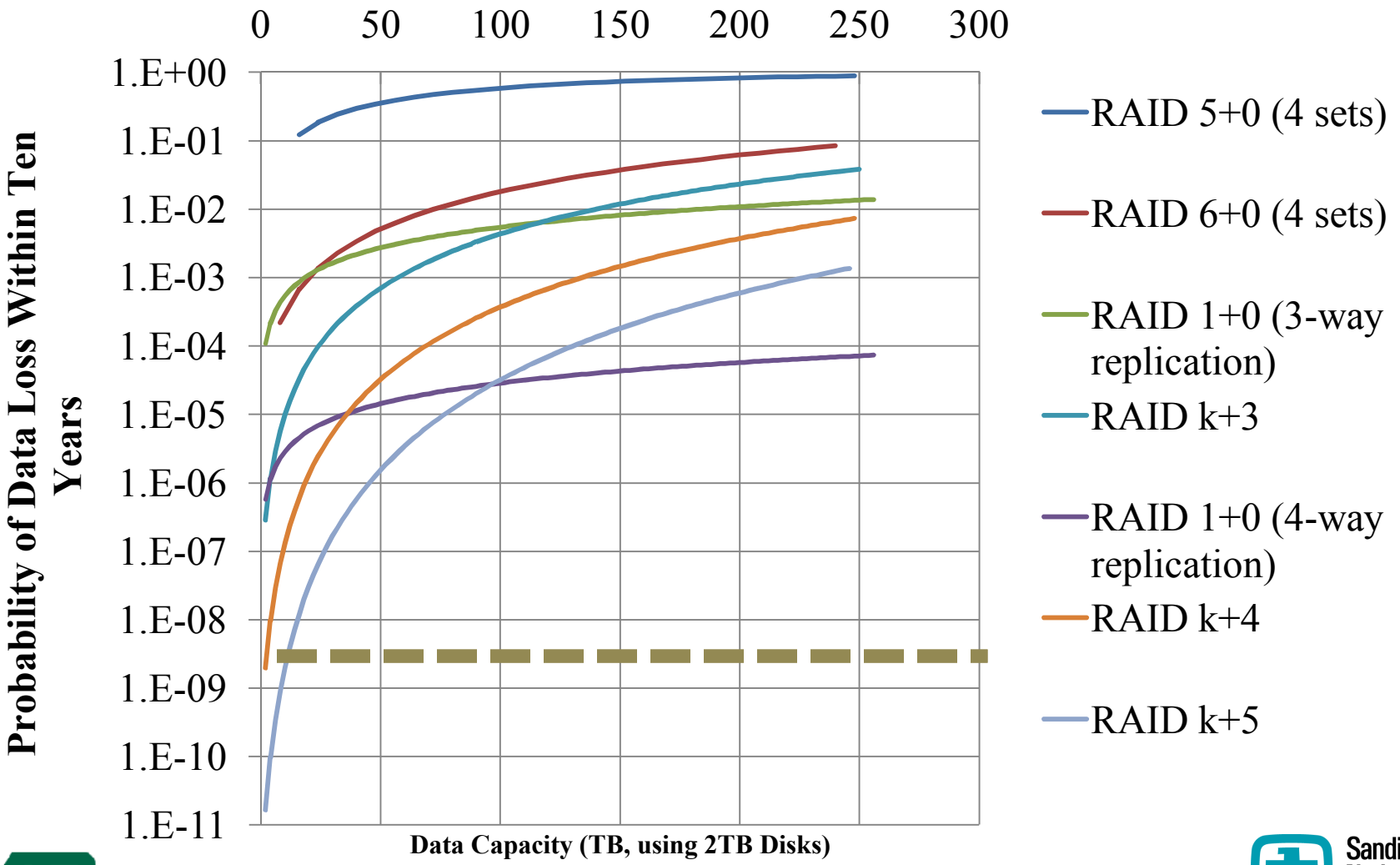
Unrecoverable Read Errors



UAB

Sandia National Laboratories

RAID Reliability, Updated Statistics

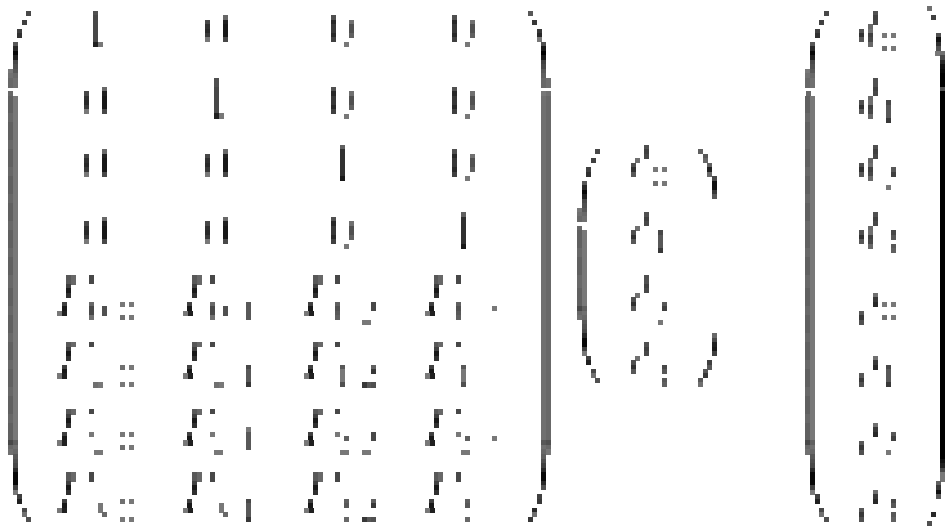


UAB

Sandia National Laboratories

Erasure Coding Computations

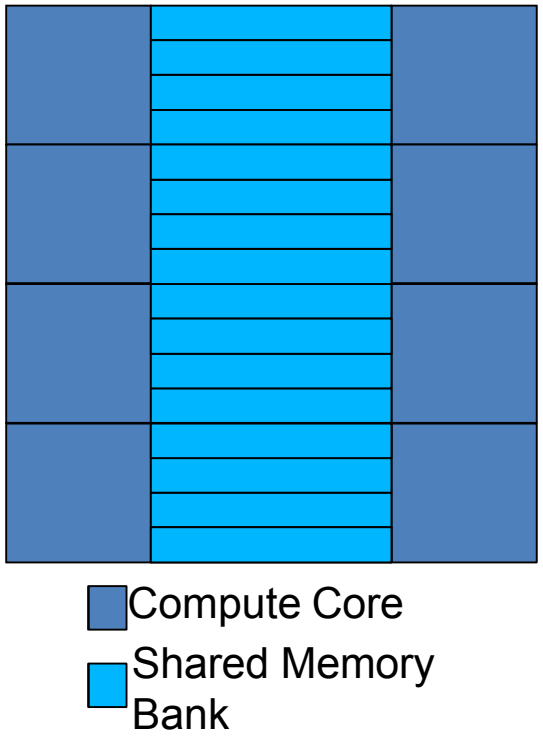
- *k+m* RAID is necessary
- Reed-Solomon coding is required to implement it
  - (*k+m*) *x k* matrix
  - Finite field arithmetic
  - Table look ups abound



GPU Architecture Does This Well!

- Each GPU contains some number of Streaming Multiprocessors (SMs)
- Banks allow parallel non-conflicting accesses
- GPU RAID simulation results: 1.82 look-ups per cycle per SM on average
- GeForce GTX 285 has 30 SMs, so can satisfy 55 look-ups per cycle for device

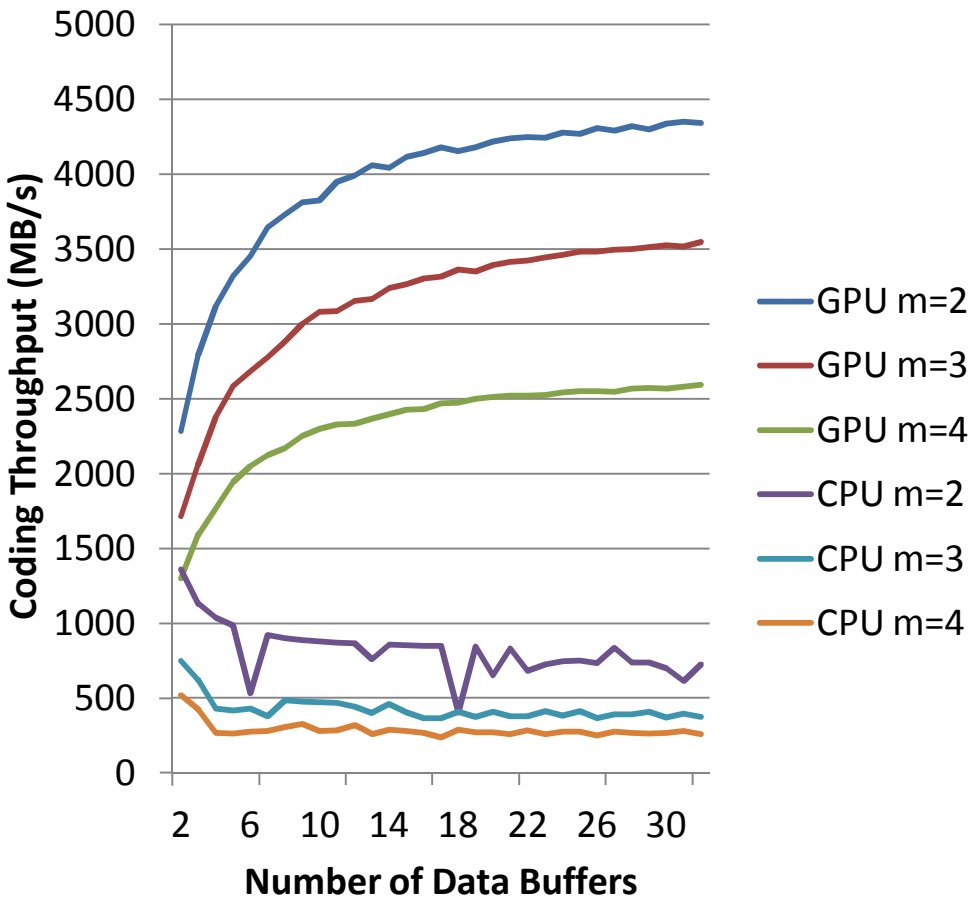
CUDA 200-Series Streaming Multiprocessor



UAB

Sandia National Laboratories

GPU-Based Reed-Solomon



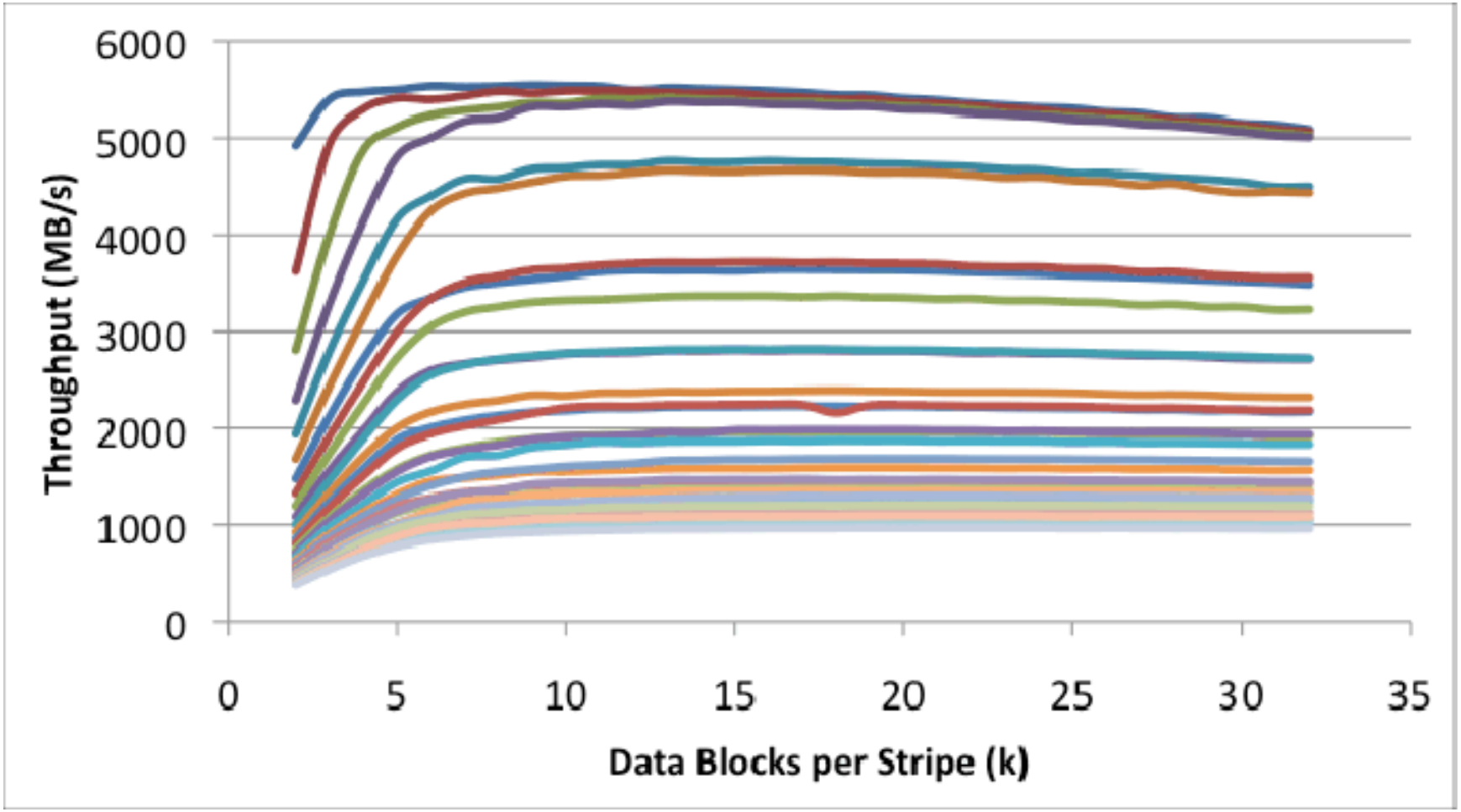
- GPU capable of providing high bandwidth at 6x-10x CPU speeds
- New memory layout and matrix generation algorithm for Reed-Solomon yields equivalent write and degraded read performance

UAB

Sandia National Laboratories

k+m Coding on GeForce GTX 480

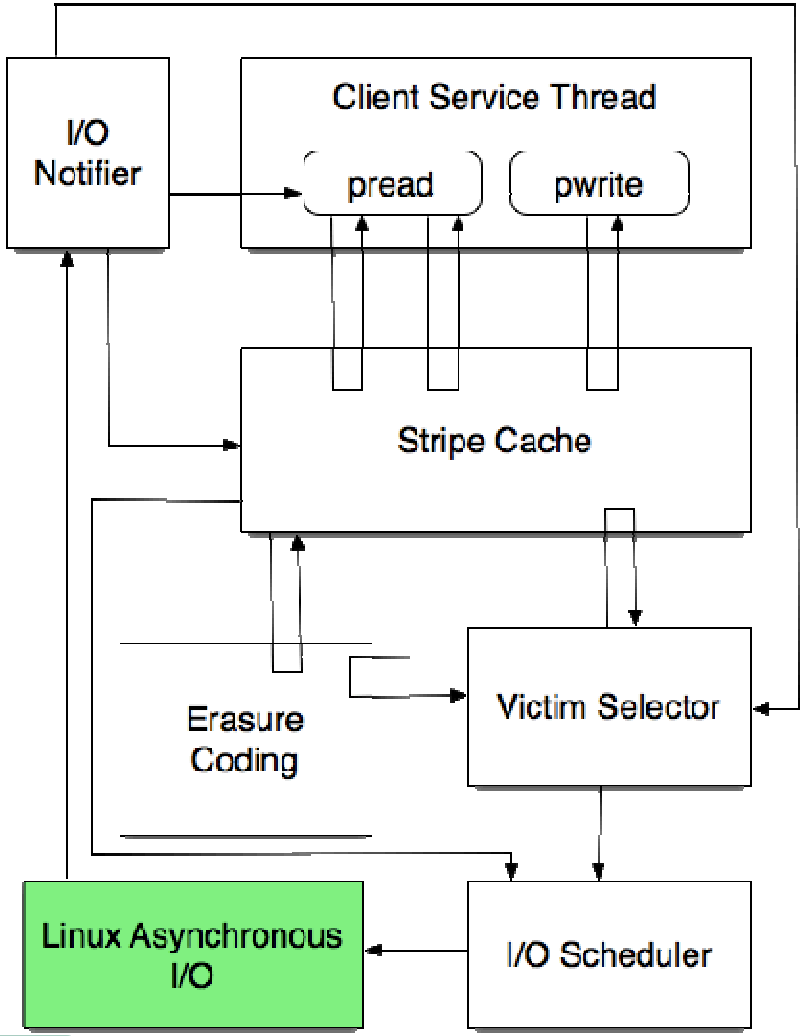
(Each line represents a value of m, 3 ≤ m ≤ 32)



UAB

Sandia National Laboratories

A User Space RAID Architecture

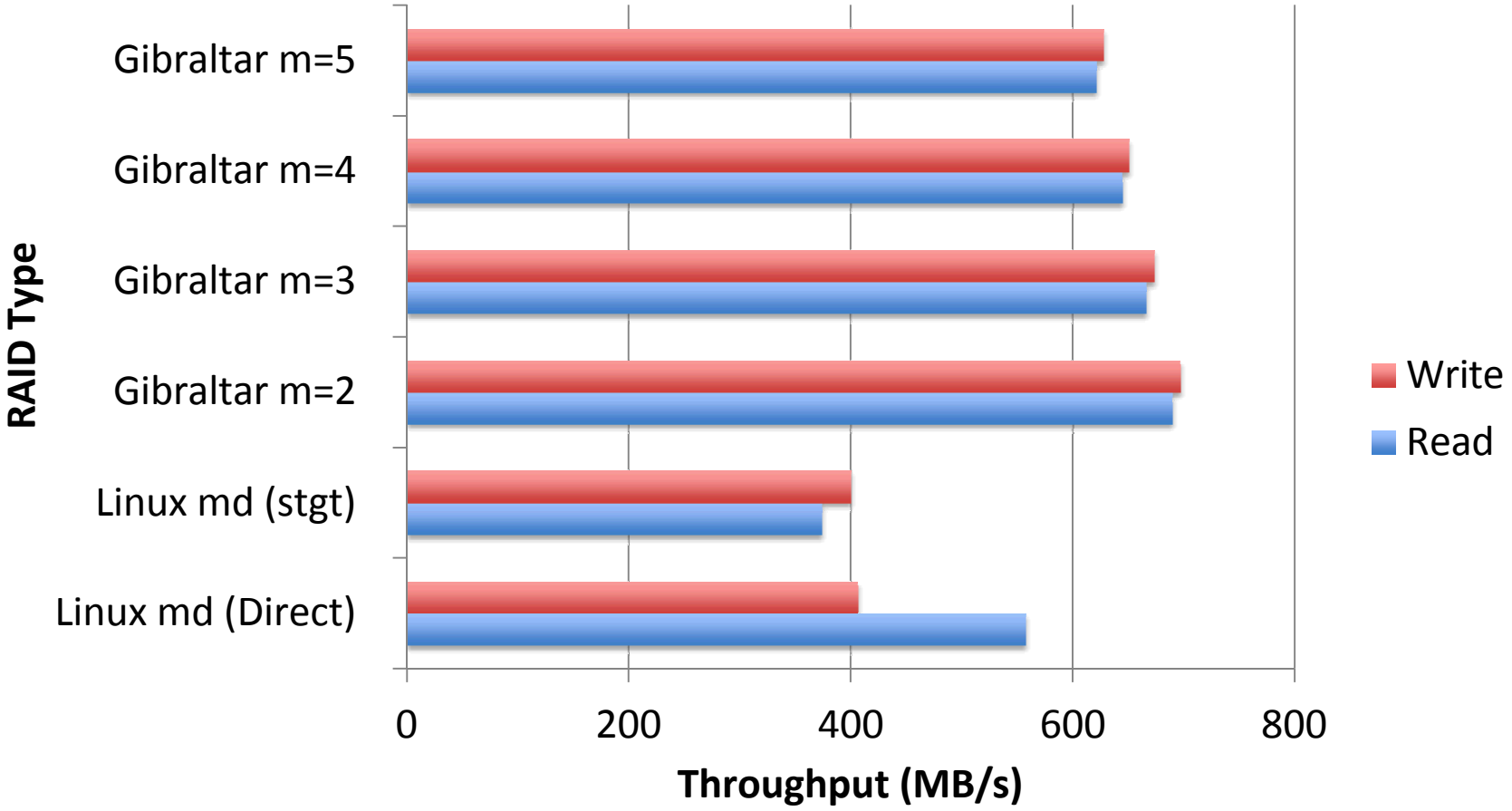


- GPU computing facilities are inaccessible from kernel space
- Provide I/O stack components in user space, accessible via iSCSI

UAB

Sandia National Laboratories

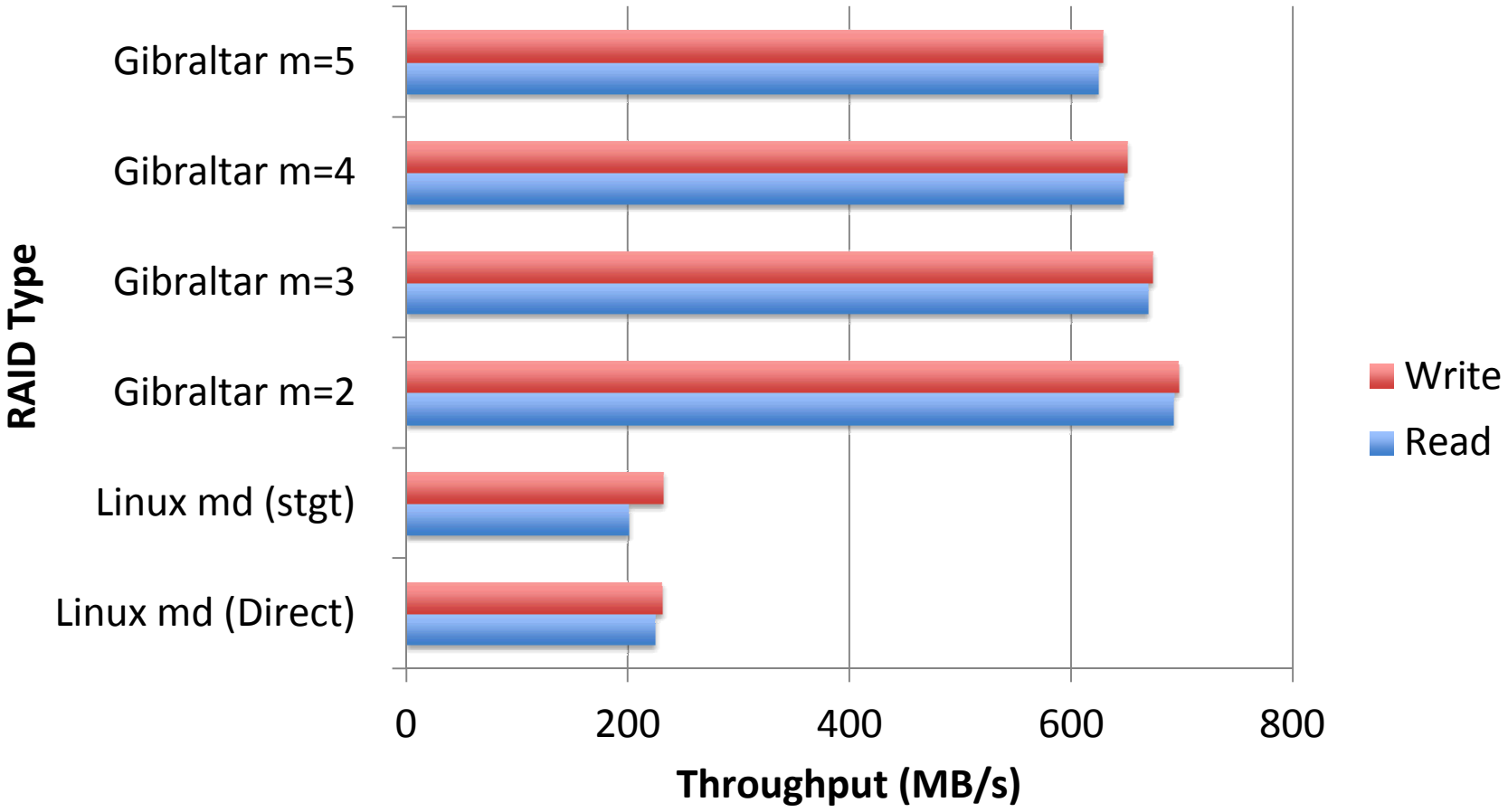
Direct Attached Storage



UAB

Sandia National Laboratories

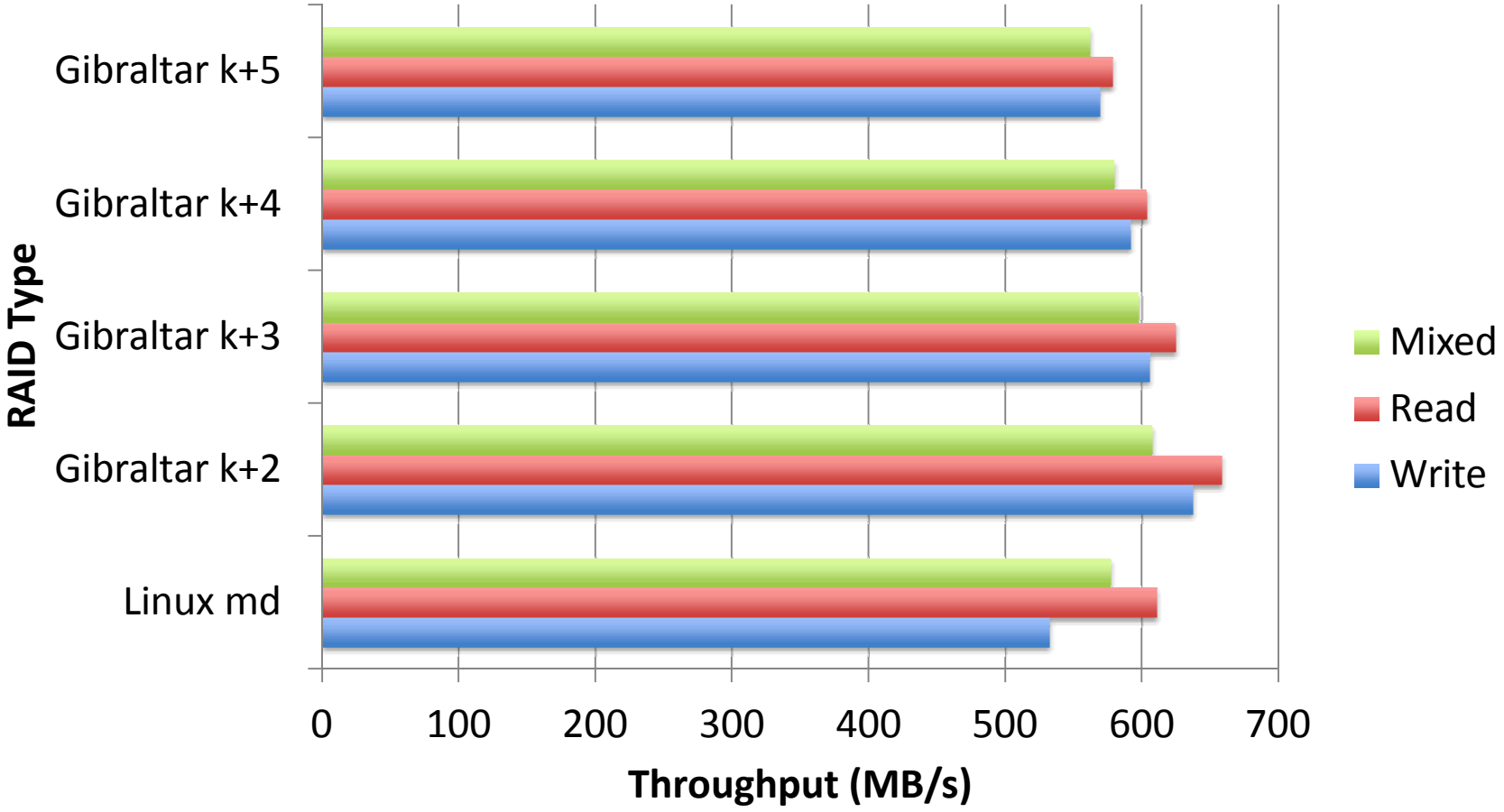
Direct Attached Storage, Degraded



UAB

Sandia National Laboratories

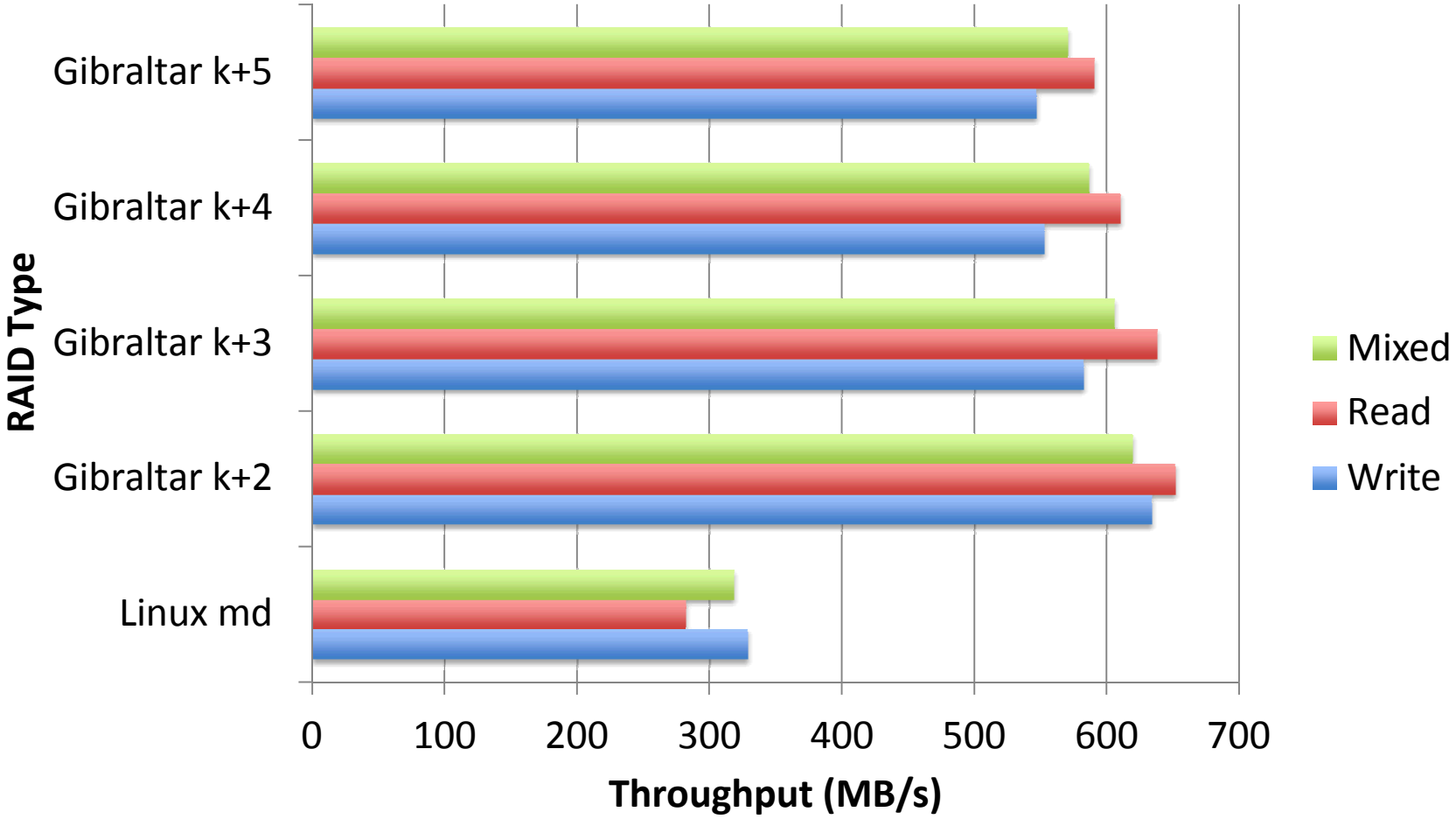
Network Attached Storage, Four Clients



UAB

Sandia National Laboratories

Network Attached Storage, Four Clients, Degraded



UAB

Sandia National Laboratories

Combine RAID with Compression

- Before or After Reed-Solomon coding is performed, it is desirable to perform other operations as well
- Initial investigations into GPU compression with Robert Cloud have yielded insight into integration
- Huffman coding can be performed at 2.6 GB/s on a GPU

UAB

Sandia National Laboratories

Combine RAID with Other Processing

- AES, allowing encryption on disk
  - "Design of a Parallel AES for Graphics Hardware using the CUDA framework," Biago et al., IPDPS 2009
- SHA-1, for block-level deduplication or content-based addressing
  - "A Brief Implementation Analysis of SHA-1 on FPGAs, GPUs, and Cell Processors," 2009 International Conference on Engineering and Computation

UAB

Sandia National Laboratories

Fail in Place

- Storage bricks in remote or harsh locations
- Degraded mode operation presents risks that affect hardware and software RAID
  - Lowered performance
  - Increased risk of multiple failures
  - Hot spare induced window of vulnerability
- High-parity RAID can mitigate the risks of degraded mode operation

UAB

Sandia National Laboratories

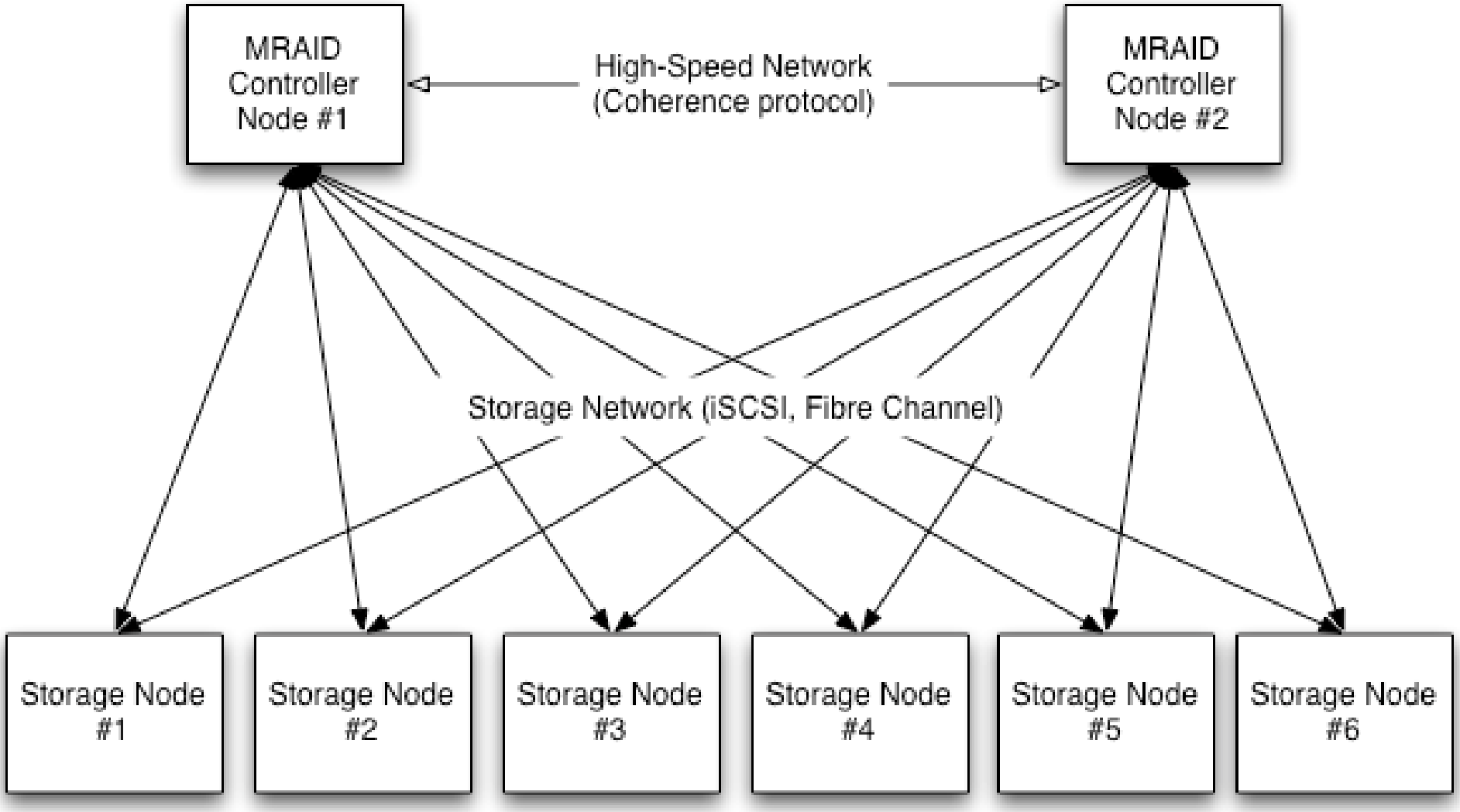
Using MRAID for Data Center Resiliency

- MRAID is similar to RAID
  - Disks → Storage servers
- Multiple controllers
  - Multiple paths of access
- Node failures
  - Power supplies
  - Disks
  - Network

UAB

Sandia National Laboratories

MRAID


MRAID Controller Node #1
High-Speed Network (Coherence protocol)
MRAID Controller Node #2
Storage Network (iSCSI, Fibre Channel)
Storage Node #1
Storage Node #2
Storage Node #3
Storage Node #4
Storage Node #5
Storage Node #6

Distributed Data Storage

- Current technologies currently use 3 x data replication, which is logically the same as RAID 1+0 with three sets
  - Google FS
  - Amazon S3
- User-managed distributed data storage can use high-speed coding
  - Multiple administrative domains
  - Low bandwidth overhead
  - Higher reliability

UAB

Sandia National Laboratories

Conclusions

- Introduced a high-parity RAID variant that is impractical in software today
- Prototyped a GPU-based library and RAID controller that can provide high parity
  - High speed
  - Low cost
  - High flexibility
- Detailed future major research areas
  - Processing in storage stack
  - Embedded/low maintenance storage
  - MRAID

UAB

Sandia National Laboratories