Redundant Array of Inexpensive Interactive Disks (RAI^{2}D)

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Knowledge that will change your world

Alexander S. Filby
University of Alabama at Birmingham
afilby@cis.uab.edu

Matthew L. Curry
Sandia National Laboratories
mlcurry@sandia.gov

H. Lee Ward
Sandia National Laboratories
lee@sandia.gov

Anthony Skjellum
University of Alabama at Birmingham
tony@cis.uab.edu

David R. Resnick
Sandia National Laboratories
drresni@sandia.gov

Architecture

We propose a novel protocol that can leverage the properties of a broadcast fabric in order to provide an inexpensive and controller-less RAID[3] solution. The proposed architecture is shown below in Figure 1.

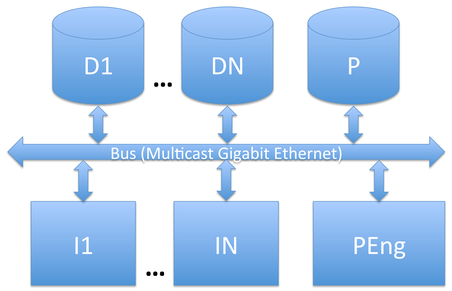

Figure 1: System Architecture – D1-DN represent data disks, P is the parity disk, PEng is the parity engine, and I1-IN are the initiators

A broadcast fabric allows each component actively to snoop on all traffic and be aware of the system state. This allows some logic and decision making to be moved to the components themselves.

Protocol

Consists of four message types: Control, Data, Status, and Sync. These support three operations: Retrieve, Store, and Sync. As implemented, the array is RAID4.

ETH_HD	Xid	Group	Offset	Length	Op

ETH_HD	I_MAC	Xid	Req_Off	Length	Dsk_Tag

ETH_HD	I_MAC	Xid	Status	Dsk_Tag

Figure 2: Content of Control, Data, and Status Messages respectively. ETH_HD is the Ethernet Frame header, I_MAC is the MAC address of the Initiator and xid is the transaction id.

Retrieve Operation

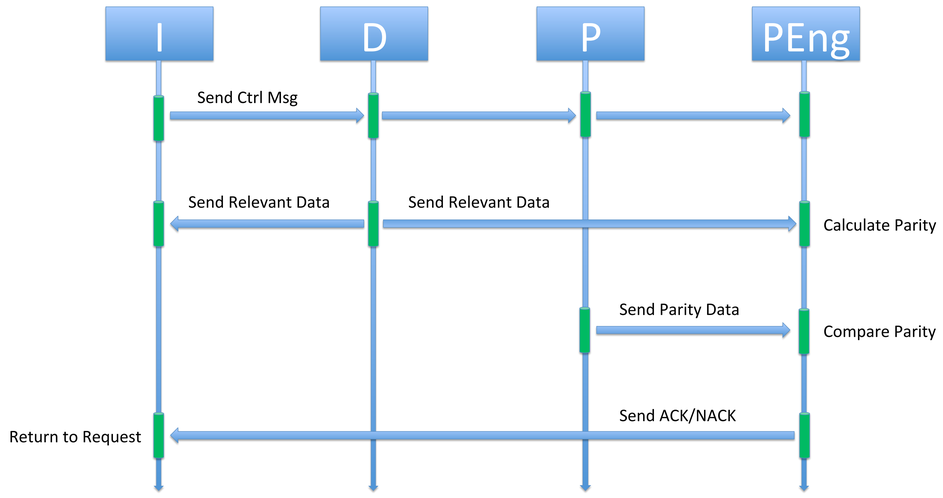

Figure 3: Retrieve operation sequence

Store Operation

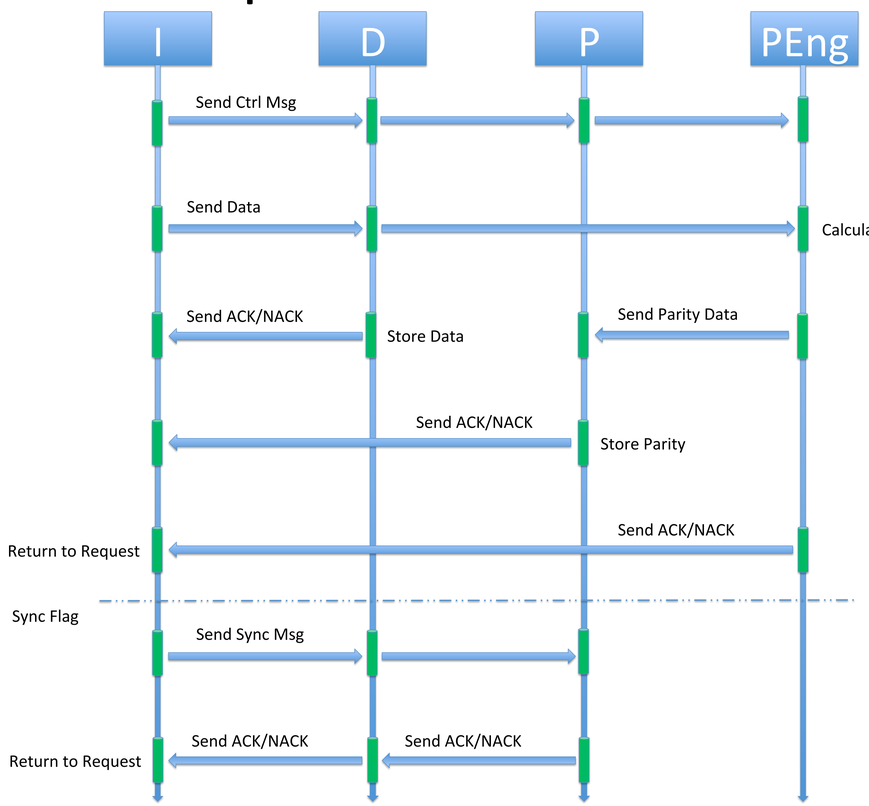

Figure 4: Store operation sequence.

Motivation

Construct a viable alternative RAID architecture that eliminates the RAID controller in favor of a consensus architecture. This would provide a more flexible system. Additionally we want to facilitate easy checking of parity on retrieve operations.

Throughput

Measured with test: write one GB of data, read one GB of data, timing each separately. Test configuration: two data disks and a single parity disk.

		Write (MB/S)	Read (MB/S)
No Parity	No Sync	94.18	69.66
No Parity	Sync	63.21	71.61
Parity	No Sync	65.64	50.69
Parity	Sync	48.08	50.95

Figure 5: Throughput results

Conclusions and Future Work

Making state transparent to the system's components and moving decision logic to the components allows for a controller-less RAID solution. Our prototype only scratched the surface of what is possible. Next, exploration of resiliency, atomicity of operations, scalability and other fabrics as well as a kernel block driver for accessibility.

References

[1] J. Hruska. Hdd pricewatch: Higher prices are the new normal, May 2012.
[2] E. K. Lee and R. H. Katz. Performance consequences of parity placement in disk arrays. SIGOPS Oper. Syst. Rev., 25(Special Issue):190–199, Apr. 1991.
[3] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks (raid). *SIGMOD REC.*, 17(3) 109-116, June 1988
[4] D. S. H. Rosenthal. Keeping bits safe: how hard can it be? Commun. ACM, 53(11):47–55, Nov. 2010.

Acknowledgements

Sandia National Laboratories is a multi-program 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.

Additional funding for the UAB authors was provided by the Department of Computer and Information Sciences at the University of Alabama at Birmingham.

Sandia National Laboratories

uab.edu