Fast Recovery Using Optimal and Near-Optimal Parallelism in Data-Intensive Computing

Dr. Jun Wang, Huijun Zhu, Pengju Shang, Peng Gu, Christopher Mitchell
University of Central Florida, School of Electrical Engineering and Computer Science - Computer Architecture and Storage Systems Group

Abstract and Motivation

With the advent of large-scale data clusters and their associated deployment of hard drives, reliability has become a major consideration in the design and development of these systems. Fortunately, the release of drives that have broken the 1 TB capacity limit and beyond lessens the concern for storage efficiency in these systems and thus paves the way for replication, not parity, to become the favored manner in which to ensure data integrity and system reliability. The switch from parity to multi-way replication is being further encouraged by the wide adoption by such mission critical systems as Google's File System (GFS), projects using Apache's Hadoop File System (HDFS), video on demand services, and Geographic Information Systems (GIS). To successfully develop replication placement schemes for use in these systems, researchers have developed several data layouts such as mirroring, chained declustering, group-rotational declustering, random declustering, and our-development in ICS'08: shifted declustering. Each data layout introduces a tradeoff between performance and reliability, a tradeoff that no research has as of yet attempted to quantify.

Prior research in replication was limited to studies of up to two replicas due to a prior lack of need or feasibility for implementing three or more replicas resulting from the smaller capacity drives and smaller scale disk arrays prevalent at the time. As previously stated, larger drives and larger storage systems are driving the number of needed replicas up thus causing us to reexamine the existing models and generalize them to work for k-way replication rather than a max of 2-way. We start by classifying the various replication data layouts as either type I (mirroring, as well as chained and group-rotational declustering) or type II (random and shifted declustering) with the difference being how we perform the calculation of the probability of loosing data on a given set of disks with of these disks failed. In type I replication layouts, this probability can be stochastically determined while in type II replication layouts, the value must be determined through random sampling of simulation data. By generalizing the models of the replication systems, we could then proceed to analyze each data layout scheme's overall reliability. We conclude that when used with a parallel recovery system, shifted declustering consistently outperforms the other layouts in terms of reliability when the same reserved recovery bandwidth is selected.

In addition to the benefits incurred in terms of reliability with replication based systems, there is an associated performance gain as a result of the increased parallelism available from these systems due to the inherent ability to read from the original as well as the copies simultaneously. However, the degree of parallelism is dependent upon the data layout in use by the storage system. The shifted declustering layout scheme is capable of leveraging the maximum degree of parallelism for the number of replicas produced and therefore is able to not only admirably perform during normal use cases but also during degraded modes when a disk failure has occurred and repairs have to be / are being made to the storage system. Thus, shifted declustering not only provides reliability for mission critical data centers but delivers it without compromising performance.

Overview of the Shifted Declustering Data Layout Scheme

The shifted declustering layout obtains optimal parallelism in a wide range of configurations. Flexibility in the system allows for any choice of available disks and number of replicas.

Shifted declustering is designed with the following properties in mind:
1.) Distributed reconstruction which balances the workload under degraded operating conditions.
2.) Maximal parallelism which ensures optimal performance during normal operating conditions.

Shifted declustering is inspired by chained declustering, which delivers maximal parallelism but not distributed reconstruction, because only neighboring disks can shoulder the workload from failed disks and as a result, becomes a performance bottleneck under degraded mode. This is due to its layout scheme which calls for distributing replicas strictly to consecutive disks. In response, the disk distances between replicas are expanded, one per iteration of the redundancy group number, to guarantee that all surviving disks share the workload resulting from a failed disk placing the system into degraded mode.

Example layout with 9 disks, and four redundancy groups:

q = 4			Disk 0	Disk 1	Disk 2	Disk 3	Disk 4	Disk 5	Disk 6	Disk 7	Disk 8	
	z = 0	i = 0	(0, 0)	(1, 0)	(2, 0)	(3, 0)	(4, 0)	(5, 0)	(6, 0)	(7, 0)	(8, 0)	offset = 0
		i = 1	(8, 1)	(0, 1)	(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)	(7, 1)	offset = 1
		i = 2	(7, 2)	(8, 2)	(0, 2)	(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)	offset = 2
	z = 1	i = 0	(9, 0)	(10, 0)	(11, 0)	(12, 0)	(13, 0)	(14, 0)	(15, 0)	(16, 0)	(17, 0)	offset = 3
		i = 1	(16, 1)	(17, 1)	(9, 1)	(10, 1)	(11, 1)	(12, 1)	(13, 1)	(14, 1)	(15, 1)	offset = 4
		i = 2	(14, 2)	(15, 2)	(16, 2)	(17, 2)	(9, 2)	(10, 2)	(11, 2)	(12, 2)	(13, 2)	offset = 5
	z = 2	i = 0	(18, 0)	(19, 0)	(20, 0)	(21, 0)	(22, 0)	(23, 0)	(24, 0)	(25, 0)	(26, 0)	offset = 6
		i = 1	(24, 1)	(25, 1)	(26, 1)	(18, 1)	(19, 1)	(20, 1)	(21, 1)	(22, 1)	(23, 1)	offset = 7
		i = 2	(21, 2)	(22, 2)	(23, 2)	(24, 2)	(25, 2)	(26, 2)	(18, 2)	(19, 2)	(20, 2)	offset = 8
	z = 3	i = 0	(27, 0)	(28, 0)	(29, 0)	(30, 0)	(31, 0)	(32, 0)	(33, 0)	(34, 0)	(35, 0)	offset = 9
		i = 1	(32, 1)	(33, 1)	(34, 1)	(35, 1)	(27, 1)	(28, 1)	(29, 1)	(30, 1)	(31, 1)	offset = 10
		i = 2	(28, 2)	(29, 2)	(30, 2)	(31, 2)	(32, 2)	(33, 2)	(34, 2)	(35, 2)	(27, 2)	offset = 11

a = 33
i = 2
disk(33, 2) = 5
offset (33, 2) = 11

Thus, shifted declustering is designed such that the following holds:

$$q = \begin{cases} 1, & \text{if } n = 4 \\ (n-1)/2, & \text{if } n \text{ is odd} \end{cases} \quad (1)$$

$$z = \left\lfloor \frac{a}{n} \right\rfloor \quad (2)$$

$$y = (z\%q) + 1 \quad (3)$$

$$\mathbf{disk}(a,i) = (a + iy)\%n \quad (4)$$

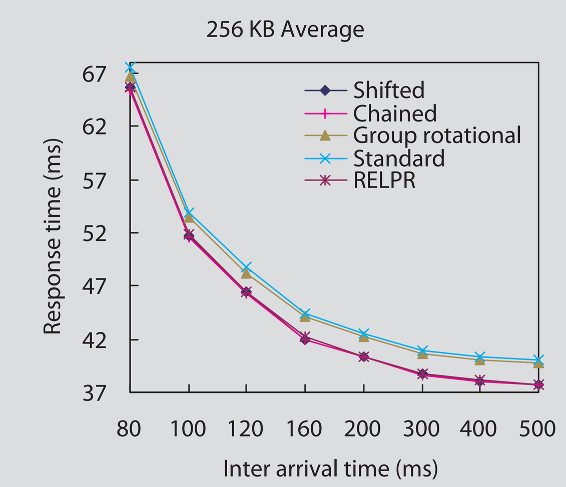
$$\mathbf{offset}(a,i) = \left\lfloor \frac{a}{n} \right\rfloor + (k-1)z + i = kz + i \quad (5)$$

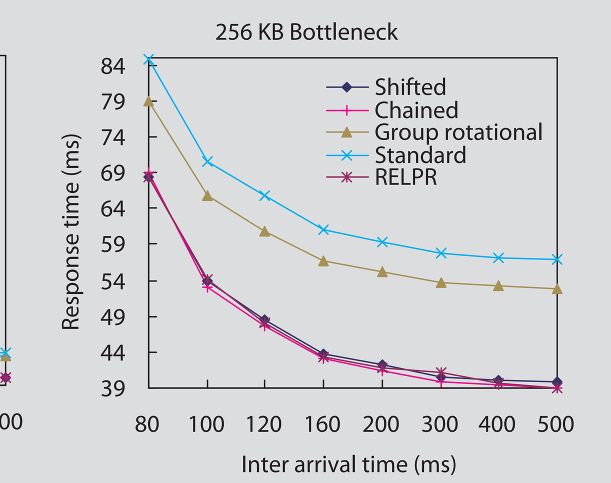
TABLE I
NOTATION SUMMARY

System configuration parameters	
n	Number of disks in the cluster
k	Number of units per redundancy group
Parameters used in computation	
a	The address to denote a redundancy group
(a, i)	The i-th unit in redundancy group a
q	Number of iterations of a complete round of layout
y, z	Intermediate auxiliary parameters
Computation output	
$disk(a, i)$	The disk where the unit (a, i) is distributed
$offset(a, i)$	The offset within $disk(a, i)$ where the unit (a, i) is distributed

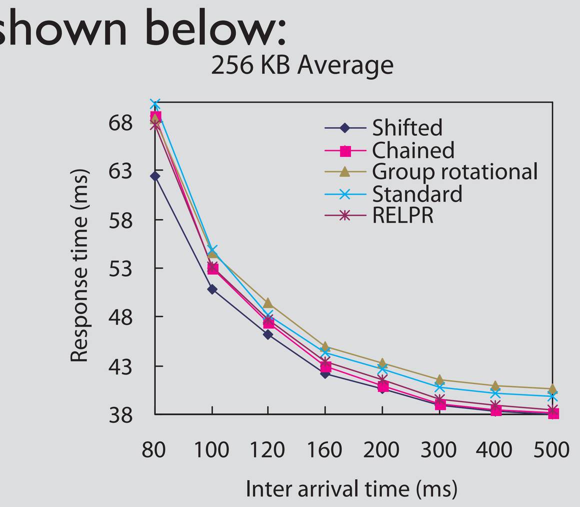
Performance Evaluation

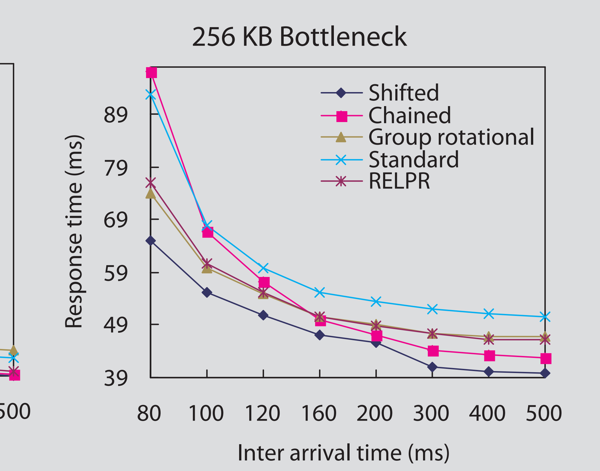
When operating in a normal operating environment, shifted declustering provides comparable performance with chained declustering, and both outperformed all other data layout schemes. The performance itself was measured in terms of response time and is illustrated in the two result graphs below:





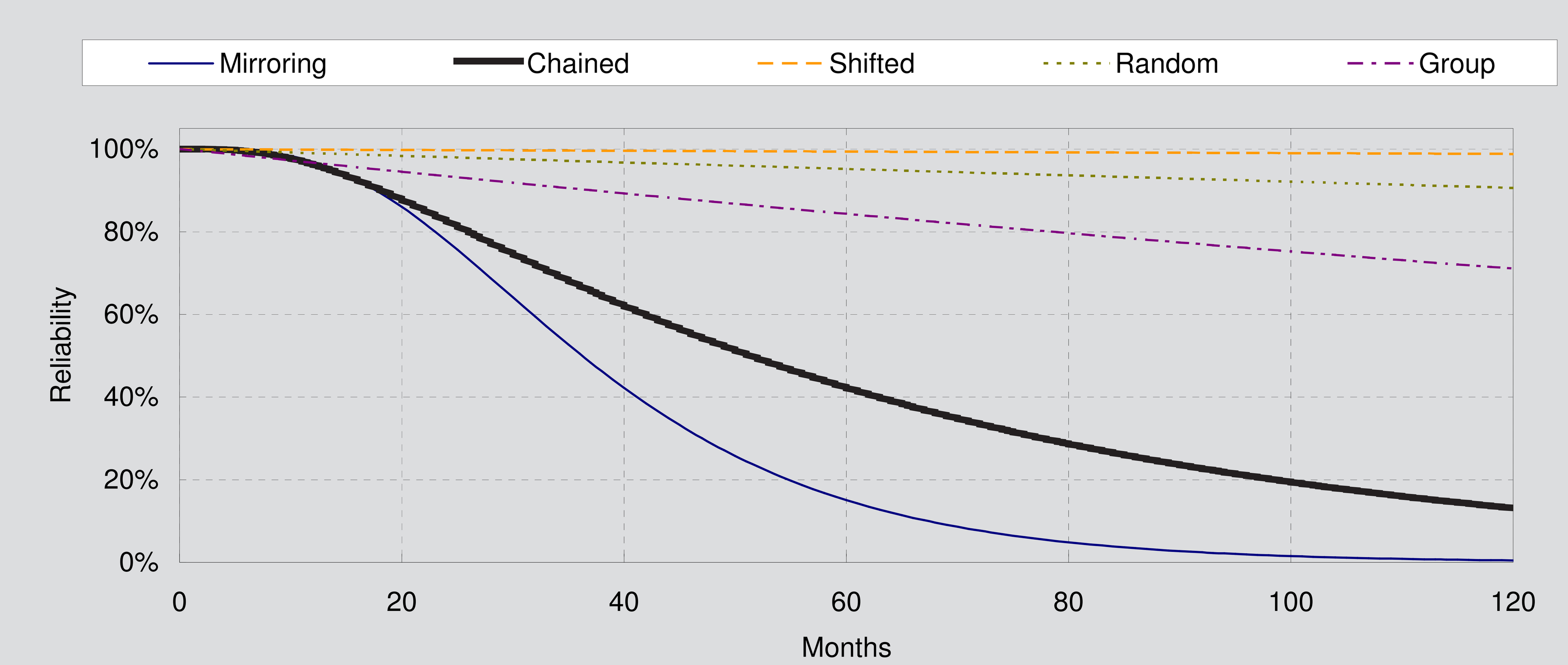
When operating in degraded mode, shifted declustering outperforms all other layouts in terms of both overall performance and bottleneck performance (performance of the slowest disk). This is illustrated in the two result graphs shown below:





Reliability Analysis

When examining the reliability of mirroring, group-rotational declustering, chained declustering, shifted declustering, and random declustering, we assume an aggressive parallel recovery scheme will be in use for all recovery options. During testing, a 10 KB/sec cap is applied for the recovery bandwidth used per disk. In this setup, shifted declustering has the highest reliability compared to all other schemes. Additionally, for all other schemes to match shifted declustering, more recovery bandwidth must be used per drive which impacts the performance of normal service requests that are also occurring at the same time. The diagram below illustrates the system reliability when 10 KB/sec is used per disk for recovery bandwidth and as shown shifted declustering maintains the highest reliability rating even as other schemes start to sharply drop off.



This work is supported in part by the US National Science Foundation under grants CNS-0646910, CNS-0646911, CCF-0621526, CCF-0811413, and the US Department of Energy Early Career Principal Investigator Award DE-FG02-07ER25747.

For Further Information, please see our two published works:

Huijun Zhu, Peng Gu, and Jun Wang. Shifted Declustering: An Ideal-placement Layout Scheme for Multi-way Replication Storage Architecture. The 22nd ACM International Conference on Supercomputing (ICS2008). June 7-12, 2008 Island of Kos, Aegean Sea, Greece.
and
Huijun Zhu, Pengju Shang, Christopher Mitchell, and Jun Wang: Reliability Analysis on Multi-way Replication Data Layouts in Large-scale Storage Systems. Work in Progess

Office of Science
U.S. DEPARTMENT OF ENERGY

UCF
Stands For Opportunity