Future Generation Computer Systems 29 (2013) 509–519

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A highly reliable and parallelizable data distribution scheme for data grids

Javad Akbari Torkestani

Department of Computer Engineering, Arak Branch, Islamic Azad University, Arak, Iran

ARTICLE INFO

Article history:
Received 18 March 2012
Received in revised form
22 June 2012
Accepted 14 July 2012
Available online 6 August 2012

Keywords:
Data Grid system
RAID
Data replication
Mirroring
Striping
Reliability

ABSTRACT

The major drawback of the replication-based RAID (redundant arrays of independent disks) architectures is that, in spite of the high redundancy level, they cannot balance the load increased by the disk failures and this results in reliability and access bandwidth reduction for processing the data access requests. Furthermore, these schemes are not able to determine the actual position of the occurring data block errors (or in-error data blocks). In this paper, to alleviate the addressed problems, we propose a new parity-based striped mirroring scheme called PSM-RAID in which the striped data blocks are replicated among the other disks, and a parity block is then associated with each stripe. In this method, a data mirroring scheme improves the reliability of the array by directing the read requests to the mirrored copy of data, and the parity data enables the controller to determine the actual position of the in-error blocks. The specific data distribution algorithm proposed in this paper also improves the access bandwidth of the array to serve many more disk requests. The proposed method is compared with similar architectures and the simulation results show that the proposed model due to the data striping scheme is able to provide a significantly higher parallelism between the disk requests, as well as a higher reliability due to the block mirroring scheme and the parity blocks.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

A data grid is a specific grid system providing users with a huge amount of storage space, and maintaining a high volume of distributed data to serve users. Generally, data grids can be classified into multi-tier data grids and cluster data grids [1]. Tremendous data size, heterogeneous structure and wide-area distributed data storage sites are the main characteristics of a data Grid system [2]. Due to huge amount of heterogeneous and distributed data in data Grid systems, managing the system to reduce access latency, to improve data locality, and to increase data reliability and availability is of great importance. Data replication used in RAID systems is a common approach to improve data availability, reliability, and to shorten the data access time. Data Grids demand new techniques to efficiently access and process the large-scale distributed data volumes. The replica location problem, i.e., seeking the best location of each copy to maximize the system performance, is one of the hot research topics in data grids. Replication-based RAID techniques in combination with data-stripping mechanisms are promising approaches to the replica location problem.

A RAID is an I/O subsystem incorporating multiple independent small disks into a considerably larger logical disk to improve the I/O performance. Disk arrays are capable of providing higher effective bandwidth to access data due to involving several independent disks to serve I/O requests. Furthermore, disk arrays, due to the redundant information, are capable of tolerating disk failures, thus providing higher reliability, as well as data availability. RAID technology is an efficient way to bridge the gap between the speed of the processor and disk access rate [3]. The tremendous growth of RAID systems has been driven by the three following important factors. First, the growth of processor speed has outstripped the growth of disk data rate. This imbalance traditionally transforms computer-bound applications to I/O-bound applications. Therefore, I/O system throughput needs to be improved by increasing the effective bandwidth to access data. Second, arrays of independent small disks often have substantial capacity and performance advantages over single large disks. Third, such systems can be made highly reliable by storing a small amount of redundant information in the array [4–13].

The data encoding procedure, mapping function and data access algorithm are three important factors in designing a RAID system, by which different RAID architectures can be distinguished and classified into five levels [14–17]. The encoding procedure specifies the type of redundancy information employed to encode the data. The mapping function introduces the pattern used to place the primary data and redundant information on the disk array. The algorithms which are used to access data can be classified as normal mode and failed mode. In normal mode, either there is no failure in the disk array or the controller knows about the failed disks, if any. In a failed mode, a disk failure has occurred in the middle of the controller operation. The controller then needs to

E-mail address: j-akbari@iau-arak.ac.ir.

0167-739X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2012.07.006

recover the error and complete the operation. This process is called error recovery [18,19,15].

A RAID system was first proposed by Patterson et al. [14] to improve the reliability and I/O performance of the independent small disks. During the recent decades, several studies have been conducted to evaluate the performance of RAID systems. Different aspects of the RAID system have been discussed in the literature. To name just a few, reliability [20–26], availability [27,21,28–30], scalability [31,32,13], energy–efficiency [33,34], and RAID modeling techniques [35,36]. Wu et al. [29] proposed an outscoring-based method to improve the availability of the RAID-structured storage systems. The proposed method improves the RAID performance by temporarily redirecting all the write requests and the popular read requests targeted at the surrogate RAID set. Surrogate RAID set is a set of spare disks or free space on another RAID set. Elerath and Pecht [24] proposed an approach to model and evaluate the reliability of the RAID architectures. Gibson et al. [14], Thomasian and Blaum [22], and Akbari and Meybodi [18,15] also studied and compared the reliability of different RAID architectures. Zhang et al. [31] presented a data redistribution approach called ALV for scaling (adding more disks to RAID-5 to have a larger storage space and higher I/O bandwidth) RAID-5. ALV exploits three techniques. First, changing the movement order of the data chunks to access multiple chunks by a single I/O. Second, updating the mapping metadata to minimize the number of metadata writes. Third, using a logical valve to adjust the redistribution rate depending on the application workload. A stripe-based orthogonal mirroring technique was proposed in [32] to boost the scalability and availability level of the distributed RAID architectures. In [15], a hybrid mirroring-based RAID architecture was proposed to improve the I/O performance of the distributed systems. Xie [33] (SEA) and Wang et al. [34] (eRAID) considered the energy–efficiency issues in RAID-structured data storage systems. On the basis of the fraction of small requests versus large ones, and fraction of updates versus read requests, Thomasian and Xu [37] presented a statistical model to select the RAID level tailored for a given RAID architecture.

Due to the critical problems with the basic disk mirroring scheme, such as low data transfer rate, low parallelism degree, and slow failure recovery, several studies have been conducted to improve the performance of the disk mirroring technique. Chen and Towsley [38] proposed the group-rotate-declustering technique, in which, similar to the basic mirroring scheme, the disk array is divided into primary and secondary sub-arrays. In this method, the data is striped and the stripe units of each primary disk are located on the secondary sub-array in a rotational manner. Akbari and Meybodi [19] proposed a striped mirroring scheme to improve the access bandwidth of the array. In this scheme, due to a diagonal data distribution algorithm, when a disk fails, the disk operations can be balanced among the surviving disks. In [15], the same authors presented a RAID architecture called RAID-RMS in which a special hybrid mechanism is used to map the data blocks to the cluster. The idea behind the proposed method is to combine the block striping, disk mirroring, and block rotation techniques to improve the parallelism of the RAID architecture. The interleaved declustering method proposed in [39] increases the probability of reconstructing the failed disks. In this method, the array of independent disks is partitioned into clusters. The mirrored copy of the primary data on each disk is evenly distributed among the other disks. An improvement of the interleaved declustering method [39] called chained declustering was presented by Hsiao and Dewitt [28]. The chained declustering method subdivides each disk into primary and secondary areas. In this method, the data is placed in the primary area of the disk array and its replication is copied to the secondary area. As compared to interleaved declustering, the chained declustering technique can tolerate a wider range of disk failures. In [40] a shift-based chained declustering method, hereafter called RAID-C, was proposed to shorten the data processing time and to improve the parallelism of the RAID system. Shifted-declustering is a replication-based RAID architecture that can be deployed in a wide variety of k-way replication configurations. It is a generalization of the chained-declustering technique from two-way replication to k-way replication. Chained-declustering simply shifts each row of the data units over the other disks in a circular fashion to enhance the parallelism.

Proportional to the amount of redundant information, different RAID architectures provide different levels of reliability and data availability for the disk array. Despite the large number of backup disks and huge amount of redundant information, the existing replication-based RAID architectures cannot provide a satisfactory level of reliability and I/O performance. The major deficiency of the current replication-based RAID architectures (in disk level) is that the load increased by the disk failures cannot be properly balanced among the surviving disks. This significantly reduces the reliability and access bandwidth of the system for processing the data queries. Furthermore, these schemes (in block level) lack the capability of determining the actual position of the data block errors. In this paper, a block-level mirroring scheme called PSM-RAID is proposed to overcome the above-mentioned shortcomings. In this scheme, the striped data blocks are replicated among the other disks, and a parity block is associated with each stripe. In PSM-RAID, every disk of the array is partitioned into primary and secondary areas [15] of the same size. PSM-RAID uses a hybrid data distribution procedure to locate the data, mirrored, and parity blocks. Block-level replication and distribution significantly improve the access bandwidth and data availability of PSM-RAID. The parity blocks distributed among the array provide a higher access and recovery rate and enable the RAID controller to determine in-error blocks. To show the performance of the proposed RAID architecture, several simulation experiments are conducted in the DiskSim simulation environment (v. 4.0) [41]. The obtained results are compared with those of RAID-10, RAID-C and RAID-5 in terms of reliability, access bandwidth, and I/O performance. Simulation experiments confirm the higher parallelism of the proposed RAID system due to using the data striping scheme, and its higher reliability due to using the block-level mirroring scheme and distributed parity blocks. The rest of the paper is organized as follows. In the next section, the RAID system is briefly reviewed. The proposed RAID architecture is described in Section 3. This section also theoretically analyzes the reliability and cost of the proposed method. In Section 4, the proposed RAID system is evaluated through simulation experiments and the obtained results are compared with those of similar architectures. Section 5 concludes the paper.

2. RAID-definition and methods

A RAID system is an I/O subsystem comprising a large number of independent small disks by which a logical large storage system as one unit is represented. It is considered superior to a single large expensive disk in terms of data transfer rate, reliability, data availability and capacity. Let N and n denote the block sequence number in the memory address space and the number of disks in the cluster, respectively. m denotes the number of blocks per disk, each of size S, and i and j are indices to specify the horizontal and vertical positions of a data or parity blocks, respectively. Formally, a RAID can be described by a quintuple $\langle \tilde{A}, \underline{B}, \underline{R}, f_B, g_R \rangle$, where $\tilde{A} = \{a_0, a_1, \ldots, a_N\}$ denotes the sequence of the data blocks, $\underline{B} = \{b_{(i,j)} | 0 \leq i \leq m-1, 0 \leq j \leq n-1\}$ is the finite set of data blocks in the disk array, $\underline{R} = \Phi \bigcup \{r_{(i,j)} | 0 \leq i \leq m-1, 0 \leq$

Download English Version:

https://daneshyari.com/en/article/10330571

Download Persian Version:

https://daneshyari.com/article/10330571

Daneshyari.com