Collaboration-Oriented Data Recovery for Mobile Disk Arrays

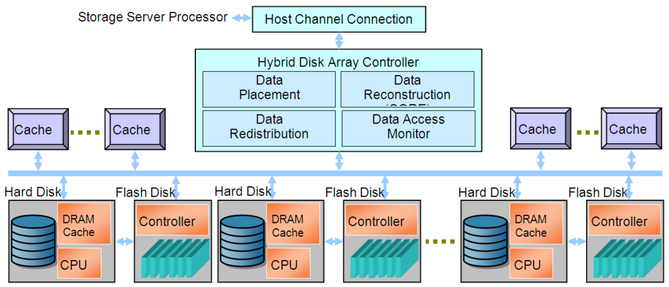
The 29th International Conference on Distributed Computing Systems, 2009

Tao Xie and Abhinav Sharma, Computer Science Department, San Diego State University

Introduction

Mobile data centers are an alternative to conventional stationary data centers that are enclosed in buildings. They could be built on self-contained trucks, airplanes, or ships that have onboard generators, UPS, multiple high-capacity servers, and satellite Internet links. For example, an NAAT MED (Mobile Emergency Datacenter) can accommodate up to 100 fully charged laptops, multiple high-performance servers, and a large capacity storage system with multiple Terabytes of data in a 20-25ft truck. Typical applications for mobile data centers include disaster recovery, live video broadcast, and homeland security, where high mobility and a fast large-volume data processing capability are intrinsically demanded. Apparently, mobile disk arrays are essential in these emergency-oriented applications because they can provide not only huge storage capacities but also high bandwidth. At present, a mobile disk array generally consists of an array of independent small form factor hard disks connected to a host by a storage interface like SAS (Serial-attached SCSI). Due to their unusual application domains, mobile disk arrays face several new challenges including harsh operating environments, very limited power supply, and extremely small number of spare disks. Consequently, data reconstruction schemes for mobile disk arrays must be performance-driven, reliability-aware, and energy-efficient. In this paper, we develop a flash assisted data reconstruction strategy called CORE (collaboration-oriented reconstruction) on top of a hybrid disk array architecture, where hard disks and flash disks collaborate to shorten data reconstruction time, alleviate performance degradation during disk recovery. Experimental results demonstrate that CORE noticeably improves the performance and energy-efficiency over existing schemes.

The hybrid disk array architecture



we propose a hybrid disk array architecture called FIT (*f*lash-assisted d*i*sk s*t*orage). Within the FIT architecture, hard disks are organized in some RAID structure such as RAID-5. Similarly, all flash disks are arranged in the same RAID structure as the hard disk array. Both hard disks and flash disks are directly attached to the system bus. Further, each flash disk cooperates with a hard disk to compose a buddy-pair through a dedicated high-bandwidth connection.

Within the hybrid disk array controller, there are four software modules that manage the data across the hybrid disk array and the controller caches. The data placement module places all newly arrived data on the hard disk array. The data access monitor dynamically records each disk zone's number of accesses, and then provides the data redistribution module with a data redistribution table. Based on the data redistribution table, the data redistribution module reallocates popular mostly-read onto flash disks. When a flash disk or a hard disk fails, the data reconstruction module (CORE) is launched to provide a fault-tolerant mechanism for the hybrid disk array.

The core strategy

Reconstruction Data Grabber

1. Sort all zones with Location 0 in DPT into a list *H* in a descending order in their total number of accesses
2. **for** each zone z_i starting from the first one in *H* **do**
3. **Repeat**
4. Issue a low-priority request to read a stripe into a buffer
5. Wait for the read request to complete
6. Submit the unit data to a centralized buffer manager for XOR, or block the process if the buffer is full
7. **Until** (all units of z_i in this disk have been read)
8. **end for**

Reconstructed Data Restorer

1. **Repeat**
2. Request the next full buffer from the buffer manager, blocking itself if none is available
3. Issue a low-priority write to the replacement disk.
4. Wait for the write to complete.
5. **Until** (the failed disk has been reconstructed)

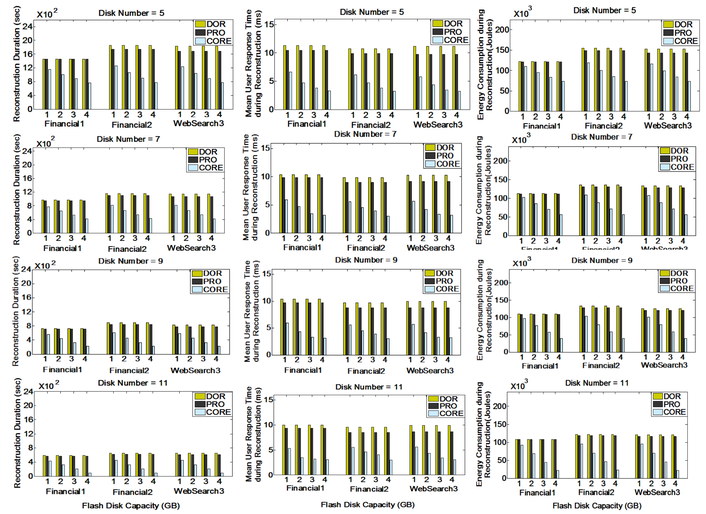
Left figure shows the workflows of the two routines of CORE: Reconstruction Data Grabber and Reconstructed Data Restorer. CORE optimizes the reconstruction workflow by fetching reconstruction data of popular stripe units from the failed disk prior to fetching reconstruction data of unpopular stripe units, a similar idea used in. In fact, the hybrid disk array controller creates n-1 processes called Reconstruction Data Grabber (RDG). Each RDG process associates with one surviving hard disk. Also, a process named Reconstructed Data Restorer (RDR) is launched in the hybrid disk array controller to write the reconstructed data onto the replacement hard disk. The functions of RDG and RDR are similar as those of the DOR algorithm except for the following difference: A RDG process always selects the next most popular "under construction" unit rather than choosing next sequential unit as the DOR algorithm. While This figure only demonstrates how CORE rebuilds data for a failed hard disk, CORE is capable of reconstructing data for flash disks as well. In fact, when a flash disk recovery starts, CORE rebuilds data on a replacement flash disk in the same manner as it does for hard disk recovery.

Performance evaluation

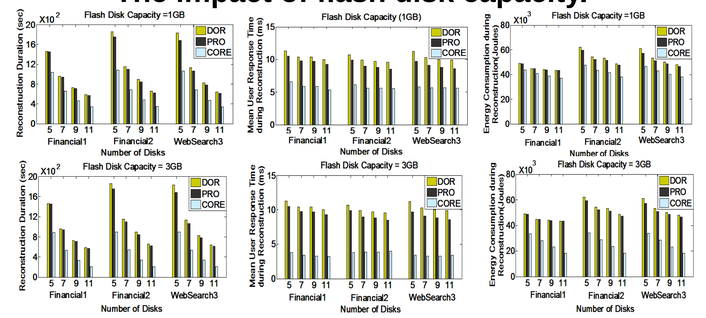
Hard disk	Seagate Cheetah 15K.4	Flash disk	Adtron A25FB-20 Flashpak
Capacity (GB)	73.4	Capacity (GB)	1, 2, 3, 4
Spindle speed (RPM)	15 K	Access time (ms)	0.272
Ave. seek time (ms)	3.5	Seek time	0
Ave. latency (ms)	2.0	Read (Mbytes/sec)	78
Transfer rate (Mbytes/sec)	77	Write (Mbytes/sec)	47
Active power (watts)	17	Read/write power (watts)	3.43
Idle power (watts)	11.9	Idle power (watts)	1.91

Simulation parameters

We developed a trace-driven simulator FITSim that models a hybrid disk array, which has one hard disk array and one flash disk array. Each disk array is made up of *m* disks organized in a RAID-5 structure. For hard disk, FITSim uses the parameters of the Seagate Cheetah 15K.4 73.4 GB. For flash disk, it adopts the specifications of the Adtron A25FB-20 Flashpak and the capacity varies from 1 GB to 4 GB with 3 GB as the default value. The main characteristics of the hard disk and the flash disk used by FITSim are shown in the table. Right figures show the performance of CORE and other two conventional strategies.



The impact of flash disk capacity.



The impact of number of disks.

Conclusions

In this paper, CORE is developed on top of a hybrid disk array architecture called FIT. Extensive experiments using real-world traces show that CORE significantly improves the performance in terms of reconstruction duration and mean response time during reconstruction over two baseline data reconstruction schemes. There are two main contributions of this paper. First, the idea of integrating flash disks into traditional hard-disk based disk arrays to not only substantially improve data reconstruction performance but also enhance mobile disk array reliability. Second, CORE also noticeably enhances energy-efficiency, which is critical for mobile data centers.

Acknowledgment

This work is supported by the US National Science Foundation under grants CNS-0834466 and CCF-0742187.