iATA 기반의 RAID5 분산 스토리지 서버의 설계 및 구현

왕숙미* · 임효택**

Design and Implementation of iATA-based RAID5 Distributed Storage Servers

Ivy Ong* · Hyotaek Lim**

요 약

iATA는 TCP/IP 네트워크상에서 ATA 명령어를 전달하기 위해 개발된 블록-레벨 프로토콜로서, 모바일 기기의 스토리지 한계를 극복하기 위한 대안으로 활용 될 수 있다. 본 논문은 RAID5 분산 스토리지 서버 개념을 iATA에 적용하여 스토리지 서버의 신뢰성과 속도를 개선하고자 한다. 분산 스토리지 서버중 하나의 서버가 다운된 경우에 나머지 서버 데이터의 XOR 함수를 적용하여 데이터 회복이 가능하며 이를 통해 데이터의 신뢰성을 높일 수 있다. 벤치마킹 실험과 시험을 통해 제안된 iATA 프로토콜은 제한된 스토리지를 가지고 있는 모바일 기기상에서 효율적이고도 신뢰성 있는 가상 스토리지 프로토콜로서 사용될 수 있음을 보여주고 있다.

ABSTRACT

iATA (Internet Advanced Technology Attachment) is a block-level protocol developed to transfer ATA commands over TCP/IP network, as an alternative network storage solution to address insufficient storage problem in mobile devices. This paper employs RAID5 distributed storage servers concept into iATA, in which the idea behind is to combine several machines with relatively inexpensive disk drives into a server array that works as a single virtual storage device, thus increasing the reliability and speed of operations. In the case of one machine failed, the server array will not destroy immediately but able to function in a degradation mode. Meanwhile, information can be easily recovered by using boolean exclusive OR (XOR) logical function with the bit information on the remaining machines. We perform I/O measurement and benchmark tool result indicates that additional fault tolerance feature does not delay read/write operations with reasonable file size ranged in 4KB-2MB, yet higher data integrity objective is achieved.

키워드

iATA, TCP/IP, 모바일 컴퓨팅, RAID5, 분산 스토리지 서버

Key word

iATA, TCP/IP, Mobile Computing, RAID5, Distributed Storage Servers

* 동서대학교 일반대학원 석사과정
** 동서대학교 컴퓨터정보공학부 교수(교신저자)

접수일자 : 2009. 10. 01
심사완료일자 : 2009. 11. 30

Ⅰ. Introduction

In recent decades, mobile technology has been growing dramatically and portable mobile devices such as PDA (Personal Digital Assistant) are widely used by corporations or individual users to access information timely. We designed a relatively cost-effective block-level protocol, iATA [12] [13] [14] to fulfill the high demand of large storage spaces required in mobile devices consequences of the rapid growth of intensive mobile applications and network services.

This paper presents the concept of distributed servers' storage with RAID (Redundant Array of Independent Disks) technology in iATA to improve its current functionality. In old practices, a PDA might have a few hundred Megabytes of internal storage spaces with a few Gigabytes of external memory card. Nevertheless, users are at risk when the internal storage spaces crashed or the external memory card broken, causing those valuable confidential data permanently lose. The novel RAID5 in iATA implementation not only benefits regular mobile users from acquiring virtual storages for typical entertainment purposes such as video or audio multimedia data streaming, but also allow corporate employees access to remote server storages to manipulate company or business data, without time and location boundary obstacles, yet in a safer way with the existence of fault tolerance feature.

In traditional RAID, a disk from a standalone server is the storage unit, however in iATA distributed RAID5 scheme, each server machine itself is a basic storage unit that builds up the server array. With this approach, both fault tolerance and parallelism access properties are achieved to increase data reliability and I/O performance. The rest of the paper is organized as follows. Section II outlines the related research works. Section III describes the design and implementation of iATA protocol with RAID5 distributed storage servers. We evaluate the functionalities in Section IV and conclude the paper in last Section V.

Ⅱ. Related Research Works

Generally, RAID is known as two or more disks to be arranged into an array, where data is split or replicated across the array accordingly to diverse level designs. The RAID concept was extended into a distributed computing environment and presented by Stonebraker and Gerhard [6] in 1990. Their RADD model supports redundant copies of data over a computer network with high availability and assistance for site failures or disasters. Furthermore, Expand [5] applies distributed RAID in NFS (Network File System) servers to increase system scalability and obtain fault tolerance function. It allows heterogeneous servers (Linux, Solaris, Windows 2000, etc.) join to create a distributed partition where files are striped in cyclic layout or RAID4/RAID5 patterns. The new Swift/RAID prototype [2] is the enhancement of the originally Swift prototype, which was firstly developed in 1991. This new approach includes RAID4/RAID5 in a distributed environment with the main purpose of getting fault tolerance benefit. Besides, S-iRAID and P-iRAID [4] were proposed to improve iSCSI performance and reliability with distributed RAID0/RAID5 configurations. Another iSCSI based iVISA [7] is an interesting framework that applied the concepts of distributed RAID and virtualization storage. It provides a dynamic block-level storage pool based on the storage nodes current workload and data layout to avoid synchronization overhead problem seen in conventional distributed RAID systems.

Among various levels, RAID5 is the most ideal level to be used with iATA due to its attractive fault tolerance and parallelism characteristics. Compare to other levels such as RAID0 or RAID0+1, RAID5 provides low cost of redundancy in which the formal gives s/2 storage capacity while the latter gives s*(n-1)/n storage capacity, where s is the sum of storage spacge capn drives been used [1ra. Besides, it over s es the sloweer ata transaeerrate due to the dedicated type capparity in RAID4 [1a. Data blocks are stripped into chunks and stored across all membeer isks as well as parity blocks that are calculated using XOR logical

cpera well an oua. cenario, iATA data blocks are chunked at the host e calc before distributed to targetre calcu, including itselfl an the case of he machine down or s illocly des,ed, the s calc array is not destroy whiutrets,l able toks as wel in a degradation lats wel igrio, adv/n age not hely s c it oine dxisting data bus pasokprovides an dasier interfacg and steps arereplacg the fes,ed machineludingdegradatios much efal givereliable compare to the traditional way of using external memory cards in mobile devices, where data may permanently lose or destroy due to physical damage or missing. We use Linux software RAID featuring stable mdadm tool to manage and monitor iATA distributed RAID5 medium.

III. iATA Protocol and RAID 5 Distributed Storage Servers

Technologically, there are several block-level protocols available in the market such as the popular iSCSI (Internet Small Computer Interface) and AoE (ATA over Ethernet) protocols, which normally used in SAN (Storage Area Network) to provide a simple and less performance overhead interface for nonvolatile magnetic media manipulating. In contrast of using the conventional routable iSCSI protocol and not routable AoE protocol, we introduce a new protocol, iATA for mobile users to obtain more storage spaces with lower cost and easier configuration settings.

3.1. iATA Protocol Overview

iATA protocol is an application layer protocol that converts ATA commands into iATA PDU (Protocol Data Unit) before it is routed over TCP/IP network from a client PDA to a server machine. It allows a server ATA hard disk to be mounted remotely from a PDA through wireless connection for read and write operations. When the iATA PDU is received at the server side, it will be converted back to ATA commands before proceed to upper processes. Figure 1 depicts that iATA protocol lies on the top of Transport Layer in the conceptual TCP/IP Layering Model.

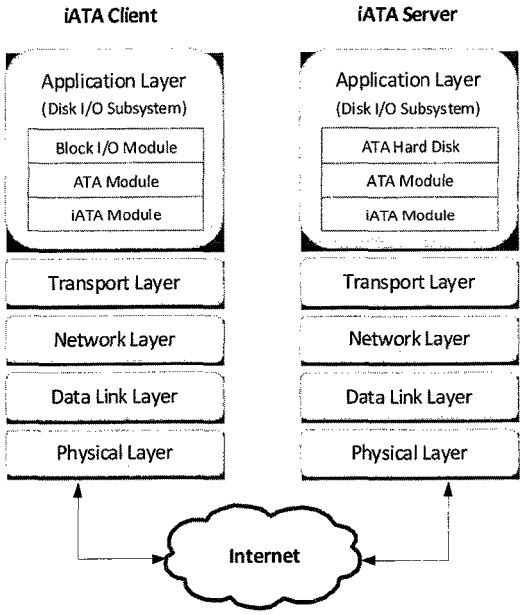

그림 1. iATA 프로토콜 스택
Fig. 1. iATA Protocol Stack

3.2. iATA RAID5 Distributed Storage Servers

To enhance the current functionality of iATA protocol, RAID5 distributed storage servers idea is embedded into iATA client-server architecture as illustrated in Figure 2. One machine is served as the host server, which plays the role to communicate with a client PDA. Three machines are served as the target servers, which are interconnected to each other to form an isolated LAN (Local Area Network) with Fast Ethernet technology, to build up a server array.

The distributed RAID5 configuration is done at the host server of array before it is ready to be mounted. Upon successfully connected, PDA user can manipulate read/write operations from/to the virtual storage, such that it is seen as a single local device but in fact those data saved in/retrieved out are relatively connected to four machines.

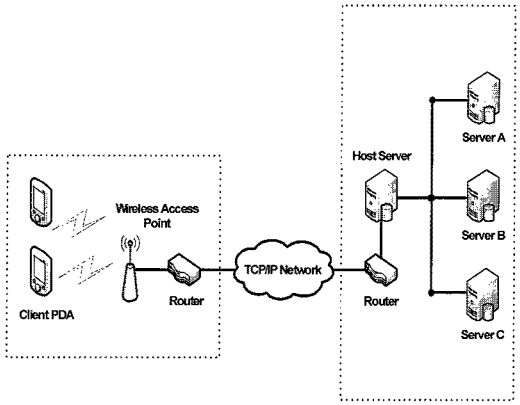

그림 2. iATA 클라이언트-서버 구조
Fig. 2. iATA New Client-Server Architecture

Technically when data reaches to the host server, they are chunked into specific size before distributed across the array. The parity block for each stripe is calculated and rotated to store in each machine in a cyclic pattern. To be simplify, it is notated as Parity1-i = D1 XOR D2 XOR D3 ··· XOR Di. Figure 3 illustrates RAID5 Left-Synchronous data and parity blocks organization.

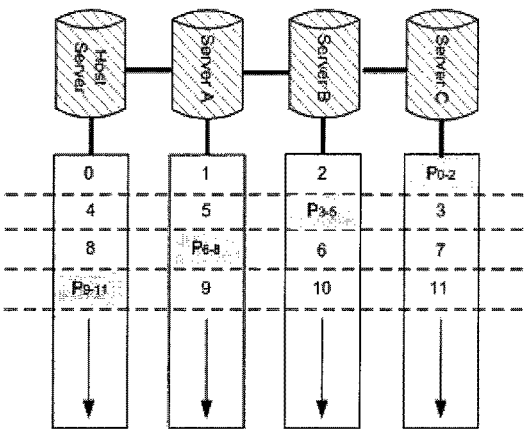

그림 3. RAID5 좌측-동기 데이터 및 패리티 블록구조
Fig. 3. RAID5 Left-Synchronous Data and Parity Blocks Organization

3.3. Platform Setup and Configuration

To set up the environment, we use four personal computers PCs, a PDA, a wireless access point and routers. The PCs and PDA specifications are listed separately in Table 1 and Table 2.

표 1. 서버 사양
Table. 1. Server Machines Specifications

Parameter	Value
Processor	Intel(R)Core(TM)2Duo2.53GHz
RAM	256MB
Hard Disk Capacity	8GB (IDE)
Operating System	Fedora Core 6 (Linux Kernel 2.6.18)

표 2. 클라이언트 PDA 사양
Table. 2. Client PDA Specifications

Parameter	Value
PDA	HP iPAQ hx2700
Processor	Marvell PXA270 624MHz
RAM	64MB
ROM	320MB
Secure Digital (SD) Card	4GB
Operating System	Microsoft® Windows Mobile® 5.0

We utilize Linux software RAID and mdadm tool to create and maintain the array. For instance, each PCs exported 8GB of storage spaces to make up a total of 24GB available spaces for use. We change the default chunk size of 64KB to 128KB, which is considered reasonable to accommodate RAID level 5. The data layout remains in Left-Synchronous mode, which is fine used to support large reads or writes. While at client side, the number of sectors per command (NSPC) is set to 2048 for both read and write requests, which is analysed able to achieve the optimum performance in iATA protocol [12].

Ⅳ. Experiment Methodology

4.1. Fault Tolerance Evaluation

Summarized below are steps used to assess the data recovery scheme.

i. Create a distributed RAID5 server array and activate it as shown in Figure 4.

ii. Export the available spaces and mount it to a client PDA through TCP/IP network. Save it with a number of files. Subsequently, fail one machine and remove it from the list as shown in Figure 5.

iii. The array continues working without interruption in a degradation mode, data able to save/retrieve on/from the PDA.

```
/dev/md0:
        Version : 00.90.03
  Creation Time : Tue Sep 29 10:51:09 2009
     Raid Level : raid5
     Array Size : 25165440 (24.00 GiB 25.77 GB)
  Used Dev Size : 8388480 (8.00 GiB 8.59 GB)
   Raid Devices : 4
  Total Devices : 4
Preferred Minor : 0
    Persistence : Superblock is persistent

    Update Time : Tue Sep 29 11:06:05 2009
          State : clean
 Active Devices : 4
Working Devices : 4
 Failed Devices : 0
  Spare Devices : 0

         Layout : left-symmetric
     Chunk Size : 128K

           UUID : 74b1f432:23ddfb7a:db3f55a3:cbf3b659
         Events : 0.6
```

그림 4. 분산 RAID5 서버 배열
Fig. 4. A Distributed RAID5 Server Array

iv. Replace with a new machine and re-assemble the array. Figure 6 depicts the rebuilding array process is in progress.

v. No data lost from the entire assessment session. Figure 7 pictures the iATADisk Explorer where different kinds of files saved in the virtual disk.

```
[root@localhost ~]# mdadm /dev/md0 --set-faulty /dev/etherd/e0.0
mdadm: set /dev/etherd/e0.0 faulty in /dev/md0
[root@localhost ~]# mdadm /dev/md0 --remove /dev/etherd/e0.0
mdadm: hot removed /dev/etherd/e0.0
```

그림 5. 서버 배열에서 고장서버 제거
Fig. 5. Removal Faulty Machine from List

```
md0 : active raid5 etherd/e0.0[0] hdb[3] etherd/e2.0[2] etherd/e1.0[1]
      25165440 blocks level 5, 128k chunk, algorithm 2 [4/3] [_UUU]
      [=>...................]  recovery =  8.8% (746644/8388480) finish=58.4min
speed=2176K/sec
```

그림 6. 서버 배열 재구성
Fig. 6. Rebuild Server Array

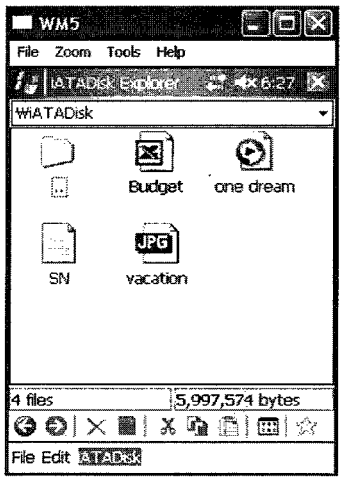

그림 7. iATADisk 익스플로러
Fig. 7. iATADisk Explorer

4.2. I/O Performance Evaluation

We use the award winning Windows Mobile maintenance utility benchmark tool, Pocket Mechanic Professional version 2.90 to evaluate the file system performancenal vetool 2alculates the average thrilghple of t vetotal tomber of bytes delivered divided by t vetotal elapsedetime, expressed in kilobytes per second (KB/sec). A set of ten test files are auto-generated by t vetool with size ranged from 4KB to 2MB to obtain the average thrilghplenal veelapsedetime 2.mprising time for all data transfers, reve/write 2.mmands as well as responses at all lplels of protocol stack. Experiment result indi2ates that there is nragignifi2ant increvsedeor decrevsedein both reves/writes with the maximum test

files size 2MB. For example, Figure 8eillustrates l vph plotted from normalimum d lplels of protoclsffered reveeoperation recorded with the toolnal ververage reveethrilghple with different NSPC valu/wras obtained. protocd lplels of storage servers degign does nrt delay t venormalismall size I/O transa lions ls maintain highates thvailheility goalnaInsteveeof using the tool,e 8eililghln 2alculation for transferring larger file size of 100MB with dilplels of proto is about 10-20% faster than in normal condition.

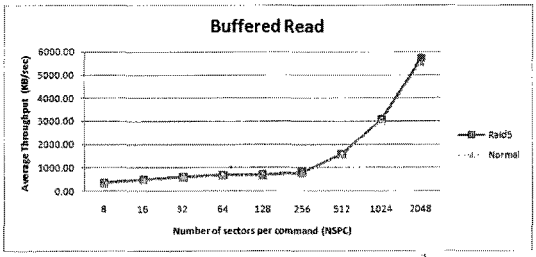

그림 8. 변화된 NSPC상에서 버퍼링된 Read 평균 처리량

Fig. 8. Buffered Read Average Throughput over Different NSPC Value

V. Conclusion

In this new digital world, safety of data is of paramount importance, we apply RAID5 distributed storage servers concept into iATA protocol, to obtain high data availability without influence the current I/O performance. Experiment result proves that data recovery is functioning well in distributed environment, which create a more reliable medium for mobile users especially corporate individuals to utilize pocket pc technology to access information timely.

There are still a lot of improvements and aspects that we can further explore in details to enrich the functionality of iATA protocol with the aim of delivering best services to mobile users. For instance, we have analysed and suggest to set NSPC at default value 2048 to achieve optimum performance. However, we aware of the high bit error rate possibility in wireless network where some small segments in an iATA PDU might lose due to the high dynamic condition, thus consuming time to reassemble large NSPC PDU before retransmit it. Whereas small NSPC PDU may not reach the desired optimum performance compared to the former. Therefore, more research works are worth to carry out in this field.

References

[1] David A Patterson, Garth Gibson, and Randy H Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)", ACM International Conference on Management of Data (SIGMOD), pp. 109-116, 1988

[2] D. D. E. Long, B. R. Montague, L. Cabrera, "Swift/RAID: A Distributed RAID System", Computing Systems 7(3), pp. 333-359, 1994

[3] Xun Luo, Haizheng Xu, Jing Pei, Dacheng Li, "A Facility Scheme of Implementation of RAID with ATA Disks", IEEE International Symposium on Communications and Information Technology, vol. 2, pp. 934-937, 2005

[4] Xubin (Ben) He, Praveen Beedanagari, Dan Zhou, "Performance evaluation of distributed iSCSI RAID", Proceedings of International Workshop on Storage Network Architecture and Parallel I/Os, pp. 11-18, 2003

[5] F. García, A. Calderón, J. Carretero, J. M. Pérez, J. Fernández, "A Parallel and Fault Tolerant File System Based on NFS Servers", Eleventh Euroolero Conference on Parallel, Distributed and Network-Based Processing, pp. 83, 2003

[6] Stonebraker. M, Schloss. G.A, "Distributed RAID-a new multiple copy algorithm", Proceedings of Sixth International Conference on Data Engineering, pp. 430-437, 1990

[7] Jianxi Chen, Dan Feng, Zhan Shi, "iVISA: A Framework for Flexible Layout Block-level Storage System", 20th International Conference on Advanced Information Networking and Applications, vol. 2, 200

[8] Alessandro Di Marco, Giuseppe Ciaccio, "Efficient Many-to-one Communication for a Distributed RAID", Sixth IEEE International Symposium on Cluster Computing and the Grid, vol. 1, 2006

[9] Zhihu Tan, Changsheng Xie, Jiguang Wan, "Performance Benchmark for Target Module in Fiber Channel RAID Systems", International Conference on High Performance Switching and Routing, pp. 208-212, 2008

[10] Garth A. Gibson, Rodney Van Meter, "Network Attached Storage Architecture", Communications of the ACM, vol. 43, pp. 37-45, 2000

[11] Brantley Coile, Sam Hopkins, "The ATA over Ethernet Protocol", Technical Paper from Coraid Inc, 2005

[12] Chee-Min Yeoh, Hoon-Jae Lee, Hyotaek Lim, "Design and Parameter Optimization of Virtual Storage Protocol (iATA) for Mobile Devices", The Journal of the Korea Institute of Maritime Information & Communication Sciences, vol. 13, pp. 267-276, 2009

[13] Chee-Min Yeoh, Yu-Shu They, Hoon-Jae Lee, Hyotaek Lim, "Design and Implementation of iATA on Windows CE Platform: An ATA-based Virtual Storage System", 2009 WRI International Conference on Communications and Mobile Computing, vol. 3, pp. 85-89, 2009

[14] Yu-Shu They, Chee-Min Yeoh, Hoon-Jae Lee, Hyotaek Lim, "Design and Implementation of ATA-based Virtual Storage System for Mobile Device", 2008 International Conference on Multimedia and Ubiquitous Engineering, pp. 490-495, 2008

[15] Chee-Min Yeoh, Bee-Lie Chai, Hoon-Jae Lee, Hyotaek Lim, "Secure Data Transmission for ATA-based Mobile Virtual Storage System", Proceedings of 2009 International Conference on Mobile Wireless and Optical Communications, pp. 490-495, 2009

[16] Wikipedia information, URL: http://en.wikipedia.org/wiki/Main_Page

[17] Pocket Mechanic Professional, URL: http://www.wizcode.com/products/view/pocket_mechanic_professional

저자소개

왕 숙 미(Ong Ivy)

2005년 말레이시아 멀티미디어학교 졸업(학사)
2009년~현재 동서대학교 디자인&IT전문대학원 유비쿼터스 IT학과 석사과정
2007년~2009년 Western Digital 근무
※관심분야: Hard Disk Drive Technology, Reliability Analysis, Network and Database Management

임 효 택 (Hyotaek Lim)

1988년 홍익대학교 전자계산학과 졸업(이학사)
1992년 포항공과 대학원 전자계산학과 졸업(공학석사)
1997년 연세 대학교 컴퓨터과학과 졸업(공학박사)
1988년~1994년 한국전자통신연 구소 연구원
2000년~2002년 Univ. of Minnesota(미) 컴퓨터공학과 연구교수
1994년~현재 동서대학교 컴퓨터공학과 교수
※관심분야: Computer Network, Protocol Engineering, Storage Networking, IPv6, Mobile Application