Research on massive data storage in virtual roaming system based on RAID5

RAN Feipeng, DAI Huayang, XING Wujie, WANG Xiang, Li Xuesong

College of Geoscience and Surveying Engineering, China University of Mining &Technology, Beijing 100083, China

Keywords: virtual roaming system, RAID5, PDSS, panoramic data

Abstract. A massive data management method was proposed based on RAID5 against limitations of data expansion, and a lack of automatic data recovery capability in current virtual roaming system: the panoramic data is separated from the roaming system as an independent panoramic data service system (PDSS), and distributed to multiple nodes to performance parallel service taking advantage of disk parallelism of RAID5. In this way, we can enhance the data storage capacity and fault tolerance of the server in a large extend. Finally, a prototype virtual roaming system was designed based on PDSS. The results showed that this method had a good application prospect.

Introduction

Virtual roaming technology based on the Panorama and combined with the advantage of GIS is an important approach to realize the visualization of GIS[1]-[2]. With the continuous areal expansion covered by the virtual roaming and greater detailed panorama spatial information, the storage nodes in data center under the massive data storage environment are very large,its efficiency and fault tolerance influence the efficiency and stability of virtual roaming system. It is difficult to meet the practical demands [3], though there are some solutions such as data compression or the improvement of the broadband to release the pressure of the server. These methods are not the best way to solve these problems, because the P2P structure has a large number of nodes and complex encryption algorithm[4]-[5].Redundant Arrays of Independent Disks(RAID)[6] puts a plurality of disks together to form a unified logical storage device, the common methods of which were disk imaging, striping and error correction[7]. For example, RAID5 scatters data stored in different disks arrays for the purpose of massive storage and high transmission rate, data security and storage cost, so it has been widely used [8]-[9]. In this paper, a new method was proposed to solve the massive data problem in virtual roaming system based on RAID5.

Data storage structure and fault-tolerance Mechanism

Taking one section of the road for example, supposing that there are i panoramic data and n isomorphic corresponding storage units, the throughput is k. The loading time is differ under the same network due to the different size of the panoramic data, named $t_1, t_2, ...t_n$, the whole loading time is $T = t_1, t_2, ...t_n$; the average number of the requesting access is λ, theoretically, the rejection rate is:

$$\rho = N_r / T_r = 1 - k/(1 + \lambda \times T) \quad (1)$$

N_r represents the number of data rejections, T_r the total access time. Therefore, to reduce the rejection rate ρ, we need to reduce T. If the panoramic data are stored on n different storage units, the service time of each units would reduce to T/n, the acceptance successful request rate would rise from $k/(1+\lambda \times T)$ up to $k/(1+\lambda \times T/n)$, which means that the actual parallelism of the application server is nearly n times compared to single disk.Hence, in order to improve data services performance, a server structure was adopted which is similar to the features of RAID5 [6]. As is shown in Fig1, the panoramic data is separated from the roaming system and distributed to multiple nodes to establish a data server connected by high speed network between service nodes and control nodes. The service nodes are used to dynamically store panoramic data, while the control nodes

used to response to the request from the roaming system to assign the service node. Data nodes store and send the data to service nodes via the high speed network according to the instruction from the service nodes.

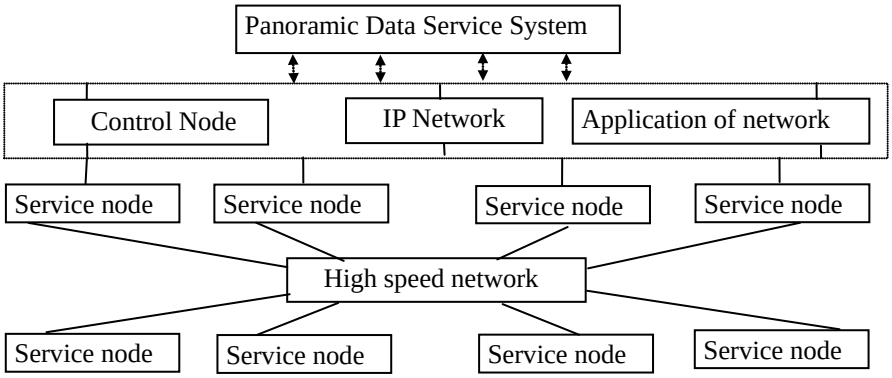

Fig.1 Structure of clustered parallel panoramic data service system

The Fig.2 shows the panorama data unit and the parity unit block. In each row, P represents parity section information, and the data in each node are stored in the remaining modules. The panorama data are divided into several StripeUnit, and a row data unit and verification units are combined into one stripe. The local panorama data are stored in data unit, and the parity unit which is useful in data recovery is used to store information on other nodes. We take advantage of random allocation algorithm to assign the panorama data, in other words, the storage unit are deposited to each node through a pseudo random function to solve the problem of node overload.

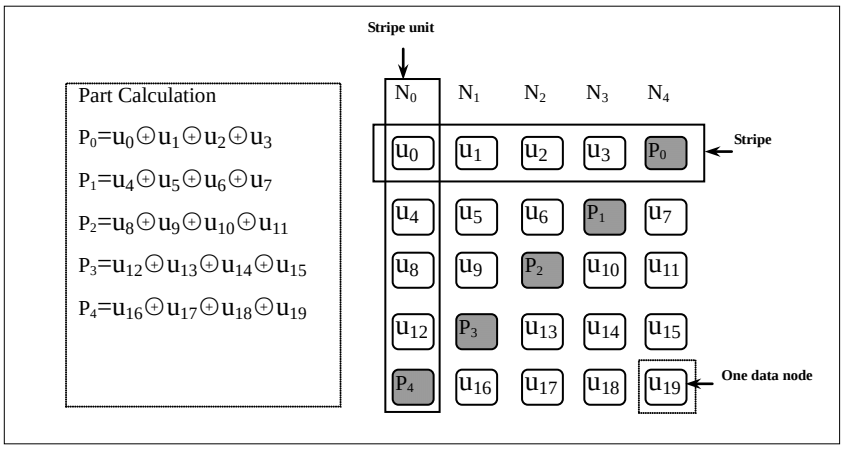

Fig.2 Storage configuration for parallel panoramic data

The data and the system are often combined together in traditional virtual roaming system which will reduce the stability and fault tolerance, that is to say the whole system will be paralyzed once fatal error or disk damage occurs on the local server. EVENODD code [10] is a dual fault-tolerance coding algorithm which is commonly used in array storage system. If the error occurs on any two disks, the redundant information of two disks in the system will be increased to recover the lost data. EVENODD code is based on the array of $(p\text{-}1)\times(p\text{+}2)$, and the first P columns is used to store raw data, and the last two columns which is the checking column is used to store the redundancy checking data. $d_{i,j}(0\le i\le p\text{-}2, 0\le j\le p\text{+}1)$ represents the line i data of disk j, The column P is called for row-best checking column. The checking block $d_{i,p}$ was gained by exclusive OR operation with all the original data in the line i; The p+1 column is diagonal-best checking list, we can get the checking block through the exclusive or between the regulators and the corresponding raw data blocks. The structural formula is as follows:

$$d_{i,j} = \bigoplus_{j=0}^{p-1} d_{i,j} \tag{2}$$

$$d_{i,p+1} = s \oplus (\bigoplus_{j=0}^{p-1} d_{<i-j>_p,j}) \tag{3}$$

$$s = \bigoplus_{j=1}^{p-1} d_{p-1-j,j} \tag{4}$$

Among them $< i - j >_p = (i - j) \mod p$,we need to use the redundant information to recover the data when a disk error occurred. Similarly, the row-best checking column is used for recovering the data in damaged disk. If error occurred on checking disk, all data in disks should be read to recover it according to the formula (2)-(4).

The virtual roaming system based on PDSS

The logical structure design. As is shown in Fig3, the system consists of image processing subsystem, parallelled storage subsystem and virtual roaming subsystem. The virtual roaming system consists of IFRAM module and GIS module which is responsible for spatial information display, query, analysis and output. This paper use the method in reference [11], which transmits the panoramic data to virtual roaming system as the source of IFRAM, meanwhile the panoramic data service system will be linked together with the virtual roaming system. This paper only discuss the storage method of panoramic data, so the storage method about geographic data will not be discussed.

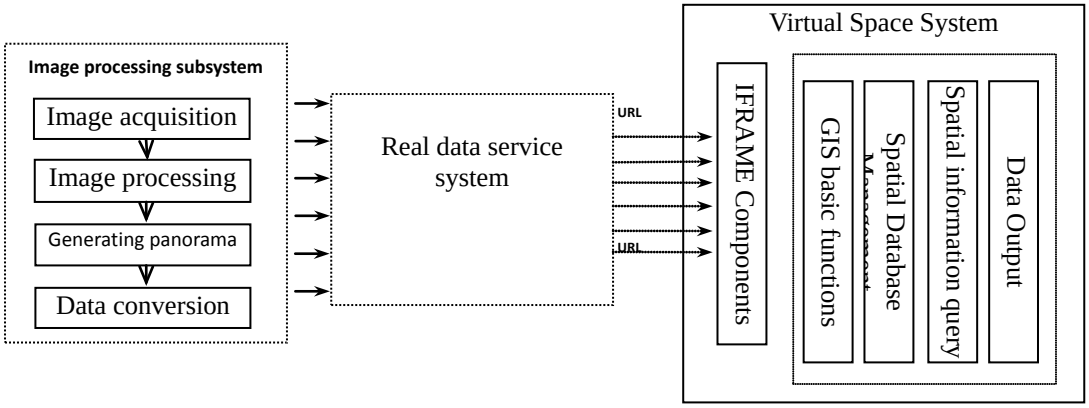

Fig.3 Framework of the system functions

Design of System Function Module. From the practical, easy operability and other principles, the system is divided into nine modules including document management, map operation, layer settings, tools, geographic analysis, data management, image editing, visual field labels and output, as shown in Fig4. The system will load the corresponding panoramic view when a user choose the start point and end point along the road in the electronic roaming map, at the same time, panoramic view can be set independently in the important region. The system interface includes display window, electronic navigation map and view control toolbar.

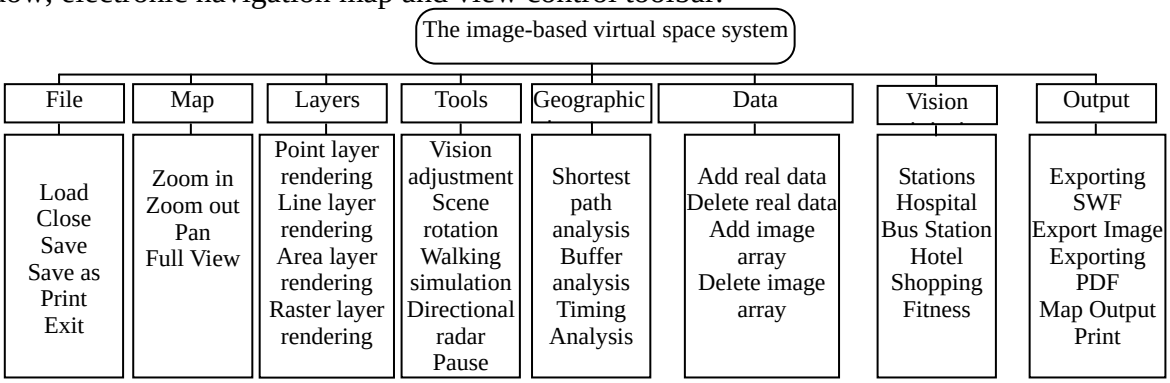

Fig.4 Framework of the system functions

System performance analysis

The performance of PDSS has a direct effect on whole roaming system. We are to evaluate the system performance separately from throughput rate, reliability and response time. We Selected 80 hard disks with the capacity is 80GB (the hard drive speed is 10000RPM, channel seeking time

is 5ms, the rate of data transmission is 200MB/sec), SCSI/FC interface, MTBF 1200000h, and 16 parity groups.

Data throughput. Throughput represents the speed of data transmission in storage system, which used to be expressed in two ways: I/O rate and data transfer rate. This paper adopted I/O rate, because each volume of request is small in virtual roaming system. 450 network services started on each node in client terminal at one time. There is a simulation test in which data is requested by several users at one time. Experiments show that the server application can easily handle 10000 data request and support 4200 concurrent requests, which broken through the limits of traditional storage system.

The response time of data acquisition. The response time refers to the total time of obtaining the data. Users and reconstruction process have a higher priority in contention for the disk bandwidth, on the contrary, the write operation will have to wait for the end of the reconstruction process. The response time is relatively faster. Simultaneously the operation of data reconstruction has a limit on the number of reconstructed Strip submitted by disks, which can ensure that the recovery workflow can evenly distributed among the disks. The working state of the system is relatively balanced.

Reliability analysis. Reliability is one of the most important indexes[12] to the performance of the system. The indexes of reliability consist of *MTTF*(average time of failure), *MTTR* (mean time to repair) and MTBF (mean time between failures), among them:

$$MTBF = MTTF + MTTR \tag{5}$$

For each disk, *MTBF*=1,200,000*h*, *MTTR*=1*h*. From *MTTR*<<*MTBF*，*MTTF*≈*MTBF*=1,200,000h, we can get the system failure rate is λ =*MTTF*-1, then the system reliability *t* is:

$$R(t) = e^{-\int_0^t \lambda dt} \tag{6}$$

$$MTTF = \frac{MTTF_{disk}^2}{N \times (G-1) \times MTTR} \tag{7}$$

We can get the average time: $MTTF_{\mathrm{RAID}}$=1,200,000,000h, therefore the reliability of the system peryear is: RAID=0.9999928. That is, the *MTTF* has increased to 1000 times of traditional copying backup. The failure time per year is 3.8 min, which is meet the requirements.

Conclusions

Taking advantage of the RAID5 such as easiness in storage extension, storage efficiency, cheapness, and automatic data recovery, a solutions of the massive data storage and fault-tolerance in virtual roaming system method was presented. The panoramic data was separated from the roaming system and distributed to multiple nodes. In this way, the expansion capability of the virtual roaming spatial would be enhanced in large extend. In addition, the system had automatic data recovery capability. Finally, the PDSS throughput, reliability and response time was tested, the results showed that this method provided a new way of solving the data limitation of virtual roaming technology.

References

[1] J Surong, W Song, F Gang, Technology of Panoramic View Based on Images, Computer Applications. 16(2002) 85-87.

[2] DEREK BRADLEY, ALAN BRUNTON, MARK FIALA, GERHARD ROTH. Image-based Navigation in Real Environments Using Panoramas [C], IEEE International Workshop on Haptic Audio Visual Environments and their Applications, Ontario, Canada, 2005.

[3] ZHU B, LI K, PATTERSON H. Avoiding the Disk Bottleneck in the Data Domain Reduplication File System [C]. Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST'08), San Jose, CA, USA, 2008.

[4] Jibf Tian, YaFei Dai .study on durable peer-to-peer storage techniques.[J]Journal of software 2007, 18(6): 1379-1399.

[5] E.Pinheiro, W.D.Weber, L.A.Barroso. Failure trends in a large disk drive population[C]. Proceedings of the 5th USENIX conference on File and Storage Technologies, San Jose, 2007, 17-28.

[6] HAYES B. Cloud Computing [J].Communications of the ACM, 2008, 51(7): 9-11

[7] HuanQing Dong,ZhanHuai Li,Wei Lin.RAID VCR: A New RAID Architecture for Tolerating Triple Disk Failures[J]. Chinese Journal of computers 2006, 29(5): 792-800

[8] WenWu Na,Jian Ke,XuDong Ke,et al, A Network RAID System with Backend Centralized Redundancy Management[J],Chinese Journal of computers, 2011, 34(5): 912-923

[9] SUNG HOON BAEK, KYU HO PARK. Prefetching with adaptive cache culling for striped disk arrays[C]. USENIX Annual Technical Conference, USA, 2008

[10] Blaum M, Brady J, Bruck J. EVENODD: an efficient scheme for tolerating double disk failures in RAID architectures [J].IEEE Transactions on Computers, 1995, 44(2):192-202

[11] Feipeng Ran, Tao Jiang, Huayang Dai, et al, Key technology of digital campus establishment based on flex framework, Journal of Geo-Information Science, 15(2013) 123-127(In Chinese).

[12] Liang Zhao, Key technology of Redundant Arrays of Independent Disks (RAID) [D]. National University of Defense Technology.2002.