Research on Automatic Archiving and CD-ROM Burning System of Massive Electronic Documents

Shiyang Deng[1,2,*], Yaoming Li[1] and Boxue Zhang[3]
[1]College of Computer Engineering, Weifang University, Weifang 261061, China
[2]College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[3]Beijing Guoruishi Technology Co., Ltd., Beijing, 100176, China
*Corresponding author

***Abstract*—According to the feature of huge data, massive files and complex handling in massive e-documents archiving, an auto archiving and burning system is developed. The multi-mission parallel mechanism based on multi-thread pipeline would solute the problem of resource competition by taking full advantage of multi-core processor, arithmetic unit and the I/O interface. It would shorten the time cost of classifying, compressing the files, generating disc data and burning disc, and thus improve the processing effectiveness of the massive E-documents archiving. The MD5 technology is used to guarantee the file security. A hierarchical structure is used to the database inquiry design to improve performance on data access. The blue-ray technology is used to accelerate the burning speed. A management system is developed to handle the process of disc archiving, accessing and inquiry. A disaster recovery system is used to solve the problem of invalid hard disc storage and database. A remote monitoring system is used to monitor the arching process in real-time. The whole system is running for more than one year and received high appraise.**

Keywords—massive electronic documents; automatic archiving; CD-ROM burning; multi-thread pipeline

I. Introduction

The archiving of massive electronic data has the characteristics of large data volume, massive files, complex data processing and large real-time throughput. At present, the research on the archiving of massive electronic data is mainly to improve storage efficiency and storage capacity by using efficient storage devices [1-4] or database technology [5, 6]. It doesn't pay much attention to data processing ability of software, so it is difficult to adapt to complex archival management requirements and high real-time data throughput requirements.

In the project of this paper, the data center receives more than 300GB of data every day. The number of files is over millions per day, and the number of documents in one year will reach 1 billion. The system needs to classify all kinds of historical data and real-time push data, and automatically store them into permanent storage media (Blue-ray discs). There are a large number of small text files with only dozens of words in the system, which wastes a lot of storage space and reduces the speed of data transmission. Therefore, the system should be able to do lossless compression on the specified classification data. In order to prevent data tampering, we also need to use MD5 digital fingerprint to verify consistency of stored data. The whole process of data processing is complex and time-consuming, while CD-ROM burning, verifying and cover printing are time-consuming hardware I/O operations, and the single task processing system will not meet the requirements of the burning system. Therefore, we use the multi-thread pipeline technology [7-10] to divide the burning task into multiple steps to make each step manipulate different hardware resources as much as possible. Then, each step acts as a thread, the multiple threads of the same task are executed sequentially, and one step of the latter task must wait until the same step of the previous task is completed. In this way, it can make full use of the advantages of multi core CPU, make each thread operate different arithmetic parts, hardware devices and I/O interface in parallel, solve the problem of resource competition, and improve the performance of the system in an all-round way.

II. System Design

A. Hardware Architecture

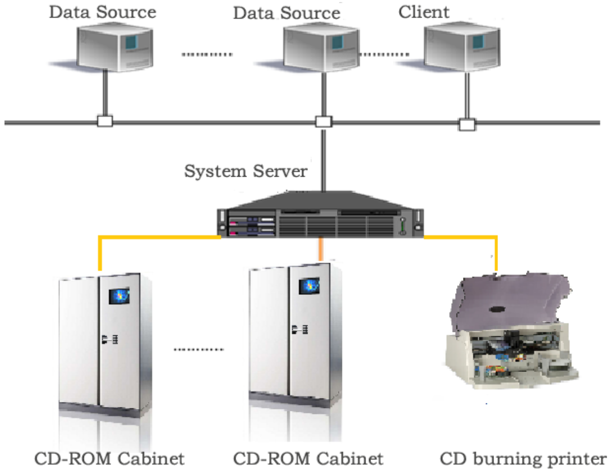


FIGURE I. HARDWARE ARCHITECTURE

The hardware of the system includes a system server, some CD-ROM cabinets and CD-ROM burning printers. The CD-ROM cabinet is equipped with an electronic indicator based on single chip computer. It can indicate the position of the disc according to the main control computer instruction. The CD-ROM burning printer can be controlled by the independent disc burning service program to record the data file and print the disk cover. The hardware architecture is shown in Fig.1.

The way of communication between the CD-ROM cabinet and the master computer is RS485. The longest control distance between the master computer and the cabinet is 1000 meters. A master computer can control up to 127 CD-ROM cabinets.

B. *Software Architecture*

The software is composed of four main parts, including automatic burning service platform, CD-ROM management system, disaster recovery system and remote burning monitoring system.

The system is developed by C++ and VB.NET, and the database is SQL Server 2008.

The blue ray disc printer of Bravo4102 Archive Disc Publisher is choosing for the CD-ROM burning machine. It can place 100 blue light discs at one time, and can print the cover of the disc dynamically.

The automatic burning service platform automatically completes data acquisition, data processing, CD burning, cover printing and other functions. It is the core of the whole system, which is divided into 6 modules.

Data processing: file directory analysis, generating MD5 digital fingerprints, and database processing.

Data compression: file classifying, compression property settings, and directory compression.

Data generating: burning information (CD-ROM description and file indexes), burning data (classified files, compressed files, two-dimensional code image file, data index files, tool files, etc.)

Burning task processing: task script generating, task checking, task submitting, and double machine joint burning management.

CD-ROM burning: burn state checking, error handling, success information processing, data validating, and disk cover printing.

Data cleaning: cleaning success information and expired burning data.

The CD-ROM management system mainly includes two sub-systems. The first sub-system is used to manage the stock in-out of CD-ROMs, communicate with the lower computer to control the indicator lights in CD-ROM cabinet, control information displaying on the monitor which is embed in the door of the cabinet. The second sub-system is used to set the basic parameters of the automatic burning, manage burning projects, classify the file directories and define CD-ROM data, query the burning state, browse CD-ROM data, and burn CD-ROM manually and etc.

Disaster recovery management system can run independently and does not need to access database server. It includes three functions of recovery setting, data recovery and data validating.

Disaster recovery setting includes directory setting, disc selection, etc.

Data recovery includes directory recovery and file recovery, which are used to restore all files in the specified directory or to restore specified files.

Data validating is used for file verifying and digital fingerprint validating.

Through remote monitoring system, administrators can monitor all kinds of information of automatic burning system locally or remotely in real time. The remote monitoring system can communicate with the burning service program by client/server mode and does not need to access the database server.

The software architecture and work flow are as shown in Fig. 2.

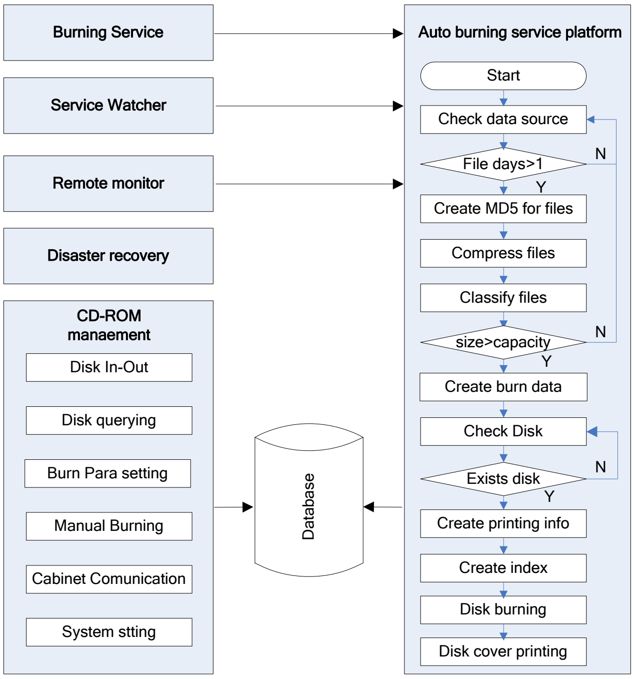


FIGURE II. SYSTEM ARCHITECTURE AND WORK FLOW

III. SYSTEM IMPLEMENTATION

A. *Database Design*

As the number of files is very large, at least 100 million files are submitted each year, and the information of these files must be saved to the database for later query, it is clear that the traditional data processing method cannot handle such huge data.

The system solves the problem of data processing and query through the following ways.

The partition lookup method based on the layered structure. The data blocks are built according to the project, the CD set, the annual date, the directory, the file and so on. This can greatly reduce the scope of the data query, and greatly reduce the total amount of data query.

Aggregation indexes are established to improve the speed of data query.

The data summary index is set up. A large number of operations are processed through summary data. It can reduce the frequency of the access to the final file and greatly improve the efficiency of the data execution.

B. *Automatic Archiving and Burning System Based on Multi-thread Pipeline*

The automatic burning system automatically completes data acquisition, data processing, CD-ROM burning, disc cover printing and other functions of the burning system. The system uses Windows Service program to ensure 24 hours of continuous operation and restart without login after power off reboot

The process of automatic archiving and CD-ROM burning of mass E-documents has many time-consuming operations such as data processing, file compression, CD-ROM data generating, burning and so on. Therefore, we take each step of the task as a thread, the threads of the same task are executed sequentially, and one step cannot be started until the same step of the previous task is finished. This is a multi-thread pipeline execution structure. It can take full advantage of CPUs, arithmetic units, hardware devices and I/O interfaces to solve the problem of resource competition and improve the performance of the system.

The main feature of multi-thread pipeline structure is that the system will not decrease the processing power due to the increase of thread number, but maintain at a constant level of [8]. This is similar to the transaction processing mode of the queue mode. When the processing requirements are increased, the requests that cannot be processed temporarily are placed in the queue, instead of assigning the processing resources directly in the system as in the time-sharing system. Using multi-thread pipeline technology, the system resources can be used in parallel processing, and the conflict between high speed equipment and low speed equipment is fully coordinated, and the throughput of the system is greatly improved.

A burning task in this system has six steps of data processing (including file directory analysis, MD5 digital fingerprint generating, database writing, etc.), file compression, generating the CD-ROM data (classifying and copy files according to CD definition), burning task management, CD-ROM burning (reading the specified directory data to the record machine, burning data to CD-ROM and printing the disc cover) and cleaning up data. These six steps are executed sequentially, with different I/O components or different hardware operations during execution, so the six-level pipelining structure can be used to deal with each of the burning steps in parallel and improve the execution efficiency.

The tasks are expressed as T_1, T_2, T_3, …, T_i,…, T_n. The six steps of the task T_i are expressed as six threads of T_{i1}, T_{i2}, T_{i3}, T_{i4}, T_{i5} and T_{i6}, are executed successively. When data processing is done by the thread T_{ij}, the thread synchronization technology is used to guarantee the same level thread of the previous task $T_{i\text{-}1,\ j}$ does not perform at the same time to avoid the resource competition between threads. In this way, at the same time, 6 burning tasks can be processed in parallel at the same time, and the processing advantage of multi-processor CPU is fully utilized, the competition of resources is reduced and the speed of data processing is improved.

The execution process of six-level multi-threading pipeline is shown in Table 1.

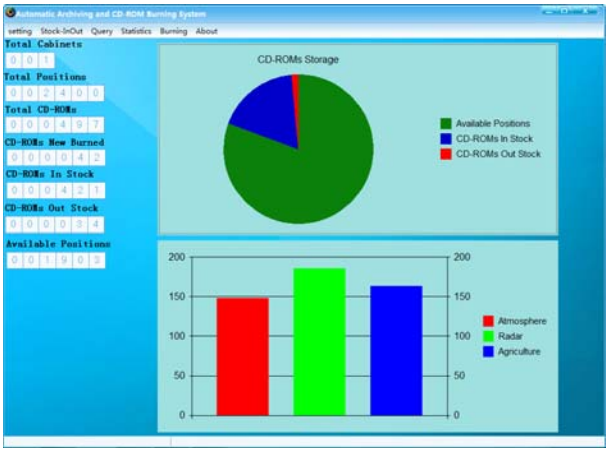
TABLE I. SIX-LEVEL MULTI-THREADING PIPELINE

T_1	T_{11}	T_{12}	T_{13}	T_{14}	T_{15}	T_{16}						
T_2		T_{21}	T_{22}	T_{23}	T_{24}	T_{25}	T_{26}					
T_3			T_{31}	T_{32}	T_{33}	T_{34}	T_{35}	T_{36}				
T_4				T_{41}	T_{42}	T_{43}	T_{44}	T_{45}	T_{46}			
T_5					T_{51}	T_{52}	T_{53}	T_{54}	T_{55}	T_{56}		
T_6						T_{61}	T_{62}	T_{63}	T_{64}	T_{65}	T_{66}	
T_7							T_{71}	T_{72}	T_{73}	T_{74}	T_{75}	T_{76}

It is easy to see from Table 1 that after the six-level multithread pipeline, the time for 7 tasks is equal to the time used by 2 tasks, and the execution speed is increased by 3.5 times. That is to say, the speedup [9, 10] is 3.5. Suppose the time of every step is 1, then the time of a task is 6. If there are x tasks, the processing time of single thread is 6x, and the processing time of six-level multithread pipeline is x+5. Then we can get the speedup=6x/(x+5). When the x is large enough, the speedup of the six-level multi-thread pipeline is close to 6.

C. *CD-ROM Management System*

The management system is used to manage the stock in-out, communicate with the lower computer in CD-ROM cabinet, set the basic parameters of the automatic burning, and query information in CD-ROMs. The main window is shown in Fig.3.



Total Cabinets 0 0 1

Total Positions 0 0 2 4 0 0

Total CD-ROMs 0 0 0 4 9 7

CD-ROMs New Burned 0 0 0 0 4 2

CD-ROMs In Stock 0 0 0 4 2 1

CD-ROMs Out Stock 0 0 0 0 3 4

Available Positions 0 0 1 9 0 3

FIGURE III. MAIN WINDOW OF CD-ROM MANAGEMENT SYSTEM

IV. System Function Test

A. Environments

The processor uses Intel Xeon E5 2609 V3 6 kernel CPU, dual CPU configuration, a total of 12 processors, memory 32GB, hard disc 3*2T, Raid 5 disc array, 4T hard disc space available.

B. Parallel Processing Ability

Single threaded test: only a few processors can be well utilized, and the utilization rate of CPU is between 10%~25%, as shown in Fig.4.

Multi-thread pipeline parallel processing test: all the 12 processors have been well utilized, and the utilization rate of CPU is over 90%, as shown in Fig 5.

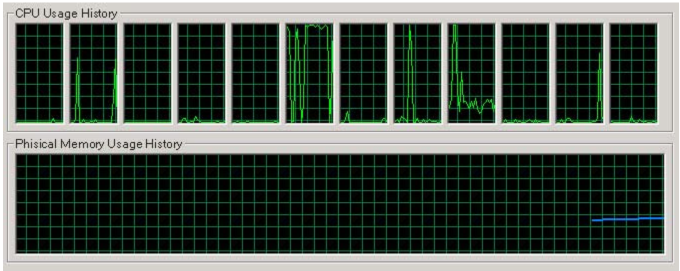


FIGURE IV. CPU AND MEMORY USAGE HISTORY UNDER SINGLE THREAD PROCESSING

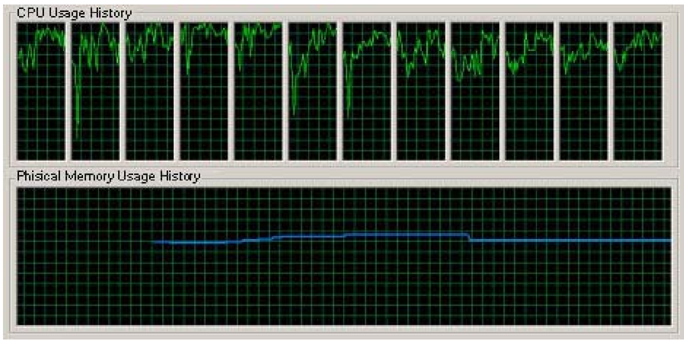


FIGURE V. CPU AND MEMORY USAGE HISTORY UNDER MULTI-THREAD PIPELINE PARALLEL PROCESSING

V. Pressure Test

Data sources: meteorological radar real-time data from 2017-05-07 to 2017-07-03.

The results are shown in Table 2.

Test conclusion: through the pressure test of each execution node, the system can satisfy data processing capacity of no less than 300GB per day.

TABLE II.RESULTS OF SYSTEM PRESSURE TEST

No.	Test Item	Test Result	conclusion
1	Data processing	981.8GB/day	satisfied
2	Data compressing	406.2GB/day	satisfied
3	Data generating	601.7GB/day	satisfied
4	CD-ROM burning	577.2GB/day	satisfied
5	Data cleaning	829.0GB/day	satisfied

A. Function Test

After the completion of the system development, the system has been tested in full function according to the system requirements and instructions. The intermediate process and results are strictly compared. The errors and shortcomings are solved. After that, the system has been tested for more than 1 year, and the functions and indexes after the test are all better the original requirement.

VI. Conclusion

This system uses the thread pipelining technology to deal with the time-consuming operation of classification, compression, data generation and burning of mass E-documents data. It can make full use of CPU, all kinds of operation parts and hardware equipment to solve the problem of resource competition. At the same time, the database processing ability is improved by the sub layer structure of the database. On this basis, we designed and developed a mass E-documents data archiving management system which integrates data processing, disc burning, disc printing, CD management, remote monitoring and disaster recovery. In July 2017, the system was accepted by the expert group. All the functions and indexes of the system reached the contract requirements, and the degree of automation was high. It is stable and reliable.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grants 61170078 and 61472228; the “Taishan Scholar” Construction Project of Shandong Province, China; the key research program of Shandong Province under Grant 2016GGX101031, the promotive research fund for young and middle-aged scientists of Shandong Province under Grant BS2015DX010 and BS2015ZZ006.

References

[1] W.R. Yan, Q. Cao, J. Yao and C.S. Xie. A novel file system for large-scale optical library. Chinese Journal of Computer Research and Development, 52 (2015) 1-8

[2] J.J. Miu, Y.J. Fu, P.Y. Yu and H.D. Mao. Design and realization of energy-efficient hybrid magneto-optical filing system. Chinese Journal of Computer Technology and Development, 8 (2017) 52-56

[3] W.D. Zhao. Research on Electronic Archives Blu-ray Storage Practice. Archives Science Study, 3 (2015) 88-95

[4] P. Yang. Storage and Management of Meteorological Electronic Archives Based on Blu-ray Storage Technology. China Computer&Communication, 2015, 16: 84-86

[5] R.T. Wang, W.C. Li. Research on Technology of Basic Large Data Storage System. Chinese Journal of Computer Technology and Development, 2017, 27 (8): 66-72

[6] L.N. Song. Research on Multi-tiered Storage Based Key Technology in Mass Storage System. Master's Thesis of National University of Defense Technology, Changsha, China, 2011

[7] X.F. Hu, L.H. Zhang, W.Y. Liu and H.D. Wang. Design and Implementation of Data Storage System Based on VC Multithreading and Pipelining. Chinese Journal of Electron Devices, 2016, 39 (4): 964-967

[8] Y.S. Liu and S.J. Chen. Pipeline Mode in Query Process Based on multithread. Chinese Journal of Computer Application, 2004, 24 (6): 54-56

[9] Y.Y. Ji, T. Zhang and H.P. Wang. Multi-core Thread Schedule Strategy Based on Thread Pipeline. Chinese Journal of Computer Engineering, 2013, 39 (2): 279-287

[10] J. Wang, X.Y. Fan, S.B. Zhang and H. Wang. A Survey of CMT processor thread Schedule Polices. Chinese Journal of Computer Science, 2007, 34 (9): 256-258