Based on Multi-Thread Dynamic Self-Healing Technology Application Research in Big Data Image Processing

Guangyu Xu and Xiaohua Qiao

School of Computer Information Management, Inner Mongolia University of Finance and Economics, Hohhot Inner Mongolia 010070, China
xgy@imufe.edu.cn

Abstract

With the rapid development of information technology, the data scale expands unceasingly, produce a large amount of image data, a massive image data processing has also is an important technology. Traditional processing techniques, in dealing with such a large scale of image data, have been unable to meet the requirements. One of the important reason is that whether the c ++ or Java, there is memory leak problem, when the data is the relatively small size, these problems may be insignificant, but when the data is very large, will highlight these shortcomings, will lead to serious program crashes, the system outage, ultimately unable to achieve expected goal. In considering the data processing efficiency, stability, is proposed in this paper, based on multithreading, a dynamic self-healing capacity of mass image data processing technology, compared with the traditional technology can improve the efficiency and improve the stability. Finally, after the experimental results verify that the method is effective.

Keywords: Multithreaded big data image processing dynamic self-healing

1. Introduction

With the rapid development of information technology, the scale of our daily contact information data has become increasingly large, traditional data units MB, GB cannot meet the requirements, by 2012, the amount of data from TB (1024 GB = 1 TB) levels jumped to PB (1024 TB = 1 PB), EB (1024 PB = 1 EB) and ZB (1024 EB = 1 ZB) level, according to the characteristics of the data scale growth, global consulting firm, McKinsey, puts forward the concept of the era of "big data", then appeared frequently in people's field of vision.

In the era of "big data", data processing scale is increasing, and the image as a kind of information carrier, than text contains a greater amount of information, the more intuitive, image data has become the era of "big data" a very important data type. Image data size will be increasing at the same time, by 2011, facebook users to upload pictures only has reached 100 billion, in the Internet, at the same time, the world will produce 500 million images every day [1]. To process the image data on such a scale, will eventually image data or photo thumbnails data in the database is an important part of the application, which is according to the image features for data retrieval, data mining, the basis of data analysis.

Any program in the process of running the risk of memory leaks, when the program size is very small, it may not harm is not serious, but the program size, such as hazard will into the geometric level increases, while the processing efficiency is lower, weight is system downtime. On the Internet the most used technique is based on Java, Java, although the risk of memory leaks than c ++ has decreased, but it can't be completely avoided. Although the Java virtual machine has a garbage collection mechanism to protect the system, but in the case of high load, the low level of the Garbage Collector system is not effective, so that the system throws out of the exception of the available memory after

the Java virtual machine exit. In this paper, we try to establish a multi - thread, with the ability of self-healing, and the high stability of the operating environment to meet the needs of large-scale image data processing.

2. The Processing of Memory Leaks

2.1. Using Virtualization Platform Monitoring Memory Leaks

The basic idea is that the first design of a virtual machine manager, and then through the virtual machine manager platform insert assembly instruction, to intercept application for and release of memory resources and function call, can build the function call, the program runs after the use of dynamic memory resources list, if a memory leak, memory can't be used by other applications [2]. Memory use by the use of the list of memory monitoring, if the memory is not released for a long time is not to be released, then it is likely to be leaked memory, through technical means to its forced recovery, to achieve the purpose of eliminating memory leaks. This method should not be restricted platform, through experimental verification is effective, is a kind of better solve the mechanisms of memory leaks, but its implementation is more complex, on the one hand to the corresponding research program assembly code, on the other hand also virtual address, physical address, machine processing of complex mapping relationship, for some specific application difficult to achieve.

2.2. Using the Page Replacement Monitoring, Deal with Memory Leaks

The basic idea is to learn from the operating system page replacement in thought, if detected when the virtual machine heap memory is less than the threshold, it will be part of the object, from the heap memory written to the disk, and record the state of these objects, at the same time start the garbage collection mechanism spill recovery of memory, after will be written to disk object to loaded into memory [3]. The method is verified by the experiment is also effective, but there is a problem, that is, objects may often need between memory and disk status from a conversion when the data size is large, program performance dropped significantly.

2.3. Using Static Test Method of Monitoring the Memory Leaks

The basic idea is to establish a static detection model of memory leakage based on interval arithmetic, which is based on the syntax tree abstract and Control Flow Graph [4].Through the experiment, this method can effectively detect the program that is running in the process of the memory leak problem.

3. Based on Multi-Threading Memory Management

3.1. Multithreading Technology

Traditional single-threaded program processing efficiency is very low, in view of the large data processing, when the data size is larger, efficiency decreases quickly, so can use multithreading technology in the large data processing.

Multithreading is a mechanism that allows concurrent execution of multiple instruction stream in the program, each instruction stream can be called a thread, thread between independent each other. From a logical point of view, a multithreaded means that a program of multi-line statement execution at the same time, but they are not equal to several times starts a multithreaded program, the operating system will not treat each thread as a separate process to, in a nutshell, is that each program perform multiple tasks at the same time.

Concurrent perform multiple tasks at the same time is multi-threaded Java language one of the main features. Under the condition of the previous hardware, really it is impossible to completely achieve concurrency. Today, however, an era of rapid development of computer software and hardware, Multi core CPU has been very common, cloud computing technology have begun to gradually popularization, in this case, in order to improve the efficiency of the program, can put the paragraph into a thread of execution at the same time, the distribution to different processor, let them in a different virtual machine implementation, parallel computing, to complete the task.

3.2. Memory Management Based on Multi-Threading Technology

For multithreading technology, memory scheduling and management is a key and complex problems, According to the task, whether we can be divided into online and offline model, according to whether the task must be put into memory, it can be divided into mandatory model and selection model. Currently using more memory scheduling model is Uniform model, Cost model, Fault model, Bit model, General model [5].

3.3. The Memory Scheduling Based on Multi-Thread Technology

Based on multi-threaded memory scheduling, there are several basic strategies and methods.

Scheduling strategy based on approximation algorithm. The core idea is that only a chain of memory scheduling problem with the same space model can find the optimal solution in polynomial time [6].

Scheduling strategy based on mining. The strategy instruction scheduling according to the results of the sampling behavior, advantage is not need to know the character of load beforehand, the disadvantage is that mining space as the growth of the thread and the number of core and rapid growth, is suitable for the core, combination of thread, a limited number of situations [7].

The strategy Based on heuristic .such as sorting with scheduling algorithm, random merger scheduling algorithm, the shortest public chain scheduling algorithm [8].

Scheduling strategy based on instruction type. The basic idea is to use mixed mode instructions as inspiration basis, mainly for long-delayed operation instruction threads, functional unit not used effectively, will these threads and computing-based, cooperative scheduling delay fewer threads [9].

Multi kernel thread scheduling based on thread pipelining. Based on the multi thread processor, drawing the parallel advantages of the pipeline technology, the concept of thread pipelining is introduced. The degree of polymerization degree and the thread of the thread are calculated [10].

4. Software Self-Healing Mechanism

Real application software, in the actual use process, will encounter many unknown circumstances, these may be unexpected software designers before, if not ahead of some processing methods or techniques, may produce a lot of negative effects, and may even cause irreparable damage, which requires some software can have self-healing ability, is in the process of program execution, have some problems, to launch the corresponding function, without human intervention, dealing with problems, solve the problem, the final program can be carried out smoothly and realize the expected goal. At present there are five kinds of technical means to achieve self-healing technology, data structure repair, repair software precision, reflection technology, bionic technology.

4.1. Computational Reflection Technique

Computational reflection technique is in the execution of a program's own, programs like data says program state of a processing capacity [11]. This process includes two aspects, introspection and intercessory. Usually the computational reflection is short for reflection. Introspection is the application ability to observe and status is derived according to the observation. Mediation is a program to modify their execution status, or change its explanation and the ability to change the connotation. Introspection and mediation requires a mechanism, the execution status of as data encoding: provide this code is called the reification. Using reflection programming, can make your program at runtime by introspection operation requirements, using specific ability to mediate themselves (in popular culture, is the runtime automatically modify the program), the dynamic ability to obtain a new behavior, the program can reflect on yourself, to think of itself as the data, can modify themselves. Language or application with reflection ability, can put the definition of the partial function application delay to the runtime sure, also is to let the end user custom applications.

4.2. Computer Bionic Technology

Computer bionic technology, the computer software, as a kind of imitating nature biological model, of course, can also pass the day after tomorrow learning to enhance their ability of environmental adaptation, bionic software in the network environment the survival, competition and so on itself is also a process of continuous learning [12]. Learning is an important technology of artificial intelligence, the bionic software engineering as a tool to make software become smarter, learning technology is an integral part of the program function, by learning about some of the problems, according to the known solution, to solve the problem, as the program's problems and solve the problem on the rise, program your ability to solve new problems will also strengthen, make software with the adaptive ability gradually.

4.3. Self-Healing by Repairing Software Accuracy

According to with the development of software industry, quality, cost and schedule of software put forward more and more high demand, some areas, may be because the precision of the software can't meet the requirements, which leads to some serious problems, such as the shuttle exploded, missile launch failure, these problems, you can define the software to improve accuracy to solve problems. Concrete is generally by static testing technology, in order to determine the scope of an operand or parameters in the program, the need to traverse the entire control flow chart, according to the abstract syntax tree each node control flow information, for in the process of each variable to track data flow analysis, its scope is interval arithmetic. Eventually find the need to improve the precision of the place, in order to improve the correctness of the software as a whole.

4.4. Self-Healing by Data Structure Repairing

Repair data structure refers to the application software running, the program list, tree, such as data structure and data, may be the defects of itself, the influence of the incident, and so on and so forth, led to the incorrect, the connection relationship between the data node or node within a data field value is not correct to change and the like, referred to as a data structure errors [13]. Error data structure is one of the main defect affecting the application of software reliability, can cause application software breakdown seriously affect the service availability of events. Data structure repair there are many related research and articles published, but as far as I know, the existing method of repair most take the form of separate tools, or is designed to assist developers, haven't updated dynamically repair data structure in the wrong way.

4.5. Self-Healing Through Dynamic Reconfiguration

Dynamic reconfiguration technology, one kind is based on the change of the program running in the process of running state, and changing the structure of the program itself, or configuration properties, has reached to improve software flexibility, the program has certain adaptability, the technology, need before the program is running, a lot of static model, the state one by one definition of possible solutions [14]. Another technique, the program is based on the structure of the component, the program will according to running status in the procedure, by component to add, delete components, component replacement and technology to change the original software components in the joint relationship, change the topology of the software, mapping relationship between elements, making process has certain adaptability [15].

5. Based on Multithreading, Dynamic Self-Healing Mechanism of Memory Management

This paper uses a dynamic reconfiguration combination of technology and multithreading technology, finally realizes the self-healing mechanism. The principle of which is to imitate the operating system process scheduling and management mode, using multi-threading technology at the same time, in the image processing program for program establishment and the simultaneous execution of process, the process is independent of the other processes running, works in a similar daemon. Work the foreground image processing process, background process automatically, the current stage, all image processing process has been completed the process automatically exit. The process can be real time no longer need to use the object, real-time monitoring of memory usage, analysis the total memory in the System, the Java virtual machine maximum memory, Java virtual machine airspace, the relationship between memory to determine if a memory leak, if you find memory leaks, immediately for processing, simulation operating System process scheduling for intervention, and make the process of image processing program temporarily sleep time, at the same time compulsory recycling has marked no longer use the object, then the scheduling process of image processing, if the process scheduling for many times, still fail, you can back to the front of the thread scheduling failure state, through the above strategy, solved completely done by the operating System process scheduling, System Garbage Collector is responsible for recycling of the memory leak problem. The multi thread technology and dynamic self-healing technology combining can ensure stability of the large image data processing program, and avoid the emergence of system downtime and other serious problems.

6. Based on Multi-Threading Memory Management

6.1. Traditional Single Thread without Self-Healing Capability of Image Processing and Multi-Thread Self-Healing Ability of Image Processing

In order to study the theory put forward by above, need further verification by experiments. Experimental software for Windows XP environment, database is MySQL server 5.6.24. In order to validate the program processing capability in a greater pressure, in the case of no image data compression, for a specified number of image data processing, the deposit to the database. The first experiment, the experiment compared the hardware environment, in the same single thread image processing and the difference of multithreading, specific see Table1.

Table1

Contrast program	CPU	Memory	hard disk	Image file number	Image total capacity	success/fail	using time (second)
Single thread, no self healing ability	Intel I5 2400 3.1GHZ	2G	500G	2386	4.6G	success	4640
Multi thread, no self healing ability	Intel I5 2400 3.1GHZ	2G	500G	2386	4.6G	success	771

In the process of the experiment, two different programs were monitored by the performance monitoring software to compare the CPU usage (Figure 1 and Figure 2), the memory usage (Figure 3 and Figure 4).
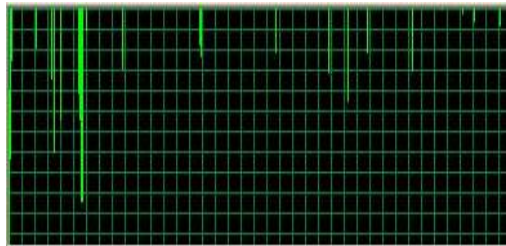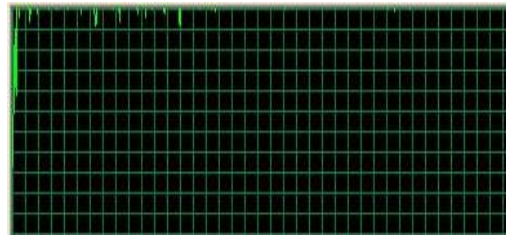

Figure 1. Single-Threaded Program CPU Usage

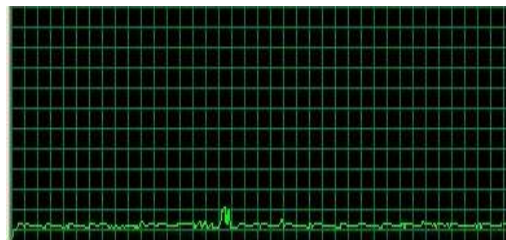**Figure 2. Multithreaded Program CPU Usage**

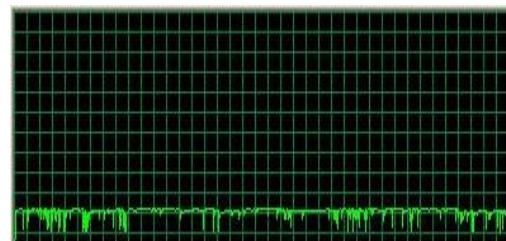**Figure 3. Single-Threaded Program Memory Usage**

Figure 4. Multithreaded Program Memory Usage

According to the experimental results, through the contrast can be seen that for the same size of the data, although multi-threaded memory usage than single thread slightly tall, multi-threading CPU utilization is slightly higher than single thread, the overall performance difference is not too big, but the multithreaded processing time is only 16.6% of the single thread, efficiency was 83.4% higher than that of single thread, by comparison with the experimental results, it can be seen, in a large image data processing, based on multi-thread programming under the condition of the system resource usage is not increased, but efficiency than single-threaded program has a larger increase.

6.2. Image Processing and Multi Thread with Self-Healing Ability of Multi Thread and Non Self-Healing Ability

In the era of big data, the size of data has become very large, so in a more massive image data processing, multi thread program and what? Please look at the following second experiments, see Table 2.

Table 2

Contrast program	CPU	Memory	hard disk	Image file number	Image total capacity	success/fail	using time (second)
multithread, no self-healing capability	Intel I5 2400 3.1GHZ	2G	500G	25142	50.2G	fail	
multithread, with self-healing capability	Intel I5 2400 3.1GHZ	2G	500G	25142	50.2G	success	29889

In the performance monitoring results, the use of CPU in Figure 5, the use of memory is shown in Figure 6.
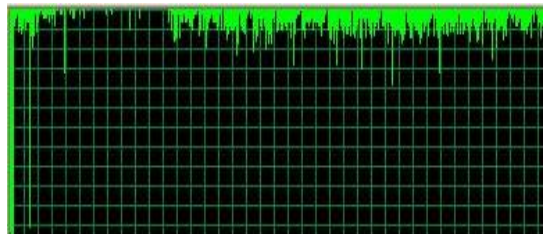

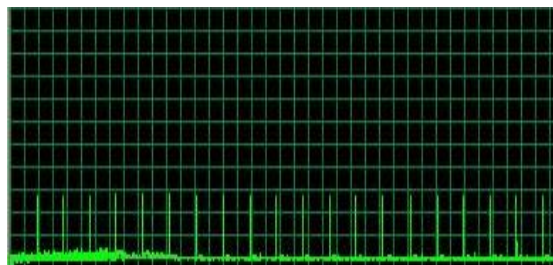**Figure 5. Multithreaded Program CPU Usage**

Figure 6. Multithreaded Program Memory Usage

During the experiment, based on multithreading, not program with self-healing ability, while higher efficiency than single thread, but because this is a large amount of experimental data, based on multithreading, not a self-healing capacity procedures after

many experiments, finally ended in failure. The reason is that, when the data scale is very large, the program has a memory leak, resulting in the final failure.

Trough monitoring results can be seen in Figure 5, based on multithreading, program, a self-healing capacity due to large data size, CPU utilization rate has been high. Through monitoring results can be seen in Figure 6, during the processing of memory usage when a warning threshold is reached, intervention system, rapid drop, memory to continue processing after, when dealing with a certain scale of data, memory footprint and achieve early warning threshold, intervene again, this program in systems have a memory leak, and at the end to the brink of collapse, timely processing, make the program has the function of self-healing, eventually the smooth completion of massive image data processing, and show that this method is really effective.

7. Summary

In terms of big data image processing, multi-threading technology than single thread technology with high efficiency. But when the data size increases further, because the Java virtual machine System level is low, the Java Garbage Collector itself can lead to memory leaks, makes the processing task easy to fail. In the big data in image processing, the processing of massive image data of the test proved this method is effective, The direction of future research will consider multithreading technology and based on AOP and Java reflection mechanism of dynamic self-healing technology, on the one hand, can make the program will affect the performance of the system further reduced at the same time, can withstand greater memory pressure, on the other hand also enables the program to better able to reuse, to solve similar problems.

Acknowledgements

References

[1]   L. Cao, "Design and implementation of massive image system", Shandong University, master's degree thesis, vol. 4, **(2014)**, pp. 1-2.
[2]   X. Wang, "The use of virtual platform for memory leak detection", Journal of computer science.
[3]   W. Lu, Z. Yian, L. Zhang, D. Wang and L. Chen, "Ma. A solving memory leaks the self-healing mechanism", Journal of Northwestern Polytechnical University.2011.8:649-650.
[4]   P. Ke, G. Yunzhan and Y. Zhaohong, "Static memory leak detection model", Computer science, vol. 4, **(2009)**, pp. 175-178.
[5]   H. Xu, "Multithreaded memory scheduling Zhejiang University PhD thesis", vol. 9, **(2011)**, pp. 11-18.
[6]   H. Xu, "Multithreaded memory scheduling Zhejiang University PhD thesis", vol. 9, **(2011)**, pp. 37-40.
[7]   X. Fan, Z. Shengbing, J. Wong and W. Hai, "Study on the thread scheduling strategy for multi core and multi thread structure", Computer science, vol. 6, **(2007)**, pp. 256-258.
[8]   Y. Ji, "Multi core thread scheduling strategy based on pipeline thread", Computer Engineering, vol. 2, **(2013)**, pp. 279-282.
[9]   Q. Fu, "Research on the software self-healing architecture based on Java dynamic programming technology", **(2007)**, pp. 15-18.
[10] P. Chen and Y. Wang, "Adaptive technology and software", Computer engineering and design, vol. 10, **(2003)**, pp. 25-30.
[11] D. Zhao, "Computer engineering", Repair of error in software dynamic update, vol. 7, **(2008)**, pp. 46-48.
[12] B. He and M. Bo, "A support component for the deployment of dynamic reconfiguration framework", Computer engineering and applications, vol. 5, **(2006)**, pp. 82-84.
[13] X. Chen, "Study on the configuration of dynamic adaptive software architecture", Computer science, vol. 6, **(2015)**, pp. 185-188.

Authors

Guangyu Xu, received the BS degree in Computer Science and Technology from Inner Mongolia Normal University and the MS degree in Information Engineering from Inner Mongolia University of Technology, China in 2002 and 2006 respectively. He is currently researching on Computer Application Technology.

Xiaohua Qiao, received the BS degree in Planning and Statistics from Inner Mongolia University of Finance and Economics and the MBA degree from Inner Mongolia University, China in 1991 and 2005 respectively. She is currently researching on information management and information economics.