The analysis and research about data transmission system

Yang lin[1,a]

1Department of Control Engineering, Naval Aeronautical and Astronautical University, Yantai Shandong 264001, China,

[a]email: liudi5388466@163.com,

Key words: data transmission; communication management; transmission efficiency; interrupt handling; the buffer;

Abstract. Big data transmit within the wan efficient and steady is the premise of the big data applications.The solution of traditional wan data transmission is mature, but there are many limitations in the era of big data. Transmission efficiency is poor and it is not apply to multi-source data transmission.The relevant content of big data transfer is studied in this paper.Based on the analysis and research of relevant data transmission technology and wan acceleration program, the thought of push and charge is used to ensure the reliability of data transmission.Through testing function and performance of transmission system, that transmission system has a certain advantage on the transmission rate is proved.

The Introduction

With the rapid development of computer and network technology, network bandwidth、storage capacity and the computer processor power has a fly. The degree of people depending on network is becoming more and more high, network is used to carry on work、study and entertainment, people enjoy fast and convenience brought by the network, the network further integrates into the life of people[1-4]. Meanwhile, web-based applications product not only can provide convenience for people's daily lives, but also can meet the needs of scientific research. However, the data produced by network applications is more than the usual size. The speed of producing data is very fast. Sources and structure are different. People often call this kind complicated and various data as big data. Big data is large capacity, fast speed, species diversity and cross regional information resources. It need to use the advanced processing technology to improve the insight of rule, process optimization and making decision. According to the data application steps, big data technology can be divided into extraction, transmission, storage, calculation, mining and show.Among them, the transmission of large data is a prerequisite for the application of subsequent data. However, data transmission of big data applications is quite different with traditional data transmission. The main performance for three aspects: one is sudden characteristics of the data flow, the second is the data transfer tasks have priority, the third is data distribution has dispersion characteristic. So, realizing big data efficient transmission in network applications, especially the big data transmission within the wan, is the current major issues that need to be solved[5-9].

The Research Status of Home And Abroad

A wide area network data transmission technology is the foundation of big data applications,it is the core technology of the information age network application. The speed of its development and evolution is very fast. Transmission technology based on FTP protocol is to use the FTP protocol for transmission and sharing data between different platforms. FTP communicate need to generate two different connection, first, TCP generates a virtual connection used as the control of information. And TCP generates a separate TCP connections, mainly for the exchange transmission of data. FTP transmission can use two modes: active and passive mode. In active mode, the client and the server must be open at the same time, and listen to the same port, in order to establish a connection. But in this case, there will be effect the use of firewall. So FTP increases passive mode

again. Passive mode only need servers to set up a monitoring connection process. The client is not listening on port[10-15]. This can avoid the problem of client to install firewall.

The Big Data Transmission System

To achieve stability and highly effective transmission of the large data within the wan, the system uses Google Protocol Buffer, BDMQ message queue, TCP long connection and related technology, and combined with improvements of the flow control and congestion control methods, the distributed big data transmission system based on message queue is achieved. The overall structure of big data transmission system is shown in figure 1. According to the different function, transmission system is divided into four parts. The first part is the communication management module, mainly responsible for controlling communication thread. The second part is the datagram encapsulation module, mainly responsible for the packet encapsulation of transmission data. The third part is the node data transmission module, mainly responsible for the transfer of data transmission. The last part is the center data synchronization module, synchronization node data is mainly used for it. Communication management module is the foundation of the whole system, the other three modules complete big data transmission reliable and efficient based on it[16-20]. The design and realization of the four modules is discussed respectively in following four sections.

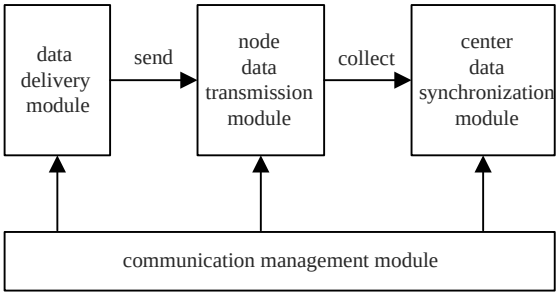

figure 1 The overall structure of big data transmission system

The Design of Interrupt Processing

When the sender sends data to the node receiver, as the disconnect of TCP connection has a certain delay, so the sender will continue to send data. The sender sends the data, but the recipient will not receive any news. This creates waste of resources. Even if the sender know, which part of the data needs to be retransmitted, but when the connection restored, the data can't insert in the forefront of the queue. If the rushing, jostle to resources, it can cause congestion. Due to the size of send and receive buffer limit, every once in a while, IO operation must be done for data, otherwise it will lead to a buffer overflow. When the system sends a large amount of data continuous, it can cause great burden to the communication connection. Transfer process will experience many jitter, affecting transmission efficiency. Due to excessive jitter, it also can cause the sender or the receiver shock phenomenon and interrupt transmission. In order to solve this problem, when the interrupt recovers, reply to the sender with a record number of message, the sender continues transmission from the breakpoint according to received the serial number[21-23]. As the transmission network is complex, data transmission may appear problem on the way such as cut off cable, hang up the other process, loss packet frequent. TCP connection can not be used in this time, but the application layer does not know. Heartbeat is that TCP sends determine whether there is connection after a period of time interval. If exist, the sender will return back a packet to ensure network effectively. If there is some wrong with the heartbeat package, the upper layer current network has the problem[24-25].

The Size of Buffer Setting

When the sender communicates with the receiver, data will be temporarily stored in the data buffer. Buffer capacity directly affect the rate of data transmission. The buffer size of sender can be set in the connected distributed nodes, and in a handshake connection, the receiving node will inform the sender to receive the size of the buffer. Through comparing buffer of sender with buffer of receiver, the big value is chosen as the size of the two buffer. The size of buffer must match with the size of send data. If the send data is less than the buffer capacity, when the node confirm to receive data, buffer will be tucked, prevent other data misinformation. When data is greater than the buffer capacity, the buffer will be expanding in order to achieve the purpose of collecting data. Otherwise it will cause data loss. The adjustment of buffer capacity is finished by the program automatically. The program will automatically detect whether data quantity match with the size of buffer. The expansion and contraction of the buffer is completed based on the results. The logic diagram of buffer automatic adjust is shown in figure 2.

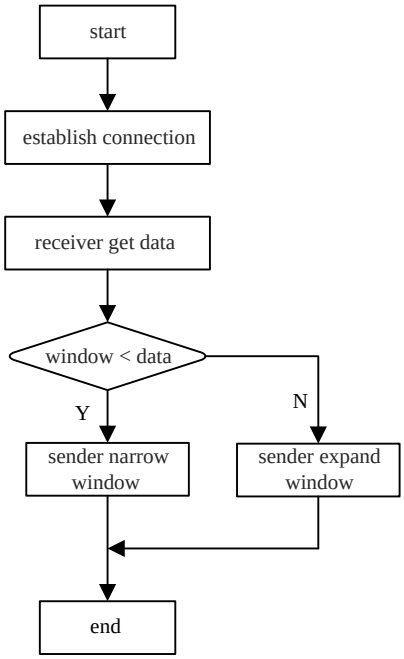

figure 2 the logic diagram of buffer automatical adjust

Conclusion

Big data era has come. The transmission technology is the basis of big data technology. It plays a crucial role for applications and research of the big data. Large data transmission technology based on wan is researched deeply in the paper.The distributed big data transmission system based on message queue is designed and implemented. It trys to satisfy most internet application products for large data transmission function and performance requirements. Through function and performance test of the transmission system, transmission system designed in this paper is proved to have a certain advantage on the transmission rate.

Reference

[1] D.X.Wei, C.Jin, S.H.Low eta1. Fast TCP: Motivation, Architecture, Algorithms. Performance. IEEE/ACM Transactions on Networking, 2006, 14(6): 1246-1259.

[2] J.LILT, W.BANG. Research SOA-Based Public Order Management Information System of Special Trade of Public Security. Computer Knowledge and Technology, 2008, 32(16):37-40.

[3] J.U, C.Chao. An Efficient Super Peer Overlay Construction and Broadcasting Scheme Based On

Perfect Difference Graph. IEEE Transactions 011 Parallel and Distributed Systems,2010,2l(5):594-606.

[4] G.FEDAK, H.HE, F.CAPPELLO. Bit Dew: A Programmable Environment for Large-Scale Data Management and Distribution. 2008 ACM/IEEE Conference on Supereomputing, 2008:10-12.

[5] W.Tan, Y.S.Fan, M.C.Zhou. A Petri Net-Based Method for Compatibility Analysis and Composition of Web Services in Business Process Execution Language. IEEE Transactions on Automation Science and Engineering, 2009,6(01):94-106.

[6] G.KOLA, M. IC VERNON. Target Bandwidth Sharing Using End Host Measures. Performance Evaluation, 2007,64(9-12):948-964.

[7] P.T.LIN, J.N.SHADID, M.SALA, eta1. Performance of a Parallel Algebraic Multilevel Preconditioner for Stabilized Finite Element Semiconductor Device Modeling. Journal of Computational Physics, 2009:228, 6250-6267.

[8] Quinlan J.R. Induction ofdecision tree[J]. Machine Learing, l 986, (1):81～106.

[9] Rajeev Rastogi and Kyuseok Shim. PUBLIC: A decision tree that integrates building and priming[C]. In Proceedings of 24th International Conference on Very Large Data Bases, New York, USA, Aug l 998, 404--415.

[10] Olaru C, Wehenkel L. A complete fuzzy decision tree technique[J]. Fuzzy Sets and Systems, 2003, 138(2):22 1-254.

[11] Pawlak Z, Skowron A. Rough sets and boolean reasoning. Information sciences, l77(2007):41-73.

[12] Lingras P, Hogo M, Snorek M. Temporal analysis of clusters of supermarket customers: conventional versus interval set approach. Informationsciences, 172(2005): 215-240.

[13] JinMao Wei. Rough set based approach to selection of node[J]. Intemational Journal of Computational Cognition, 2003, 1(2): 25-40.

[14] JinMao Wei, Shuuqin Wang. Novel approach to decision tree construction[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2004, 8(3):332-335.

[15]Quinlan J R. Discovering rules by induction from large collections ofexamples[J]. In Expert System in the Micro Electronic Age, 1979, 26～37.

[16] Niblett T, Bratko I. Learning decision rules in noisy domains[A]. Proceedings of Expert Systems[C].Cambridge University Press, 1986, 25-34.

[17] Barbara D, DuMouchel W, Faloutsos C, Hass P.J. The New Jersey data reduction report. Bulletion of the Technical Committee on Data Engineering, 20:43-45,1997.

[18] Palmer C.R, Faloutsos C. Density biased sampling：An improved method for data mining and clustering[C].ACM SIGMOD Record, 29(2):82-92,2000.

[19] Starzyk J, Nelson D.E, Sturtz K. Reduct generation in information system. Bulletin ofInternational Rough Set society,3(1/2):19-22,1999.

[20] Tan M. Cost—sensitive leaming of classification knowledge and its application inrobotics. Machine Learning,1993,13(1),30-33.

[21] Kira K, Rendell L. A practical approach to feature selection[J]. In:Proc.Internat.Conf. on Machine Learning,ICML, 1992:123-135.

[22] Tan M. Cost-sensitive learning of classification knowledge and its application in robotics. Machine Learning, 1993, 13(1), 1～33.

[23] Robnik—Sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RreliefF, Machine Learning, 2003, 23-69.

[24] I.Konoenko, S.J.Hong. Attribute selection for modeling. Future GenerationComputer Systems 13:181-195,1997.

[25] Lam W, Keung C.L. Leaming good prototypes for classification using filtering and abstraction of instance[J].PattemRecognition, 2002,35:1491-1506.