Evaluation of Objectivity/AMS on the Wide Area Network

SATO, H[1], MORITA, Y[1]

Institute1, High Energy Accelerator Research Organization (KEK), Tsukuba, JAPAN

Abstract

As the number of physicists increases and will be spread over the world in the coming generations of HEP experiments, the needs to distribute the obtained data near to the physicists' home institute also increases. Thus the importance of the effective utilization of the wide area network (WAN) increases at unprecedented scale.

In this report we present the details of behavior and the performance of Objectivity/Advanced Multithreaded Server (AMS) on the dedicated link including the satellite network with relatively large latency. This link gave us a unique opportunity to test the distributed data analysis in the WAN environment. We have measured the performances of remote database access using the Objectivity/AMS.

Keywords: Objectivity, WAN

1 Introduction

In the next generations of HEP experiments, the needs to access the obtained data via the WAN and distribute those near to the physicist's home institute increases. However it is not efficient that the physicists in the world do an analysis in the same method. Because the round trip time (RTT) of several hundred times between the client where the user does analysis and the server where the data is stored is different. In this report we have prepared the analysis client machine and the data server machine, and located each in different site connected with the WAN. And various measurement in this environment has been done. An Objectivity/DB 5.1 system has been used as the client-server software.

2 Setup for Measurements

2.1 Machine at KEK and CERN

At KEK, we have prepared a SUN Ultra-60 workstation named "arksol1" operated by dual SPARC processors at 360 MHz with 1 GB of memory and 100GB disk. At CERN[1], a SUN Ultra-4 Enterprise 450 server named "monarc01" has been dedicated. It has quad SPARC processes at 400 MHz with 512 MB of memory and two 9 GB disks.

2.2 Network between KEK and CERN

In this measurement we have employed the different kind of WAN connection. One is the public network link dedicated by the National Center for Science Information Systems (NACSIS) of the Ministry of Education, Science, Sports and Culture. It is shared

[1]Thanks the staffs who support the machine and the network.

by many academic institutes between Japan and Europe. The other link is dedicated by the Communications Research Laboratory (CRL) of the Ministry of Posts and Telecommunications as one of Japan-Europe Gamma (JEG) project[2]. It is the 2 Mbps link of INTELSAT satellite. The activity of the other institutes is low so that we can use this link with the full bandwidth. The RTTs are about 310 ms and 655 ms, respectively.

At KEK, we have prepared two network routers with the different configurations. On the primary router a packet to the monarc01 was set in order to send in the NACSIS line and on the secondary one it was set in the JEG line. One can easily select the route to the monarc01, manipulating the network routing tables with the route command.

2.3 Objectivity/DB Configuration

Figure 1 illustrates the Objectivity/DB configuration in this measurement. A federated database file is the highest level Objectivity/DB logical storage hierarchy and contains the system database, which stores the schema and the catalog. This federated database has a boot file, which contains information used by an application or tool to locate and open it. In a journal file update information of a transaction is recorded. It enable Objectivity/DB to recover the system when the transaction is aborted or terminated abnormally. A database file is the second highest level Objectivity/DB logical storage hierarchy and is where the application's persistent data is stored. Using Objectivity's Advanced Multithreaded Server (AMS) as data server software, we can distribute these files described in the above to other remote hosts on the network.

This configuration is the simplest client-and-server model in the case of using the AMS. Since the application process uses the lock server and the federated database on the host where it runs, the resources accessed over the network are only the database files. Monitoring the packets on the network, we can know the detailed negotiation between the client and the server.

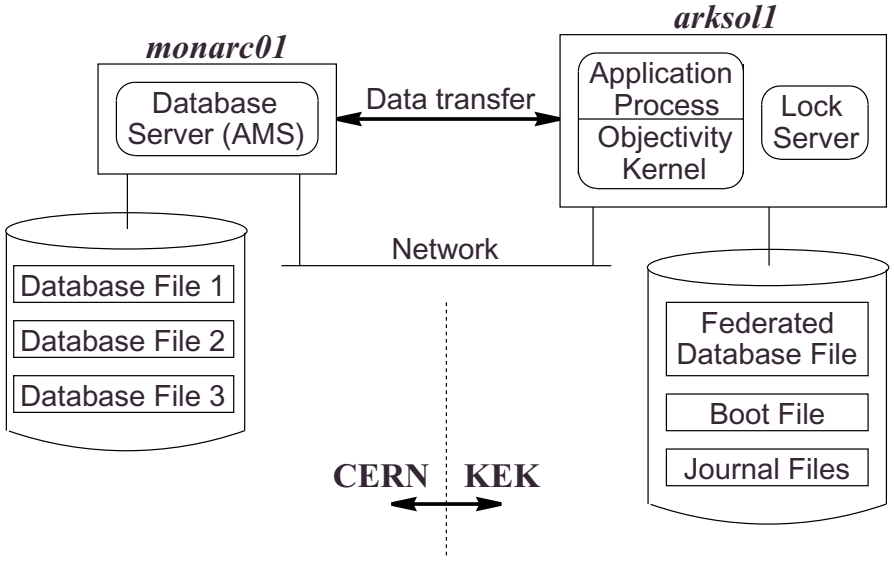

Figure 1: Objectivity/DB Configuration

[2]Thanks the JEG project team.

3 Measurement and Result

At first, we have written the data objects in the database files on the server where the AMS is running. The page size was set to 8192 bytes (default size). After writing, those have been read from the client where the Objectivity/DB application is running. While these writing and reading were in progress, we have been monitoring the packets on the network with the tcpdump command.

3.1 Writing Transfer

While a new object is being created by the application, the transfers between the client and the server can be divided into two kinds (Figure 2). In a Control-Transfer-Phase (CTP) the created object is not transferred, and only control packet is transferred to the server. In a Data-Transfer-Phase (DTP) the created object is transferred. However as soon as the object is created, it is not sent to the server. Since the Objectivity kernel has a buffer, the objects are stored there in the beginning of the transaction. Those are transferred when the buffer is full.

3.2 Reading Transfer

A reading transfer is simpler than the writing transfer. The client requires the server to transfer the page which a necessary object is within, and the server sends it simply (Figure 2). However in using the satellite link the phenomenon that transfer time for one page changes periodically has been observed (Figure 3). Using the client-server program we wrote which does the packet exchange by the same pattern, the same phenomenon has been seen. By analysing the TCP/IP packets in the transaction with the tcpdump, it was found that the initialization (set to one segment) of the "congestion window" occurred periodically due to the retransmission. This forces "slow start" to occur for the next page transaction[1].

4 Discussion and Conclusions

In the transaction of the Objectivity/DB the data transfer is done by the page size unit and it uses the handshaking algorithm on the application layer. However such an implementation does not take the usage on the WAN with the large latency into consideration and the bandwidth of the network is not used effectively. The improvement of the performance of the Objectivity/DB can be expected by the change of this handshaking algorithm.

In this measurement we found that the Objectivity's behavior on the WAN with the large latency is very different from that on the LAN. This is due to congestion avoidance and slow start on the WAN. Future development of the network applications should take into account these effects.

References

1 Stevens, W.R. 1994. "TCP/IP Illustrated, Volume 1/2: The Protocols", Addison-Wesley, Reading, Mass.
Wright, G.R., and Stevens, W.R. 1995. "TCP/IP Illustrated, Volume 2: The Implementation", Addison-Wesley, Reading, Mass.

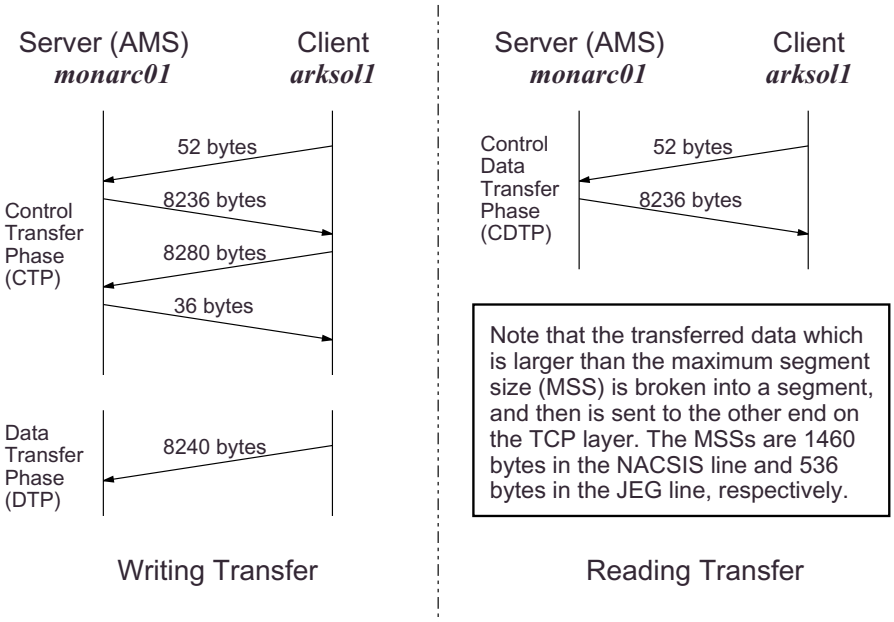

Figure 2: Data Transfer for Writing and Reading

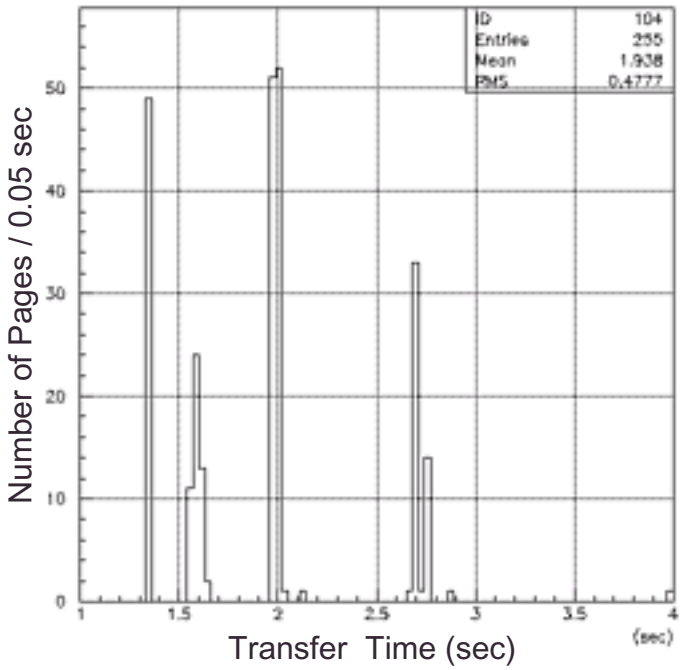

Figure 3: Transfer Time for Reading 1 Page