Journal of Southeast University (English Edition) Vol.22, No.1, pp.43-47 Mar. 2006 ISSN 1003-7985

Ground data transfer research in AMS-02

Wu Hua Gong Jian Zhou Yu

(Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

(Key Laboratory of Computer Network Technology of Jiangsu Province, Nanjing 210096, China)

Abstract To increase the performance of bulk data transfer mission with ultra-long TCP (transmission control protocol) connection in high-energy physics experiments, a series of experiments were conducted to explore the way to enhance the transmission efficiency. This paper introduces the overall structure of RC@SEU (regional center@Southeast University) in AMS (alpha magnetic spectrometer)-02 ground data transfer system as well as the experiments conducted in CERNET (China Education and Research Network)/CERNET2 and global academic Internet. The effects of the number of parallel streams and TCP buffer size are tested. The test confirms that in the current circumstance of CERNET, to find the right number of parallel TCP connections is the main method to improve the throughput. TCP buffer size tuning has little effect now, but may have good effects when the available bandwidth becomes higher.

Key words bulk data transfer; performance; TCP buffer size; parallel stream; IPv6

AMS (alpha magnetic spectrometer)-02 is an international cooperative particle experiment in space[1]. This experiment will keep a magnetic spectrometer on the ISS for a period of about 3 to 5 years, with the purpose of performing accurate, high statistics, long duration measurements of the spectra of energetic primary charged cosmic rays in space. Its early prototype, AMS-01, was carried by the US Space shuttle in 1997 to space for a 10 d operation with fruitful observation results achieved.

The measured data of the AMS-02 will be transmitted to CERN (European Organization for Nuclear research at Geneva, Switzerland) via NASA, and finally distributed to different regional centers (RC) for physics research purposes. The RC@SEU, located at Southeast University, is the only RC in the Asian region. According to the design, the measured raw data and the auxiliary ones are transferred to RCs via a global academic network for a duration of about 5 years. This is an ultra-long data transmission task which requires certain stable bandwidths, and the currently available data transmission tools are insufficient for meeting such needs because there is no QoS guaranteed at the current global Internet being used. Therefore, a special transfer system needs to be developed for this purpose. This paper introduces the implementation model of RC@SEU for the AMS-02 data transfer system, based on a series of feasibility experiments conducted on CERNET and CERNET2.

Received 2005-07-01.

Foundation item: The National Basic Research Program of China (973 Program) (No. 2003CB314803).

Biographies: Wu Hua (1973—), female, graduate; Gong Jian (corresponding author), male, doctor, professor, jgong@njnet.edu.cn.

1 Structure of the RC@SEU Data Transfer System

The RC@SEU data transfer system should support not only the data transmission with SOC@CERN, the data center for AMS-02, and data distribution to all the users involved, but also video conferencing (VRVS, developed by the California Institute of Technology, USA) and remote access services for the cooperative work among AMS-02 partners. Therefore, RC@SEU should provide guaranteed bandwidth and transmission quality to support these requirements.

The global network environment used by all the AMS-02 partners is heterogeneous one. The Chinese partners are using CERNET2, a native IPv6 network. The Internet used in the USA, e.g. Abilene, is dual-stack based, and almost all the European partners are still using IPv4 network. The BBFTP[2], an FTP-based parallel bulk data transfer tool used for the AMS-02 data transmission system, works on both IPv4 and IPv6 platforms, but the VRVS only has IPv4 version. To integrate them together, tunneling technology is used to connect the IPv4 and IPv6 systems together. The overall system structure of RC@SEU based on the Linux environment is shown in Fig.1. The core component of the system is the tunnel server (TS) which

- Integrates the IPv4 over IPv6 tunneling services with the AMS-02 ground data transmission system;

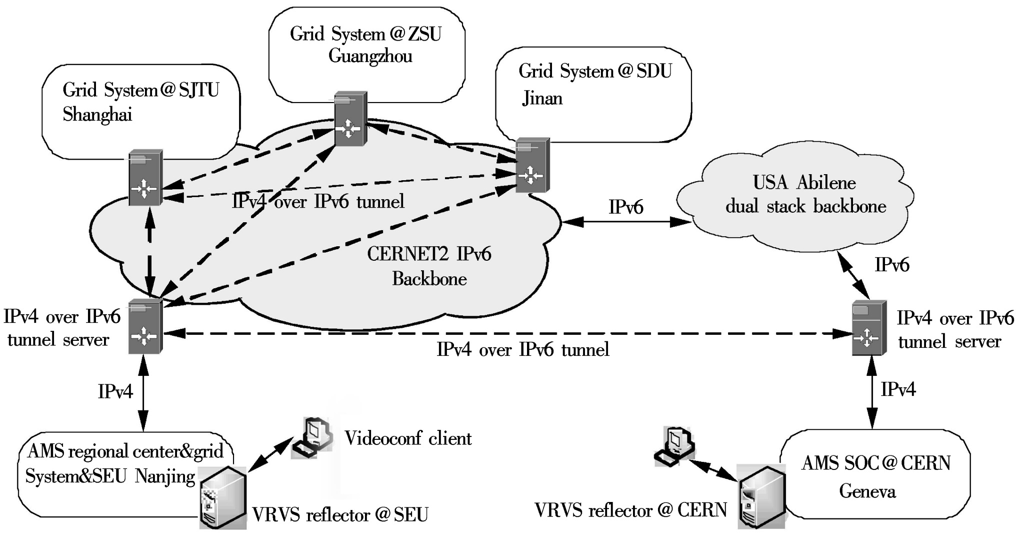

Fig 1 Overall structure of RC@SEU

- Provides priority-based QoS function embedded in the Linux kernel;
- Manages the tunnel operation, including tunnel maintenance, throughput, path MTU, tunnel soft states, etc.

2 Buffer Size and Stream Numbers

The file transfer protocol uses a TCP connection to send and receive data. And a standard TCP connection uses a window-based method to control the bandwidth being used to guarantee the stability and fairness on resource sharing. The TCP implements flow control and congestion control via a flow-control window (fwnd) that is advertised by the receiver to the sender and a congestion-control window (cwnd) that is adapted by the sender based on the inferred state of the network. The maximum fwnd and cwnd is related to the amount of buffer space that the kernel allocates for each socket. For each socket, there is a default value for the buffer size, which can be changed by the program using a system library call just before opening the socket. The buffer size can be adjusted for both the send and receive ends of the socket. If they are set below the BDP (bandwidth delay production), this will cause a performance bottleneck on a high bandwidth and a long delay link, and make it impossible to take full advantage of the available bandwidth. For the interoperability among all the RCs, we cannot modify the TCP congestion control algorithm for RC@SEU implementation. There are two main methods to improve this situation, one is to improve the numbers of parallel streams, the other is to adjust the TCP window when setting up a TCP connection, and this will have an effect on the value of fwnd and cwnd[3-5].

It has been proved by experiments that increasing the numbers of parallel streams will increase the transfer speed for sure. But if the number is over-increased, the transfer speed will decrease because there will be a shortage of resources that are needed for connection operations[6]. Besides this, to increase the parallel streams for one application is not fair to other users of this link, because it breaks the fair principle among network users, so some ISP will assume that it is a kind of "denial-of-service" attack, and refuse to provide service to such users. Therefore, it seems that optimizing the buffer size of TCP connections is a better choice to this problem[7].

WEB100[8] did much work to make full use of the link by adjusting the TCP buffer according to the BDP measured. But there should be a trade-off between the buffer size and the stream numbers. SLAC[9] compared the performance with different TCP buffer sizes and different parallel streams. The results showed that when the buffer size was relatively small, the transfer speed will increase linearity with the stream numbers. Under such a circumstance, the link was not saturated, so to increase the number was a good method. However, as the buffer size increases, the speed will decrease because every connection needs to consume system resources (memory, process and CPU). So the problem is how to achieve the best speed by properly setting the two parameters according to the network status that AMS-02 faces. Actually, since the path between RC@SEU and SOC@CERN will pass across the three largest academic networks in the world, past experiences may not be suitable for this situation. For such an ultra-long session, delay and available bandwidth fluctuate from time to time, so that the system should be able to adjust these parameters automatically.

3 Data Transfer Experiments on IPv4 Network

To find out the suitable buffer size and stream numbers for RC@SEU under the specific global Internet environment, a series of experiments have been conducted to verity the feasibility of the RC@SEU design.

The two end systems for IPv4 network experiments are located at Southeast University (Nanjing, China) and CERN (Geneva, Switzerland). The transmission path passes through CERNET, vBNS in the USA, and CERN network (interconnected through StarLight in Chicago, USA). Iperf[10] was used to measure the available bandwidth of the path. This test found that the path was composed of about 18 hops, and there was a configuration problem at StarLight which made the path asymmetrical (the problem was fixed later). The available bandwidth of single TCP connection is about 1.1Mbit/s.

BBFTP was used for the experiments between SEU and CERN, which allowed users to adjust manually the parallel stream numbers and the TCP buffer size. Six experiments were conducted with parallel number settings from 1 to 40 each time. The average result is shown in Fig. 2. We try to find a polynomial $p(x)$ (x is the number of parallel streams) of order 2 that fits the data in a least square sense. We obtain it as follows:

$$P(x) = -0.0083x^2 + 0.3915x + 1.4962 \quad (1)$$

The maximum value of $P(x)$ is 6.11Mbit/s when x is 23.36. That means on this path, we can get the maximum throughput by setting the parallel streams to 23.

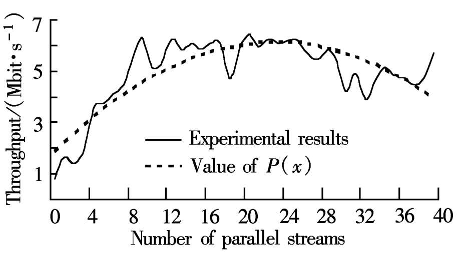

Fig. 2 Throughput with different parallel stream numbers

The test results prove the basic equilibrium principle of TCP, so that the throughput can be increased by increasing the parallel streams. However, according to Ref. [11], if an application uses n TCP streams between two hosts, the aggregate bandwidth of all n TCP connections B can be roughly expressed as

$$B \leqslant \frac{S_{MSS}}{t_{RTT}} \left(\frac{1}{\sqrt{p_1}} + \frac{1}{\sqrt{p_2}} + \dots + \frac{1}{\sqrt{p_n}} \right) \quad (2)$$

where S_{MSS} is the maximum segment size, t_{RTT} is the round trip time and p_i is the packet loss ration of the i-th connection. Among these parameters, the packet loss rate p_i is a primary factor in determining aggregate TCP throughput of parallel TCP connection sessions. As discovered in Ref. [12], the loss rate over a wide area network was mostly caused by physical error and/or congestion. Advanced communication technologies can ensure us a reliable infrastructure for the current global Internet, so that the main reason for packet loss is congestion which makes the parallel TCP connections compete with each other as well as with other traffic, and make the bandwidth of the aggregate TCP sessions decrease.

It is also well known that when the TCP sender buffer is set to the BDP, the throughput of a single TCP connection may be the best. We used Iperf and ping to get the bandwidth and RTT of the path. The experiment lasted for about two days, to see the variation of the BDP. Fig. 3 is the BDP over the path from SEU to CERN at 30 min intervals.

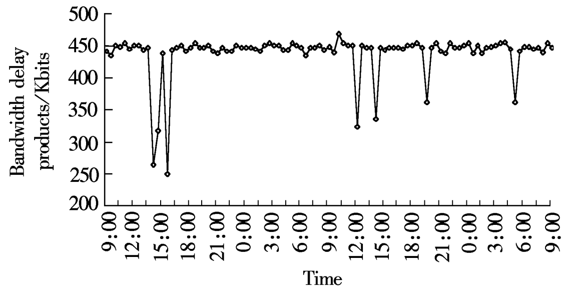

Fig. 3 BDP on the path from SEU to CERN

It can be seen that although there are 18 hops on the path, the value of BDP is relatively stable during the time of the two days. The mean value of BDP is 437 Kbits, and the standard deviation is 36 Kbits. Based on the above results, we can suppose that the BDP over the lifetime of a connection does not change very much unless something unusual happens in the global Internet, so the value of the TCP buffer size should be stable for the connection. What we need to do is to adjust the TCP buffer size when the traffic behavior becomes a long-term bandwidth utilization management for this ultra-long session.

Fig. 4 shows the experimental results measured between SEU and CERN by increasing the TCP buffer size from 10 Kbits to 800 Kbits.

The polynomials that fit the data in a least square sense are

$$P_1 = -0.0001x^2 + 0.0177x + 0.1357 \qquad 0 < x \leqslant 90 \quad (3)$$

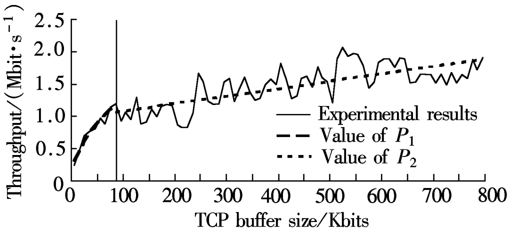

Fig. 4 Throughput with different buffer sizes

$$P_2 = 0.0012x + 0.9568 \qquad 90 < x < 800 \qquad (4)$$

It can be seen that although the BDP is about 450 Kbits, the throughput has an obvious increasing trend when the buffer is set below 90 Kbits. So during this period, the buffer size is the bottleneck of transfer, but after 90 Kbits, the increasing trend is not so obvious. There are no special changes of throughput at the point around BDP. It should also be noted that when the TCP buffer is set to 100 Kbits, the throughput has reached 1.1 Mbit/s, which is the measurement result of Iperf for one TCP connection. That reflects two facts: one is that the result of Iperf is believable; the other is after the TCP buffer is set above 100 Kbits in BBFTP, the main bottleneck changes from the TCP buffer to the available bandwidth of the network path.

In order to determine the bottleneck after the buffer size is set above 100 Kbits. We set the TCP buffer size to 100 Kbits and transfer a 10M file from CERN to SEU while using tcpdump to listen to the process of the file transfer. In Fig. 5, the broken line is the window advertised by the receiver (rwnd). This value is affected by the TCP buffer size, and the continuous curve is the value of the instantaneous outstanding data samples at various points in the lifetime of the connection. We can see that, although the rwnd is large, when congestion occurs, both the rwnd and outstanding data decrease. It proves that the throughput cannot be increased by increasing the TCP buffer size after the available bandwidth has been reached.

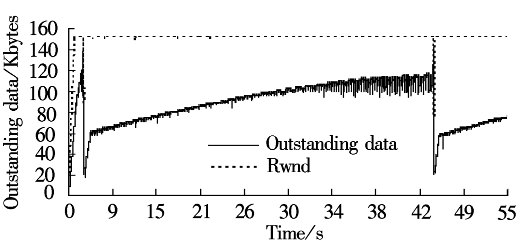

Fig. 5 Outstanding data graph

So we can conclude that for a non high bandwidth path, the performance of the data transfer system will greatly depend on the number of parallel stream numbers.

4 Data Transfer Experiments on IPv6 Network

The special features of AMS-02 ground data transmission make it a typical application for next generation Internet CNGI (China Next Generation Internet) is an IPv6 promotion program sponsored by 8 ministries of the Chinese central government. Several national IPv6 backbones will be set up by carriers and educational institutions. CERNET2 is the largest one among them, with one of its Gigapops located at SEU. The CERNET2 backbone has been in operation since December, 2004. And CERN has also been connected with several IPv6 backbones in Europe and USA. These bring us a chance to transport RC@SEU onto the IPv6 network.

By using traceroute6, we confirm that the transmission path between CERN and SEU and its RTT are consistent with the ones in the IPv4 network. Some experiments were carried out using BBFTP which had been transported to IPv6 platform. BDPs of the TCP connections for these experiments were measured. Because of page limitation, the results cannot be listed here and will be described in another paper in detail. The results confirm the observations we made on the IPv4 network. This is obvious because TCP implementations are the same for both IPv4 and IPv6 network. These experiments also confirm for us that many traditional applications on IPv4 network can be easily transported onto IPv6 network.

5 Conclusion

AMS-02 ground data transfer is an ultra-long bulk data transmission on the global Internet, so it is suitable for the IPv6-based next generation Internet world-wide. By defining a special tunnel server, not only can the system be implemented in a heterogeneous network environment with IPv4 and IPv6 backbone co-existing, but also the transfer service quality can be guaranteed to some extent when the QoS of global Internet varies dynamically in order to meet the requirements of the AMS-02 mission.

To improve the efficiency of BBFTP so as to make better use of the precious bandwidth of the global Internet, a series of experiments were conducted in the real network environment, to study the effects of parallel streams and the TCP buffer size on the throughput of TCP connections. The results show that the positive effects of parallel streams on TCP connection throughput are more advantageous than changing TCP buffer size under the current circumstances. So a measurement tool can be embedded into BBFTP to find the optimal value of the parallel streams; and when the path becomes high bandwidth in the future, the tool can also be of use in finding the BDP value termly to make the TCP buffer size adjustment possible.

We believe that the experiences gained in implementing and testing such an application will be of benefit to all similar applications which may appear in CNGI.

Acknowledgments We would like to thank Professor Li Xing of CERNET Network Center for helping us set up the experiment environment in CERNET and Dr Alexei Klimentov and Dr Alexandre Eline of MIT in the AMS project for providing the data interface. And we also would like to thank Dr. Edoardo Martelli of CERN for helping us set up the experiment environment in CERN.

References

[1] AMS 02 homepage[EB/OL]. (2005-03-22) [2005-06-22]. http://ams.cern.ch/.

[2] BBFTP website[EB/OL]. (2005-05-30) [2005-06-22]. http://doc.in2p3.fr/bbftp/.

[3] Jacobson V, Braden R, Borman D. RFC 1323: TCP extensions for high performance [EB/OL]. (1992-05) [2005-06-22]. http://www.apps.ietf.org/rfc/rfc1323.html

[4] Tierney B. TCP tuning guide for distributed applications on wide area networks [J]. Usenix & SAGE Login, 2001, **26**(1): 33-39.

[5] Jain M, Prasad R S, Dovrolis C. The TCP bandwidth-delay product revisited: network buffering, cross traffic, and socket buffer auto-sizing [EB/OL]. (2003-02) [2005-06-22]. http://www.cercs.gatech.edu/tech-reports/.

[6] Sivakumar H, Bailey S, Grossman R L. PSockets: the case for application-level network striping for data intensive applications using high speed wide area networks[C] // Proceedings of IEEE Supercomputing 2000. Dallas, TX, USA, 2000: 240-246.

[7] Semke Jeffrey, Mahdavi Jamshid, Mathis Matthew. Automatic TCP buffer tuning[C] // Proc of ACM SIGCOMM. ACM Press, 1998: 315-323.

[8] Web100 concept paper[EB/OL]. (1999-09-29) [2005-06-22]. http://www.web100.org/docs/concept_paper.php.

[9] Cottrell R Les, Logg Connie, Mei I-Heng. Experiences and results from a new high performance network and application monitoring toolkit[C] //PAM 2003 Workshop. La Jolla, CA, USA, 2003: 205-217.

[10] Iperf website[EB/OL]. (2005-05-03) [2005-06-22]. http://dast.nlanr.net/Projects/Iperf.

[11] Hacker T, Athey B, Nobel B. The end-to-end performance effects of parallel TCP sockets on a lossy wide-area network[C] // 16th IEEE-CS/ACM International Parallel and Distributed Processing Symposium. Fort Lauderdale, FL, USA, 2002: 314-329.

[12] Hacker T, Nobel B, Athey B. The effects of systemic packet loss on aggregate TCP flows[C] // ACM/IEEE Conference on Supercomputing. Baltimore, Maryland, 2002: 270-285.

AMS-02地面传输系统

吴 桦 龚 俭 周 渔

(东南大学计算机科学与工程系, 南京 210096)
(江苏省计算机网络技术重点实验室, 南京 210096)

摘要: 为了研究在高能物理试验中使用超长 TCP 连接进行数据传输的最佳效率, 分别在 IPv4和IPv6网络协议上进行了一系列的试验寻求优化方法. 介绍了 AMS-02项目中东南大学地面传输系统的总体结构, 以及在 CERNET/CERNET2和欧洲粒子实验室之间进行的数据传输试验, 试验主要研究了并行流数目和TCP缓冲大小对传输效率的影响. 结果表明: 在目前 CERNET的环境下, 找到最佳的并行流数目是提高传输效率的主要方法; TCP缓冲调节在目前的条件下效果不明显, 但是当可用带宽显著提高后会有明显的效果.

关键词: 海量数据传输; 效率; TCP缓冲; 并行流; IPv6
中图分类号: TP393.06