Analysis of Multiple Target Network Storage Access using Multi-routing VPN Connections

Nozomi Chishima
Ochanomizu University
2–1–1, Otsuka, Bunkyo-ku
Tokyo 112–8610, JAPAN
nozomi@ogl.is.ocha.ac.jp

Saneyasu Yamaguchi
Kogakuin University
1–24–2, Nishishinjuku, Shinjuku
Tokyo 163–8677, JAPAN
sane@cc.kogakuin.ac.jp

Masato Oguchi
Ochanomizu University
2–1–1, Otsuka, Bunkyo-ku
Tokyo 112–8610, JAPAN
oguchi@computer.org

Abstract

As the introduction of SAN progresses for the purpose of the storage management cost reduction, iSCSI is expected as a representative of IP-SAN that uses IP network. However, SAN is mostly used only in the server site currently. Thus we have claimed iSCSI can be employed in a WAN environment using VPN, as well as in a local environment.

In this paper, the behavior of the TCP Congestion Window is observed and the performance of the system is evaluated when iSCSI access through multiple connections is performed using a multi-routing function of a VPN router. Furthermore, we have examined the iSCSI access method with mutiple connections on VPN when a parallel storage system is used as Target of network storage.

1. Introduction

Recently, IP-Storage Area Network (SAN) configured with inexpensive Ethernet and TCP/IP is introduced. As a standard, iSCSI protocol is becoming important in IP-SAN[1][2]. iSCSI encapsulates SCSI command, within a TCP/IP packet and transports the volume of data between server (Initiator) and storage (Target). As the realization of the gigabit/10gigabit class line is expected according to the advancement of Internet, iSCSI will be effective furthermore.

In the present condition, SAN is mainly used only in the server site. However, SAN is expected to connect between server and storage of a remote site for the sake of remote backup. Thus we have claimed iSCSI used in a local environment can be applied to the WAN using Virtual Private Network (VPN). Furthermore, we propose the method to build multi-routing connection on the VPN established through WAN, for the realization of higher performance and reliability of the connection.

A large-scale storage system consists of multiple storage devices. Such system has a single access interface, single IP address for example, though it contains a lot of storage device actually. In the case of using such a high-performance parallel storage system, including autonomous disks, the network tends to become a bottleneck. Thus the multi-routing access is considered to be effecrive in this case.

2. Background of Our Research Works

2.1 iSCSI

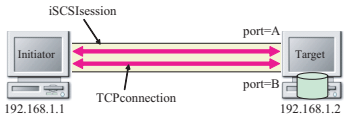

Figure 1. iSCSI multiple connections

iSCSI is the most popular protocol of IP-SAN. iSCSI is a standard to encapsulate a SCSI command into a TCP/IP packet, so we can build SAN only with IP devices when we use iSCSI. On the other hand, it has complicated hierarchical structure. In addition, although iSCSI is expected for the realization of the long-distance remote storage access, it has the problem of Long Fat Pipe in the case of using the gigabit connection. Therefore appropriate control of the TCP/IP layer is required[3].

iSCSI can be variously tuned. We can establish multiple TCP connections in a single iSCSI session by the implementation of UNH-iSCSI offered by University of New Hampshire[4], which is used in this experiment. iSCSI standard requires that all iSCSI PDUs related to a single SCSI command must all be sent and received on a single connection. Different read commands in the same session can use different connections. The choice of which connection to use for a command is made by the initiator. The target must always respond to that command on the same connection.

That is, as shown in figure 1, multiple connections having different port numbers can be connected with a single IP address and a single iSCSI drive of a target.

2.2 VPN

VPN is technology to connect with virtually closed network using a public network of a carrier including Internet. While making use of the inexpensive public network, VPN is realized as a "substantial exclusive line" by compensating problem of insecurity with another method such as encryption. On the other hand, unlike an exclusive line, QoS of a network is not guaranteed.

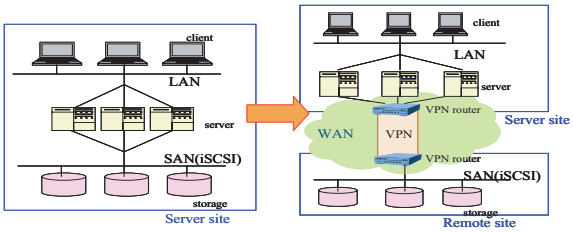

Figure 2. Utilization model of VPN

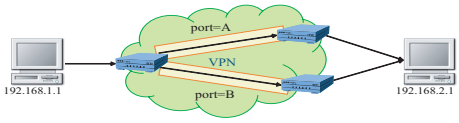

Figure 3. VPN multi-routing function

In this study, we have installed network storage in the remote environment connected with VPN router, and access over VPN in the WAN, in order to perform secure remote backup using iSCSI. This is shown in figure 2. In this case, by going through a VPN router, network bandwidth is restricted so that a throughput should degrade remarkably [5]. Furthermore, it is assumed that communication path in the WAN is unstable. Thus in this paper, we have considered to connect the multi-routing connection inside the VPN. Thereby, the reliability of the data transmission and the fault tolerance of the network should be improved.

VPN router used in this experiment, Fujitsu Si-R570, has the multi-routing function [6]. With this function, it becomes possible to transmit packets through multiple routes to a destination which has a single IP address, using information such as port numbers. This is shown in figure 3. Since a communication route can be divided based on each communication contents, we can set up, for example, one connection is used for transmitting voice data and another connection is used for the rest types of data. In this paper, we use this function for configuring a different route for each connection that is correpondent with iSCSI multiple connections.

2.3 TCP CWND

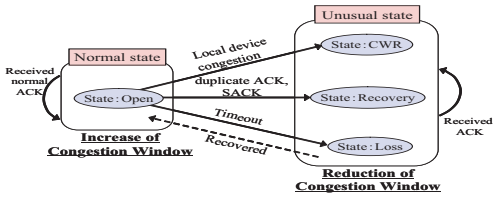

Figure 4. State transition of LinuxTCP implementaion

TCP uses the concept of Congestion Window (CWND) for congestion control. CWND means the number of maximum packets that can be sent consecutively without receiving a reply packet of Acknowledgement (ACK) from a data sender. That is to say, CWND is a parameter, which limits the behavior of a data sender, for the purpose of the network congestion control. Generally, CWND increases whenever a data sender receives one ACK. If the state of communications of TCP is judged as normal, CWND increases every time the data sender receives one ACK. However, If TCP implementation detects an error and judges the state of communication as unusual, CWND reduces dramatically. In Linux TCP implementation, the cases in which CWND reduces are as follows (Figure 4).

1, CWR: Detecting Local Congestion error in which device driver buffer of the data sender overflows.

2, Recovery: Receiving duplicated ACKs or SACK.

3, Loss: Detecting timeout.

Linux TCP Implementation doesn't increase the window size unless CWND is consumed completely before receiving ACKs. In such a case, we confirmed the throughput remains stable.

3 Multiple Target Network Storage System

3.1 Parallel Storage System

Recently, a large-scale storage system consists of a parallel storage connecting a lot of storage nodes with a network. Data is divided, distributed, and stored into each file, extent, page, or block. In order to hide data disposition management from user, the system is virtualized with the storage virtualization mechanism using a metadata server and a distributed directory. This is offered as a single huge volume. Access requests from each client are sent to an appropriate storage node through the virtualization mechanism and they are processed on it. The parallel storage system is constituted that it maintains the backup data of other nodes so as to prevent the data loss by the node failure.

In the study of the parallel storage system, the autonomous disk as a high-performance parallel storage system, which is superior in availability and scalability, has been proposed until now [7]. With the autonomous disk, reliability is raised by the Chained Declustering replica arrangement strategy, in which backup data is maintained for obstacle recovery by having a copy of Primary-Backup between storage nodes.

3.2 Multiple Target Network Storage System

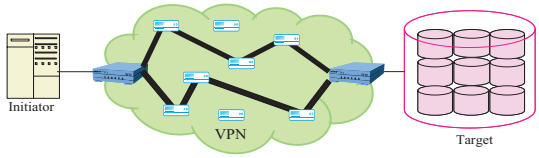

Figure 5. Multiple Target network storage system

With the parallel storage system described in the previous subsection, multiple storage devices are bundled, and offered as a single huge volume storage. When we use it as a target of network storage such as iSCSI, although they are multiple targets actually, they look as a single storage having a single IP address from the server side.

In the case of using a high performance storage system as a target like a parallel storage, it is likely that the bottleneck of the network storage is at a network part. Therefore, it is important to raise the ability of the network part in order to improve the total performance. For this purpose, the method of the multiple route connection such as the VPN multi-routing described in section 2.2 is effective.

Therefore, we propose and evaluate a method of the multiple route connection on the VPN for the better performance, when we configure a multiple target network storage system using a parallel storage as a target. This is shown in figure 5. Even though the target has a single IP address and the server accesses it through a single interface, the network part is configured with multiple routes and the target is a parallel storage system including multiple targets.

3.3 Previous Research Works

We have observed behavior of the network and TCP CWND, and evaluated the performance on iSCSI storage access using VPN [5][8].

In order to control TCP CWND on iSCSI storage access, the technique of controlling a CWND value dynamically is proposed [9]. In this method, monitor functions are inserted in the kernel of Target OS. Target monitors CWND and observes its change, and notifies the value to Initiator. When Initiator receives the notification, the middleware decides the block size based on CWND, and the application on Initiator modifies the block size and performs sequential read access. In a long delay environment, iSCSI network throughput has improved about 28% maximum in this experiment.

Furthermore, with iSCSI multiple connections, protocol tuning of iSCSI and a related protocol layer is realized. As a result, its validity is confirmed[10].

In this paper, we have examined the access of iSCSI multiple connections on VPN multiple routes. With this access method, not only the improvement of performance but also the improvement in reliability is expected. In addition, we have proposed an access method in the case of using a parallel storage system and obtained its ideal value.

4 Experimental System

4.1 TCP CWND Monitor Tool

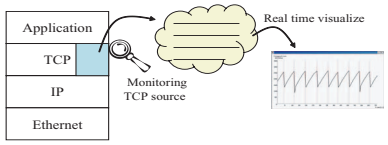

Figure 6. TCP CWND monitor tool

Generally, user programs cannot recognize the size of CWND because CWND is a parameter controlled in a Kernel space of an operating system. Therefore we have inserted monitor functions in TCP source code and implemented a recording mechanism of TCP parameters within a Kernel memory space, so that they are accessible from user space as shown in figure 6. With this mechanism, we can observe TCP parameters by reading a special file for accessing Kernel memory space. What we can monitor with this tool is the value of CWND and various error events (Local device congestion, duplicate ACK/SACK, Timeout). Besides it is able to visualize the value as a graph using the X11 window system library function on real time.

4.2 VPN Multiple Routes Access Control System

In this experiment, we have configured multiple routes using VPN routers to evaluate performance and CWND in iSCSI storage access. We have inserted 4 VPN routers between Initiator and Target, so that multiple routes accesses can be performed. Furthermore, we have inserted 2 FreeBSD Dummynets, which are artificial delay equipment simulation a long distance access [11].

We have set up this experimental system using iSCSI multiple connection and VPN multi-routing function. In

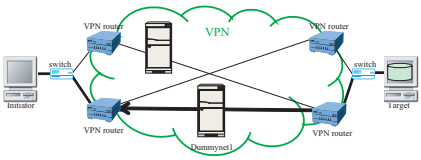

Figure 7. iSCSI Single connection/ VPN Single route

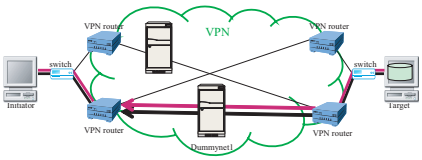

Figure 8. iSCSI Multiple connections/ VPN Single route

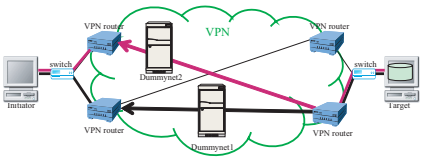

Figure 9. iSCSI Multiple connections/ VPN Multi-routing

the following, while we explain iSCSI read access case, in which data is transferred from Target to Initiator, write access case is almost the same basically.

First, in the case of the iSCSI Single connection/VPN Single route, packets route is as shown in figure 7. Second, in the case of the iSCSI Multiple connections/VPN Single route, 2 connections of iSCSI are on the same route, as shown in figure 8. Finally, in the case of the iSCSI Multiple connection/VPN Multi-routing, each connection is on the different route, as shown in figure 9.

As Initiator and Target of our experimental system, OS is Linux2.4.18-3, CPU is Intel Xeon 2.4GHz, Main Memory is 512MB DDR SDRAM, NIC is Intel Pro/1000XT Server Adapter on PCI-X (64bit,100MHz), and iSCSI is UNH IOL reference implementation ver.3 on iSCSI Draft 18[4]. OS of Dummynet1 is FreeBSD4.9-RELEASE, and that of Dummynet2 is FreeBSD6.2-RELEASE. VPN router is Fujitsu Si-R570 [6]. This router achieves 500Mbps maximum as 3DES encryption speed.

In order to evaluate only the performance of storage access, in the Initiator side, the influence of cache has been eliminated by using raw device in this experiment. Moreover, in order to focus on the network performance of iSCSI access, Target has been operated as Memory Mode that UNH-iSCSI offers, so as not to be accompanied with disk access.

5 Experiment With VPN Multi-routing

5.1 Performance Evaluation

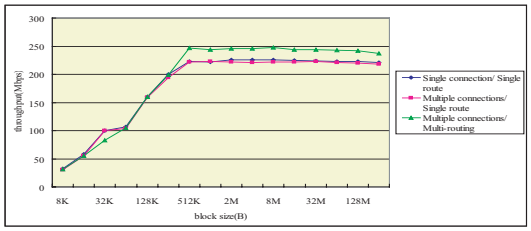

Figure 10. Comparison of throughput with various block sizes

Figure 10 shows the throughput graph with various block sizes in the case of iSCSI Single connection/VPN Single route, iSCSI Multiple connections/VPN Single route, and iSCSI Multiple connections/VPN Multi-routing. In this experiment, one-way delay time is 0ms.

As block size increases the throughput of iSCSI access increases. However, it saturates at the block size of more than about 512KB in all case. Additionally the throughput at Single connection/Single route is little less than that of Multiple connections/Single route, while in the case of Multiple connections/Multi-routing, the throughput is relatively high.

We examine the result of high throughput at Multiple connections/Multi-routing. Although Initiator and Target are connected with Gigabit Ethernet, the encryption speed of a VPN router is 500Mbps maximum. Moreover as the throughput of the sockets communication at VPN Single route in this experiment is actually measured, the performance is about 330Mbps. Therefore, the remote access bottleneck is the encryption processing of VPN router. As the route is divided into multiple ways, the load of the encryption processing of VPN router is distributed and become lightweight. The reason why the performance is not close to twice as high as the single route case is that the bandwidth of the connection does not completely double in spite of using 2 circuits. In the case of Multiple connections/Multi-routing, 2 connections are used by turns, not simultaneously, since commands are transferred on a connection decided by round robin on iSCSI.

5.2 Comparing of CWND Behavior

Figures 11, 12, and 13 show the behavior of the TCP CWND. In this experiment, block size is 2MB and one-way delay time is 2msec.

Figure 11 shows CWND in the case of Single connection/Single route when iSCSI sequential read access is performed. In this figure, ErrorNo.2 means Local device congestion (CWR), ErrorNo.3 means receiving duplicate ACK/SACK, and ErrorNo.4 means timeout. The vertical dashed line of ErrorNo.2 indicates happening of CWR at the point. This error means device driver buffer of the data sender overflows. The CWND increases up to about 350 packets, then CWR is detected and it decreases rapidly.

Figure 12 shows CWND in the case of Multiple connections/Single route. The vertical dashed line of ErrorNo.2 indicates also happening of CWR. We confirm that the

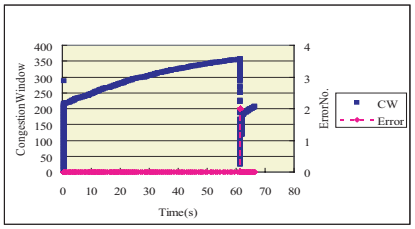

Figure 11. CWND (Single connection/Single route)

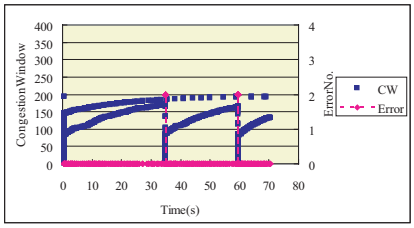

Figure 12. CWND (Multiple connections/Single route)

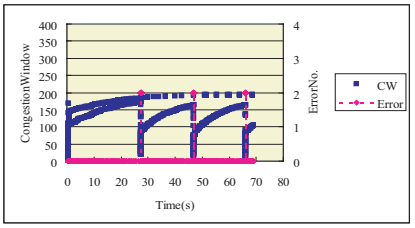

Figure 13. CWND (Multiple connections/Multi-routing)

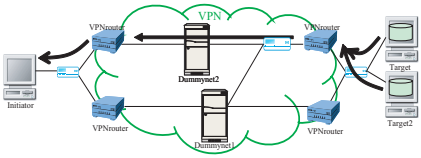

Figure 14. Multiple Targets/ VPN Single route

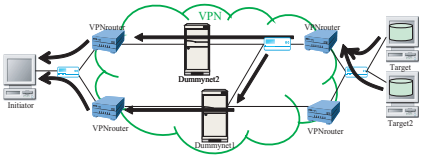

Figure 15. Multiple Targets/ VPN Multi-routeing

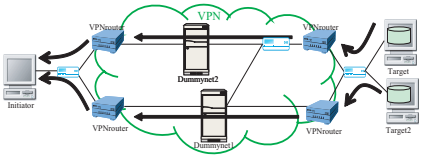

Figure 16. Multiple Targets/ Independent Multiple routes

graph changes greatly in comparison with the graph of Single connection/Single route. CWND monitor tool indicates the CWND value when the CWND value changes at any connection. For this reason each connection changes the value of CWND independently, and the timing indicating the value of graph is at random. Therefore the shape of CWND becomes a constant graph and a saw-like graph.

Figure 13 shows the case of CWND at Multiple connections/Multi-routing. The behavior of CWND is almost the same with the case of Multiple connections/Single route.

Thus we have discussed the reason why the shape of CWND became a constant graph and a saw-like graph in the case of VPN Multi-routing. If one of connections have consumed CWND, the next iSCSI access is not performed until its ACK is received. Because commands are assigned to iSCSI connections by round robin, when an access through one connection is suspended, the access through another connection also stops. Therefore, since the packet which consumes CWND is not sent through another connection, CWND of another connection becomes constant according to the feature of Linux TCP CWND.

Next, we compare the difference of the CWND of Single connection and Multiple connections. Compared with these cases, errors are frequent in the case of Multiple connections than that of Single connection. Moreover, the sum of the CWND value at the constant graph and the saw-like graph become about 350 packets, which is the maximum value of CWND at Single connection. Because of the graph becomes a constant graph and a saw-like graph in the case of Multiple connections, packets are transferred until by both of CWND values. Therefore the buffers of the Target device driver is considered to overflow frequently.

6 Simultaneous Access to Multiple Targets

In this section, we assume Multiple Targets network storage system, and we configured such a system using 2 Targets. The Targets consist of 2 storage devices in this experimental system. If parallel storage system like the autonomous disk is operating, the 2 Targets look as a single network storage system having a single IP address for an interface accessed from the server side. Since the current experimental system does not equip such a mechanism, 2 Targets look as 2 storage devices from the server side. However in this experiment, we assume that parallel storage system is operating in order to evaluate the performance of the proposed system.

Figure 14 shows the method to access Multiple Targets in the case of VPN single route. On the other hand, the route can branch to 2 routes at VPN router, as shown in figure 15. This is equivalent to the case of accessing parallel storage system having single IP address by VPN Multi-routing. Moreover, in order to confirm the ideal value of Multiple access performance, we have also measured the performance when accessing both Targets simultaneously through independent 2 routes, as shown in figure 16.

Figure 17 shows the throughput using Multiple Targets with various block size, in the case of VPN Single route, VPN Multi-routing, and Independent Multiple routes. In this experiment, one-way delay time is 0ms. T1+T2 in this graph is the sum of each throughput at Target1 and Target2 with VPN single route. It shows the maximum value using 2 Targets.

The reason why the throughput at Multiple Targets/Single route is better than that of Target1 or Target2 is simply that the Target became double. Moreover, the reason of higher throughput at Multiple Targets/Multi-routing than

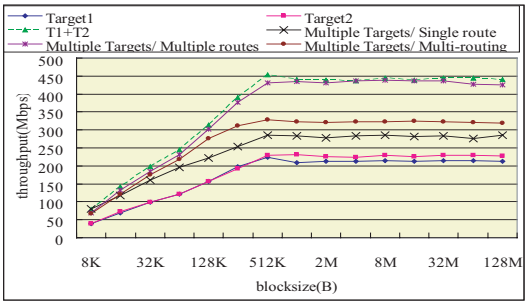

Figure 17. Comparison of throughput with Multiple Targets

Multiple Targets/Single route is that the connection on VPN has twice bandwidth and the processing data at VPN router in Initiator side is dispersed. Furthermore, the reason of higher throughput at Multiple Targets/Multiple routes than that of Multiple Targets/Multi-routing is that the processing of packets dividing on VPN router in Target side become unnecessary and the processing of encryption was dispersed. Finally, the small difference between Multiple Targets/Multiple routes and T1+T2 seems to be a processing overhead of Initiator.

In this experiment, 2 Targets have independent disks and IP addresses. However in practical use, it is desirable to access it without being conscious of a destination disk from Initiator side. Thus it is desirable that it is configured as a single huge storage system having single IP address. The performance of VPN Multi-routing in this study is applicable for such a case. On the other hand, the bottleneck of this case is the processing of VPN router. Thus this access method becomes effective as performance of VPN router improves. The ideal value of throughput should accord with the performance of Multiple routes case in this study. Furthermore, we are currently configuring the experimental environment using autonomous disk as parallel storage system.

7 Conclusions

In this paper, we have observed behavior of the TCP CWND, and evaluated the performance on iSCSI storage access using VPN with various access methods.

As we have compared iSCSI Multiple connections/VPN Single route and iSCSI Multiple connections/VPN Multi-routing, the behavior of the CWND is almost the same, while the throughput is relatively high in the case of iSCSI Multiple connections/VPN Multi-routing. This is because the encryption processing on VPN router is reduced for multi-routing.

In addition, we have examined an iSCSI access method on VPN when a parallel storage system is used. We have evaluated the performance using 2 Targets with various access methods.

As a result, we have confirmed that in the case of VPN Multi-routing, the throughput has improved compared with the Single route case and efficient communication is performed. Moreover, we have obtained the ideal value of throughput with 2 Targets parallel storage system.

As a part of future works, we will implement and evaluate a parallel storage system, and investigate an access control method. Moreover we will consider designing and implementing an algorithms other than round robin, and try using asynchronous read operations that can operate in parallel.

References

[1] iSCSI Specification ! $ http://www.ietf.org/rfc/rfc3720.txt?number=3270

[2] SCSI Specification ! $ http://www.danbbs.dk/˜dino/SCSI/

[3] S. Yamaguchi, M. Oguchi, and M. Kitsuregawa: ”iSCSI Analysis System and Performance Improvement of Sequential Access in Long-Latency Environment,” IEICE Transaction on Information and Systems, Vol.J87-D-I, No.2, pp.216-231, February 2004.

[4] InterOperability Lab, Univ,of New Hampshire, http://www.iol.unh.edu/consortiums/iscsi/

[5] N.Chishima, M. Toyoda, S. Yamaguchi, and M. Oguchi: ”Evaluation of Correlation between TCP Parameters and Communication Performance on VPN,” FIT2006, L-042, September 2006.

[6] Fujitsu IP access router GeoStream Si-R570 ! $ http://fenics.fujitsu.com/products/sir/sir570/index.html

[7] Haruo Yokota:”Autonomous Disks for Advansed Database Applications”, DANTE’99 ! $pp.441-448 ! $ November 1999.

[8] N.Chishima, M. Toyoda, S. Yamaguchi, and M. Oguchi: ”A Study of Controlling TCP Congestion Window on iSCSI Access through VPN,” DBWS2006, pp.709-712, July 2006.

[9] M. Toyoda, S. Yamaguchi, and M. Oguchi: ”Proposal and Performance Evaluation of TCP Congestion Window Control Method on iSCSI Storage Access”, IEICE Transaction on Information and Systems, Vol.J90-D ! $No.2 ! $pp.359-372 ! $February 2007.

[10] K.Fujiwara ! $N.Wakamiya ! $K.Shiga:”A study on protocol tuning for iSCSI transmission over a wide-area IP network”, The 68th National convention of IPSJ ! $ pp.155-156 ! $March 2006.

[11] L.Rizzo:”dummynet”, http://info.iet.unipi.it/˜luigi/ip_dummynet/