AoE storage protocol over MPLS network

Marek Landowski, MEng
Electrical, Electronic & Communications Engineering
University College Dublin, Ireland
marek.landowski@ucd.ie

Dr Paul F. Curran
Electrical, Electronic & Communications Engineering
University College Dublin, Ireland
paul.curran@ucd.ie

***Abstract*—ATA over Ethernet (AoE) protocol is an interesting alternative to iSCSI and Fibre Channel. AoE is a light, layer 2 protocol integrated with Ethernet frames, which makes it ideal for work inside LAN segments. Unfortunately, this advantage is also its limitation when access to the AoE storage is required to be over the internetwork. In this paper we show how MPLS can make AoE routable and thereby also independent of Ethernet itself.**

Keywords: Storage Area Networks, SAN, ATA over Ethernet, AoE, Multi-Protocol Label Switching, MPLS

I. Introduction

Server systems can mount a disc volume over the network. This is usually done when some of the resources are held on remote storage servers. If the storage permits other computers to access files only then it is described as Network Attached Storage (NAS). Popular file sharing protocols include NFS, SMB, FTP and HTTP. The file server maintains the files and directories and shares them with clients. In general this slows down access to files, because file server must take charge on every operation. For example, when a client writes a file to a file server, the file is first written in the virtual shared file space and only then is it physically written onto a real disc drive [1].

A much quicker method is provided by Storage Area Networks (SANs) where the storage server shares a disc volume, as distinct from merely sharing files with the clients. In principle SAN is similar to non-networked interfaces such as SATA or SCSI, but with the main difference being that in SAN physical connection is achieved through a network of switches and routers, whereas in SATA and SCSI it is by wire [2]. A SAN is broadly comprised of four distinct components: SAN storage server, SAN clients, a connecting network and protocols governing the sharing of disc volume. The SAN client has full rights to the mounted over the network disc volume including read/write operations as well as right to format the volume and create a partition on it. A variety of protocols such as Fibre Channel, iSCSI (Internet SCSI) and AoE (ATA over Ethernet) are employed to enable the sharing of disc volumes in this way. Fibre Channel and iSCSI are more expensive alternatives to AoE. Both are based on SCSI rather than ATA [3]. They include a higher overhead in each data packet which inevitably requires more processing. Fibre Channel was originally developed for use with fibre optics, but with extensions it may also be used with Ethernet, either by itself (FCoE) or in conjunction with IP (FCIP or iFCP) [4]. The iSCSI protocol also uses IP and provides SCSI commands over the TCP/IP network [5].

Accordingly both protocols have the distinct advantage of being routable and both protocols will effectively function in conjunction with a variety of technologies such as Ethernet, ATM, *etc*. On the other hand AoE, which is less expensive and has lower overheads, is not routable and functions only with Ethernet technology.

Both of these disadvantages of AoE can be overcome by employing tunnels. Examples of already identified tunneling methods for WAN connectivity to disc volumes are presented in [6]. They employ additional protocols which encapsulate AoE packets inside the tunnel and enable routing over the network. There is of course an increase in overhead, but overheads remain below or comparable to those of Fibre Channel and iSCSI. The financial cost is in no way affected so that AoE retains its advantage in this respect.

However, it transpires that this problem of ensuring that AoE can be routed is solvable without the construction of tunnels. Moreover, the increase in overhead in comparison with AoE is relatively low and is lower than that required by existing tunneling methods. In this paper we present a routable form of AoE which does not employ tunneling but rather uses MPLS technology only.

In section II we discuss AoE, MPLS and a method of routing AoE over the MPLS network. In section III we discuss in some detail the experimental setup. Results are presented and discussed in section IV and reveal the efficacy and relative efficiency of the method. Conclusions are presented in section V.

II. AoE and MPLS

AoE is a simple protocol for sharing disc volumes with clients over the Ethernet. The protocol is described in the AoE specification [7] which defines the header format and a set of four commands used to achieve a very basic RPC mechanism between a client and a storage. These commands are: (i) Issue ATA Command, (ii) Query/Config Information, (iii) MAC Mask List and (iv) Reserve/Release. AoE messages share a common 20 byte header format, with an additional

978-1-4577-0428-4/11/$26.00 © 2011 IEEE

header depending upon the specific command. The Issue ATA Command, which is responsible for read/write operations, the Config/Query Information command, which retrieves or sets configuration information and the MAC Mask List command, which controls access to the storage, all require a further 12 bytes in addition to the common header, giving 32 bytes in total for these commands. The Reserve/Release command, which controls locking of storage, requires at least 2 further bytes in addition to the common header. The overall header size depends upon the number of clients permitted to perform ATA commands on the disc.

In principle AoE employs Ethernet only, because it was originally designed for local networks. The Ethernet has the virtue of being simple and easy to maintain, is reliable, has the ability to connect new technologies together, and has low cost of installation and upgrade. AoE exploits all these advantages and employs Ethernet broadcasts for storage discovery. The broadcasts are naturally terminated on the router, because routers do not forward them. This feature restricts the range of AoE to the local Ethernet segment only. In cluster systems this feature is required, because it ensures that the storage cannot be externally accessed. However, this same feature gives rise to significant difficulties if external access of the AoE storage is in fact required.

When external access is required and AoE is routed along a tunnel the overall header size with a command increases and depends upon the tunnelling method employed. In the Table I we show selected protocol stacks available for governing remote access to the disc volumes. We assumed that protocols are using its mandatory header fields (with some exceptions). In life networks, this totals can be different. It is not a surprise, because some of the protocols have optional fields, which when used, increase the overall number of bytes. For better comparison to AoE we assume that other routable storage protocols employ Ethernet as transmission medium. When AoE is encapsulated in the GRE tunnel the overall header size increases from 32 bytes to 80 (32 bytes AoE header, 14 bytes GRE header, 20 bytes IP header and 14 bytes Ethernet header). GRE has a feature of encapsulating anything within IP packet [8], what allows to reach tunnel endpoint over the Internet. Unlike the MPLS, when tunnel is down, GRE has no means to re-route the traffic to a new tunnel. Instead it does not forward or process any traffic, apart from sending and listening for keepalive packets [9]. When AoE is encapsulated in the L2TP protocol the overall header size is equal to 86 bytes (32 bytes AoE header, 12 bytes L2TP header, 8 bytes UDP header, 20 bytes IP header and 14 bytes Ethernet header). L2TP protocol can be encapsulated within UDP protocol or can run directly on the top of technologies such as ATM, Frame-Relay, MPLS, *etc*. Like the MPLS it is a layer 2 protocol, and unlike, it has no means to re-route the traffic to a new tunnel when the tunnel is down. Instead it considers the connection to the peer as to be lost, and sends appropriate messages while tunnel is set into the idle state [10]. In comparison when disc access is governed by iSCSI the overall header is at least 102 bytes (48 bytes basic iSCSI header [5], 20 bytes TCP header, 20 bytes IP header and 14 bytes Ethernet header). The routable forms of Fibre Channel, namely FCIP or iFCP, have headers comprising at least 82 bytes (28 bytes Fibre Channel header, 20 bytes TCP header, 20 bytes IP header and 14 bytes Ethernet header) [11].

TABLE I
PROTOCOL HEADER SIZES FOR TRANSMISSION OVER ETHERNET

Protocol	Total header size [bytes]
AoE / Eth	32
AoE / MPLS / Eth	50
AoE / GRE / IP / Eth	80
AoE / GRE / IP / MPLS / Eth	98
AoE / L2TP / UDP / IP / Eth	86
iSCSI/Eth	102
iSCSI / MPLS / Eth	120
iFCP / Eth or FCIP / Eth	82

In sense of ISO/OSI model, AoE is a native layer 2 protocol and usually tunnelling methods involve upper layers for enabling routing over the networks, which if unnecessary must be considered undesirable. When GRE is employed, the layer processing of AoE packets starts at layer 2 → 3 → 3 → 2. When L2TP is employed the processing starts at layer 2 → 2 → 4 → 3 → 2 or it stays at layer 2. When iSCSI governs the disc access the processing starts at layer 4 → 3 → 2. When iFCP and FCIP are employed the processing through the layers is more complex (Fibre Channel layers → 4 → 3 → 2), because Fibre Channel itself has custom layers which cannot be expressed through ISO/OSI model.

To extend client access to AoE disc volumes over the internetwork it is not necessary for routers to construct tunnels such as GRE/IP or L2TP. AoE packets can be sent over the MPLS network, which is very commonly employed by the service providers nowadays. MPLS is a technology which was developed to give service providers better control over the traffic routed through their networks. This is achieved by means of a labeling mechanism which effectively selects the path. MPLS has the ability of routing traffic over the network, because it integrates network layer routing (BGP, OSPF, IS-IS) with label switching [12], [13]. When AoE is encapsulated within MPLS (Table I) the overall header size increases from 32 bytes to 50 (32 bytes AoE header, 4 bytes MPLS header and 14 bytes Ethernet header). In sense of ISO/OSI layer processing, MPLS has significant advantage to existing tunnelling methods, because it works at layer 2 only. MPLS potentially solves a fundamental problems occurring in IP networking, namely layer 3 lookups and routing data across, not necessarily the shortest, but rather the least congested paths. MPLS also has the unique ability to re-route traffic very quickly in the case of network failure, and with additional labeling, it can enable full separation from the other traffic which is routed along the same path. Moreover, we establish here that it also overcomes the problem of routing AoE over networks, and frees it from Ethernet, because MPLS runs over whichever mix of networking technologies it faces, including ATM, SDH, Metro Ethernet *etc*.

III. Experimental Setup

For the purpose of establishing a baseline of performance and to investigate hardware dependence we consider two groups of experiments. Group I represents series of live experiments when storage was directly connected to the server (Fig. 1) whereas group II represents series of live experiments when storage and server were connected over the MPLS network (Fig. 2). To artificially slow down the MPLS network we have used Pentium 3 machines with FastEthernet network cards. The MPLS network itself is not complex. We consider simple topology (three edge routers and two switching routers) to research how the performance changes when AoE is routed along MPLS path and how such topology impacts the performance when compared with results from the group I.

In each group we test two AoE (AoE-Lx and AoE-Cd) and one iSCSI (iSCSI-Lx) storage servers. Physically, the AoE-Lx and iSCSI-Lx are regular Intel Xeon, 4GB memory, 2TB HDD rack servers. In opposition to AoE-Lx, the AoE-Cd is the Coraid EtherDrive SR421 series storage with available disk volume of 4TB size. We employ two types of AoE storage servers to research how the performance changes when AoE is running on a different hardware platform. While on the contrary, AoE-Lx and iSCSI-Lx are employed to research how AoE and iSCSI perform when running on the same class of hardware. In all series of experiments the SERV1 or SERV2 acting as storage initiators remain the same (Dell 775). The MPLS routers are build on the basis of five Dell 250 servers.

For each group we investigate three performance measures such as timings of server cache reads, timings of disc volume reads and copy transfer speeds between server and storage (Fig. 3). The timing of cache reads (in [GB/s]) measures the speed with which the server reads through the buffer cache without disc access. This measurement is essentially an indication of the throughputs of the processor, cache, and memory of the server under the test. The timing of disc volume reads (in [MB/s]) measures the speed with which the server reads through the buffer cache to the disc without any prior caching data. This measurement is an indication of how fast the drive can sustain sequential data reads under Linux, without any filesystem overhead. Sequential data reads describe access part of I/O process of reading the contiguous clusters of data on the disc volume [14]. The copy transfer speed (in [MB/s]) measures the speed with which the initiator uploads 4GB size file to the targeted storage.

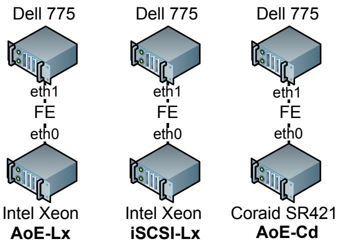


Fig. 1. Group I: servers and storage directly connected

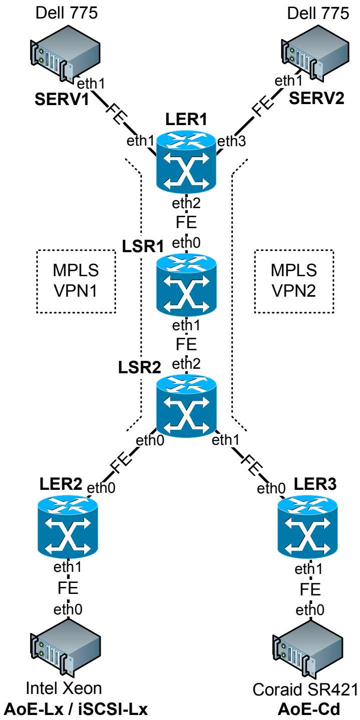
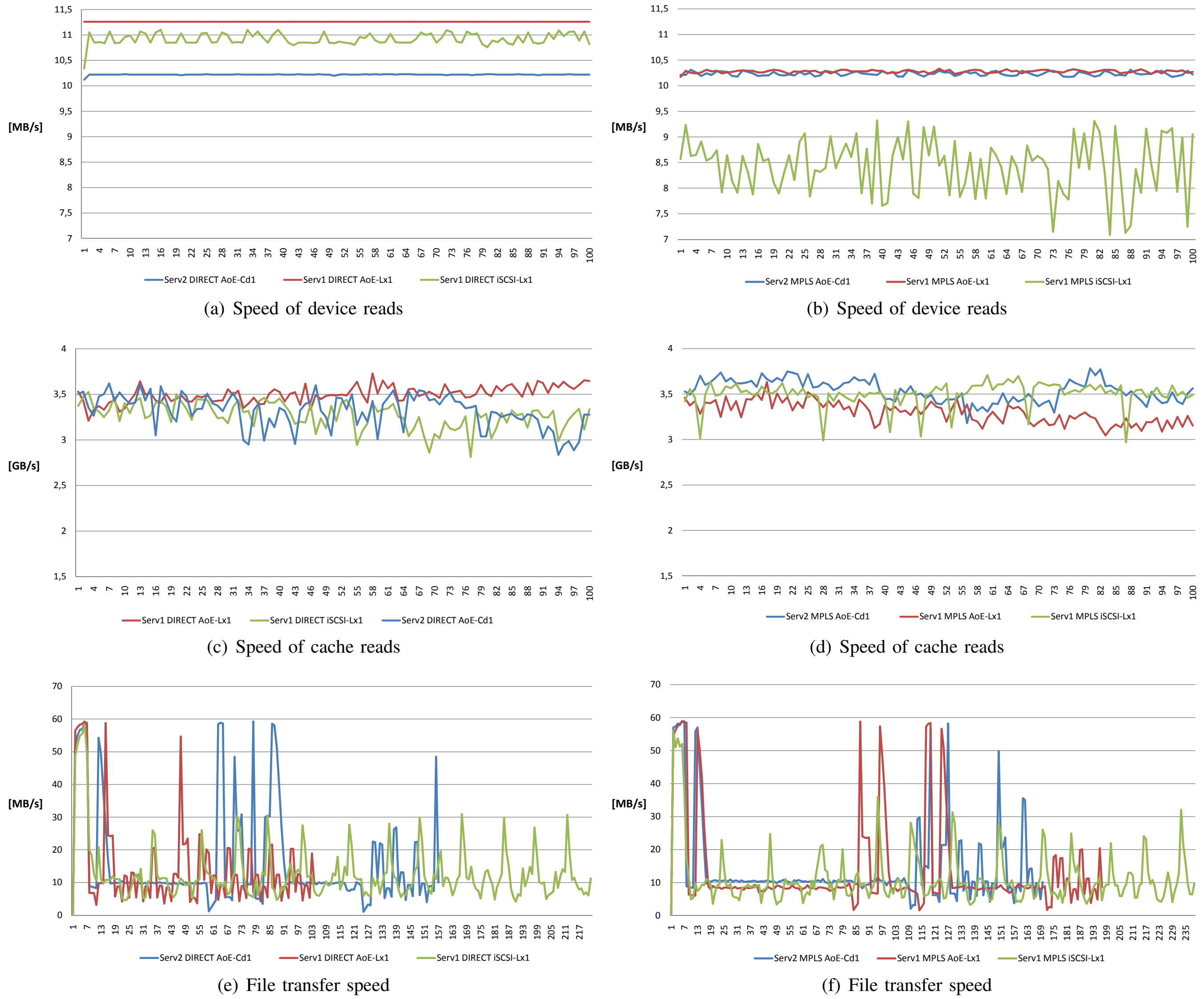


Fig. 2. Group II: servers and storage connected over the MPLS network

The experimental MPLS network consist of three Label Edge Routers (LERs) and two Label Switch Routers (LSRs). A Label Edge Router has the ability to add a label to an unlabeled AoE packet, or to remove a label from a labeled AoE packet. Internally labeled AoE packets are switched by Label Switch Routers. Switching information is retained in data structures such as Forwarding Equivalence Class (FEC), Incoming Label Map (ILM) and Next Hop Label Forwarding Entry (NHLFE). FEC is responsible for classification of incoming AoE packets to appropriate labels so that they can be appropriately switched inside the domain. The ILM structure holds information concerning the incoming labels and interfaces. NHLFE retains information concerning outgoing labels, output interfaces and next hop nodes. SERV1 and AoE-Lx/iSCSI-Lx are separated from SERV2 and AoE-Cd by the addition of a second label (called bottom) to the AoE packets. When SERV1 communicates with AoE-Lx/iSCSI-Lx, either LER1 or LER2 attaches a unique bottom label to the AoE packets and switches them along the path 1. Likewise, when SERV2 communicates with AoE-Cd then either LER1 or LER3 attaches a unique bottom label to the AoE packets and switches them along the path 2. These labels are used only for traffic separation. The actual routing of AoE packets is facilitated by a top label used for switching within the domain (between LSRs as well as between LERs and LSRs). In the experiment, the labels are added without analysing the layer 2 information (in life networks it is possible to use 802.1Q header information to label the packet appropriately). Moreover LSRs switch the labels without analysing the bottom



(a) Speed of device reads

(b) Speed of device reads

(c) Speed of cache reads

(d) Speed of cache reads

(e) File transfer speed

(f) File transfer speed

Fig. 3. Speeds when servers and storage were connected directly (a, c, e) and when connected over MPLS (b, d, f)

label. Hence intermediate nodes (such as LSR1 and LSR2) have no information concerning traffic separation into the VPNs. When LER receives a labeled AoE packet, it pops the top label and consults the remaining label. The bottom label tells LER whither the AoE packet is destined i.e. to which interface the packet must be forwarded.

MPLS VPN also has a security advantage. SERV1 cannot attach the AoE-Cd which is available in other VPN, unless there is a configuration error in the MPLS network. However, MPLS open access to the AoE storage, and expose it to the possibility of being attacked and corrupted from outside. To secure this access, AoE packets should be sent inside private VLANs in the Ethernet segment and mapped to the dedicated MPLS VPN path when sent over the service provider network.

IV. Results

The results for group I of experiments are shown in Table II. These make clear that the performance of AoE is dependant upon the hardware employed and is superior to that of iSCSI. The results for group II are presented in Table III. Clearly the introduction of the MPLS network decreases the transfer speeds in each case. Nevertheless the decrease is relatively modest ranging from 13% in the case of iSCSI to approximately 9% in the case of the Coraid AoE storage. The performance of the AoE over MPLS is really quite close to that of AoE direct, bearing in mind that the former is routed whereas the latter is not.

Of course AoE over MPLS and iSCSI over MPLS are slower than directly connected, how could they not be, but their advantage in terms of routability is really rather significant. It should be noted that the first two experiments of Table III establish unequivocally that the method is effective,

TABLE II
AVERAGE SPEEDS WHEN DISCS WERE DIRECTLY ATTACHED TO SERVERS

Initiator	Target	CONN	AVG speed of cache reads [GB/s]	AVG speed of device reads [MB/s]	Transfer speed [MB/s]
SERV2	AoE-Cd	DIRECT	3.22	10.22	15.78
SERV1	AoE-Lx	DIRECT	3.31	11.26	14.06
SERV1	iSCSI-Lx	DIRECT	3.04	10.88	12.66

TABLE III
AVERAGE SPEEDS WHEN DISCS WERE ATTACHED TO THE SERVERS OVER THE MPLS NETWORK

Initiator	Target	CONN	AVG speed of cache reads [GB/s]	AVG speed of device reads [MB/s]	Transfer speed [MB/s]
SERV2	AoE-Cd	MPLS	3.52	10.23	14.29
SERV1	AoE-Lx	MPLS	3.29	10.27	12.34
SERV1	iSCSI-Lx	MPLS	3.34	8.44	11.01

AoE over MPLS does indeed permit external access to the disc volume. We can say this with certainty since the experiments consist precisely of achieving such routing. This ability to route AoE does not appear to be hardware dependent, functioning quite effectively with both hardware platforms on which it was tested (AoE-Lx and AoE-Cd). The second and third experiments of Table III are of particular significance. Evidently, restricting to the same hardware platform, AoE over MPLS actually delivers improved performance in comparison with iSCSI over MPLS and of course it does so at a much reduced financial cost.

In addition to its advantage over iSCSI in transfer speed, AoE over MPLS has some unexpected further advantages. When the servers and storage were directly connected, the average speeds of device reads in case of AoE are constant in comparison to iSCSI (Fig. 3a). This shows that AoE is much better at sustaining sequential data reads and addition of MPLS network does not change that fact (Fig. 3b) The timings of cache reads in all cases present similar variability (Fig. 3c). There is no major difference when MPLS is introduced (Fig. 3d), because cache reads test throughput of the processor, cache, and memory of the server and in all cases the servers where the same. The transfer speed charts (Fig. 3e and Fig. 3f) not only show that in general AoE achieves faster transport than iSCSI, but also show that AoE flows are bursty and have shorter duration then those of iSCSI. It means for the network that when AoE flows are arriving, the router must have enough space in the queue to buffer incoming bursts. The iSCSI flows are less bursty, but lasting longer have lower transfer speeds. One of the reasons why AoE storage has better performance than iSCSI lay in layer processing discussed in Sec. II. The difference is not overwhelming, but shows supreme position of AoE in SAN environment.

V. CONCLUSIONS

It has been argued in this paper that AoE, with its significant advantage of reduced cost in comparison with routable protocols such as iSCSI and Fibre channel, can, in fact be routed. On its own this claim is hardly novel. Many authors have considered the routing of AoE by means of tunnels. However, tunnels involve, at a minimum, a significant increase in the size of the headers required. On the contrary we find that AoE over MPLS provides a routable protocol which can be implemented without the need for tunnels and with a very modest increase in the header size in comparison with AoE. As essentially a side benefit the resulting protocol is no longer restricted to Ethernet, working on the MPLS network which is, of course pervasive. Although the performance of this routable form of AoE is degraded in comparison with its non-routable counterpart, experiment shows that this degradation is surprisingly small, just 12% or so, given that the gain, namely routability, is so large. The suggested protocol outperforms other tunneling methods for routable AoE, such as GRE/IP and L2TP. More significantly the new method also outperforms iSCSI, a protocol which comes at a much greater financial cost.

ACKNOWLEDGMENT

The would like to thank Science Foundation Ireland for the full support of our research. We owe also thanks to Brian Turner, Senior IT engineer at University College Dublin for feedback in the field of cluster systems.

REFERENCES

[1] CORAID, Inc., “Fundamentals of Networked Storage,” June 2008.
[2] E. L. Cashin, “Kernel Korner - ATA Over Ethernet: Putting Hard Drives on the LAN,” *Linux Journal*, no. 134, June 2005.
[3] M. A. Covington, “An Overwiew of CORAID Technology and ATA-over-Ethernet (AoE),” 2008.
[4] INCITS T11 Technical Committee, “Fibre Channel Backbone - 5 (FC-BB-5), REV 2.00,” June 2009.
[5] J. Satran, et al, “RFC3720: Internet Small Computer Systems Interface (iSCSI),” April 2004.
[6] David W. Chapman Jr., “Application Note: AoE WAN Connectivity,” August 2008.
[7] S. Hopkins and B. Coile, “AoE (ATA over Ethernet),” February 2009.
[8] B. Hubert, “Linux Advanced Routing & Traffic Control.”
[9] Cisco Systems, “Document ID: 63760 How GRE Keepalives Work,” June 2006.
[10] R. Shea, *L2TP: Implementation and Operation*. Addison-Wesley Professional, 1999.
[11] J. Long, *Storage Networking Protocol Fundamentals*. Cisco Press, 2005.
[12] Bruce Gillham, Brian Wesley Simmons, Fran Singer, “JUNOSe Internet software for e-series routing platforms: BGP and MPLS configuration guide,” April 2006.
[13] V. Alwayn, *Advanced MPLS Design and Implementation*. Cisco Press, 2002.
[14] Microsoft TechNet Library, “The Distributed Systems Guide,” 2011.