Computer Communications 32 (2009) 750–765

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Spanning tree elevation protocol: Enhancing metro Ethernet performance and QoS

Minh Huynh [a], Prasant Mohapatra [a,*], Stuart Goose [b]

[a] Computer Science Department, University of California at Davis, Davis, CA 95616, United States
[b] Siemens Technology-To-Business Center, 1995 University Avenue, Suite 375, Berkeley, CA 94704, United States

ARTICLE INFO

Article history:
Received 7 March 2008
Received in revised form 2 December 2008
Accepted 3 December 2008
Available online 14 December 2008

Keywords:
Spanning Trees
Load balancing
Metro Ethernet Network
Quality of Service (QoS)
RSTP

ABSTRACT

The economics and familiarity of Ethernet technology is motivating the vision of wide-scale adoption of Metro Ethernet Networks (MEN). Despite the progress made by the community on additional Ethernet standardization and commercialization of the first generation of MEN, the fundamental technology does not meet the expectations that carriers have traditionally held in terms of network resiliency, load management, and Quality of Service (QoS). We propose a new concept of Spanning Tree Elevation Protocol (STEP) that increases MEN performance while supporting QoS including traffic policing and service differentiation. STEP manages multiple Spanning Trees as a means to control the traffic flow rates and to differentiate classes of traffic. Whenever a service level agreement is compromised, STEP redirects frames of affected flows to the next spanning tree in sequence utilizing the alternate paths. As a result, the capacity in terms of network throughput is greatly enhanced while almost avoiding any reconvergence time in the case of failures. The gain ranges from 1.7% to 7.3% of the total traffic in the face of failure; while load balancing increases an additional 12.8% to 37% of the total throughput. At the same time, STEP maintains the required bandwidth for high priority traffic class during the failure scenarios and the high congestion scenarios.

1. Introduction

The most common technology used for local area networks is the Ethernet protocol, which has been predominant for more than 30 years. Ethernet is a simple and cost-effective protocol that provides a variety of services. Despite the occasional challengers, such as fibre channel [22] and InfiniBand Architecture [23], the evolution of Ethernet has continued. The recent standardization of Gigabit Ethernet [13] protocol has propelled it for consideration in the scope of metropolitan area networks.

Metro Ethernet Networks (MENs) [12] comprise a metro core network and several access networks as shown in Fig. 1. All the access networks connect to the core at one or two gateway Ethernet switches. The subscribers' networks are connected to the access network, and the metro core helps in interconnecting the access networks. Packets hop through multiple switches in both access and metro core networks. Redundant links are used in the core as well as the access networks. The main challenges in the context of MEN include resiliency, load balancing, and support for QoS.

Current Ethernet solutions deploy Local Area Network technology for Metro Area Networks (MANs). Deploying the Spanning Tree Protocol (STP) to manage the topology autonomously are inadequate and do not meet the requirement for MANs because STP blocks redundant links leaving traffic on a single path, running the risk of slow reconvergence and link congestion. As a result, STP provides poor support for resiliency and load balancing.

The need for QoS in MEN is driven by the application requirements. Applications such as video conferencing, VoIP, and online gaming require an upper bound on the delay and jitter. Other applications, such as streaming video on demand, call for guaranteed bandwidth. Enterprises desire guaranteed services interconnecting their remote sites for their intranet services. While bandwidth could be over-provisioned on gigabits pipes, there is still a need for differentiation of services. Subcribers have different traffic profiles and needs, which reflect the range service provider pricing plans. Naturally, a demanding subscriber should not pay the same rate as an occasional web surfer.

In this work, we address poor support in Ethernet for resiliency, load balancing, and QoS, specifically in term of service differentiation. We have introduced a new approach, called Spanning Tree Elevation Protocol (STEP), which allows switching between multiple Spanning Trees (STs) without forming any cycles. This feature enhances the resiliency as well as facilitates load balancing. In addition to fast recovery, it also increases the capacity of the network in terms of the achievable throughput.

In STEP, traffic flows are managed by different STs associated with the VLANs. Based on the traffic class to which a frame belongs, it is assigned to a ST at the ingress. Before sending out to a link, a switch checks for the service level agreement. If the frame is in-profile, then it is allowed to continue, otherwise, the frame is ele-

[*] Corresponding author. Tel.: +1 530 754 8380; fax: +1 530 752 4767.
E-mail addresses: huynh@cs.ucdavis.edu (M. Huynh), prasant@cs.ucdavis.edu (P. Mohapatra), stuart.goose@siemens.com (S. Goose).

0140-3664/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2008.12.001

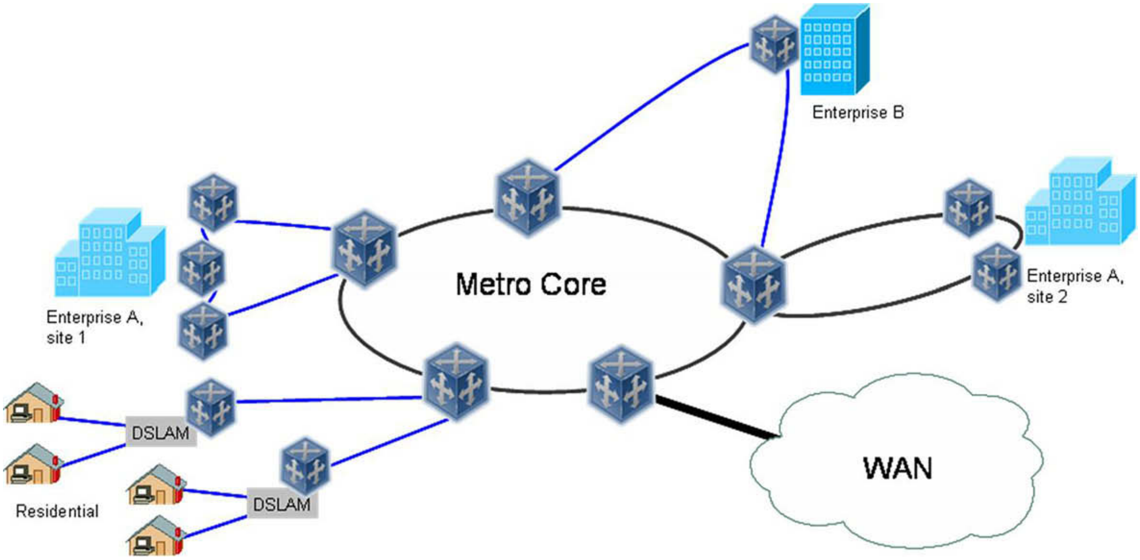

Fig. 1. A typical MEN architecture [12].

vated to the next ST that belongs to the same service class of the frame. Each crossover indicates a markdown for that frame until there are no more STs to cross. Then the frame is classified as out-of-profile and is dropped.

We believe that STEP has the potential to provide enhanced services with a low overhead. A key requirement for STEP is to be backward compatible with the current Ethernet protocols. STEP fully supports switches running the Spanning Tree Protocol family such as STP, RSTP, and MSTP. Furthermore, considering the ASIC paradigm in network processing units (NPUs), STEP was designed to have a simple implementation for efficient adaptation that requires no changes to NPUs.

The encouraging experimental results presented in this paper were obtained using the OPNET [11] simulation product to quantify the resiliency and the gain in terms of the network throughput. The behaviors of the Ethernet switches within OPNET Modeler were modified to imbue the STEP approach. In the resilience test scenarios, STEP yields an increase of 1.69% and 7.3% of the total traffic comparing to Multiple Spanning Tree Protocol (MSTP) and Rapid Spanning Tree Protocol (RSTP), respectively. In addition, when the network is overloaded and imbalanced, STEP gains an additional 12.76% and 37% of the total traffic comparing to MSTP and RSTP, respectively. For QoS tests, STEP maintains the required bandwidth uninterrupted for the higher priority class obviating the reconvergence process.

The organization of the paper is as follows: a preliminary section explains the current state of Ethernet and the motivation for STEP. It is followed by a description of the concept of STEP. STEP is then evaluated separately in three areas: resilience, load balancing, and QoS. Finally, related works are presented before the conclusion of the paper.

2. Preliminaries

Traditionally, Ethernet-based networks use STP [1], standardized in IEEE 802.1d, for switching frames in the network. STP is a layer 2 protocol that can be implemented in switches and bridges. Essentially, it uses a shortest-path approach in forming a tree that is overlaid on top of the mesh-oriented Ethernet networks. Spanning tree is used primarily to avoid the formation of cycles, or loops, in the network. Unlike IP packets, Ethernet data frames do not have a time-to-live (TTL) field. STP prevents loops in the network by blocking redundant links. Therefore, the load is concentrated on a single link which leaves it at risk of failures and with no load balancing mechanism. The root of the tree is chosen based on the bridge priority, and the path cost to the root is propagated throughout so that each switch can determine the state of its ports. Only the ports that are in the forwarding state can forward incoming frames. This ensures a shortest single path to the root. Whenever there is a change in the topology, switches rerun the protocol that can take 30–60 s. At any one time, only one ST dictates the network.

Although STP has been used for most Ethernet networks, it has several serious shortcomings for MEN deployments. These shortcomings are as follows:

1. *Low utilization*: Spanning trees restrict the number of ports being used. In high-capacity Ethernets, this restriction translates to a very low utilization of the network.
2. *Poor resiliency*: A very high convergence time (30–60 s) after a link failure.
3. *No load balancing*: STP does not have any mechanisms to balance load across the network. Such a mechanism is very useful for service providers.
4. *No QoS*: STP lacks the following features to support QoS:
   - Guarantee of services.
   - Admission Control.
   - Traffic Policing and Shaping.

An improvement of STP is the RSTP [2] specified in IEEE 802.1w. RSTP reduces the number of port states from five in STP to three: discarding, learning, and forwarding. Through faster aging time and rapid transition to forwarding state, RSTP is able to reduce the convergence time to between 1 and 3 s. It is understood that depending on the network topology, this value varies. In addition, the topology change notification is propagated throughout the network simultaneously, unlike STP, in which a switch first notifies the root; then the root broadcast the changes. Similar to STP, there is only one ST over the entire network. RSTP still blocks redundant links to ensure loop free paths leaving the network underutilized, vulnerable to failures, and with no load balancing.

MSTP or Multiple Spanning Tree Protocol [4] is defined in IEEE 802.1 s. MSTP uses a common ST that connects all of the regions in the topology. The regions in MSTP are instances of the RSTP. An instance of RSTP governs a region, where each region has its own regional root. The regional roots are in turn connected to the common root that belongs to the common ST. Since MSTP runs pure RSTP as the underlying protocol, it inherits some drawbacks of RSTP as well. However, a failure in MSTP can be isolated to a separate region leaving the traffic flows in other regions untouched. In addition, the administrators can perform light load balancing manually by assigning certain traffic sources to a specific ST.

2.1. Lack of traffic enforcer

As mentioned earlier, the IEEE 802.1 family does not define any mechanism to enforce the traffic to stay within their service level agreement. One method of traffic policing is to drop customers' packets when they exceed the service level agreement. It can be softened through the marking of packets that reach a certain threshold, such as in a packet coloring algorithm. Thus the marked packets are more likely to be dropped than the unmarked ones during the congestion period. The Metro Ethernet Forum is defining a packet marking approach that is similar to the packet coloring concept [17]. Traffic policing is managed through the token bucket mechanism.

2.2. Differentiation of service

The 802.1Q [3] standard defines eight traffic types, in order of priority: background, spare, best effort, excellent effort, controlled load, video, voice, and network control. Network control has the "no loss" requirement to maintain and support the network infrastructure. Voice must be less than 10 ms delay and video must be less than 100 ms delay. Controlled load is important business application traffic that is subjected to some form of "admission control". These traffic types are put into their respective priority queues within a switch. Depending on the number of queues supported on a switch, 802.1Q divides these traffic types among the priority queues. For example, if there are three queues, then the traffic types are grouped as follow: {best effort, excellent effort, background, spare}, {controlled load, video}, and {voice, network control}. Each traffic type is identified by a 3 bit user priority field (CoS bits) within the VLAN tag of the Ethernet header. There are a maximum of eight priorities can be supported on a switch. However, supporting eight priority queues per port on a switch can be very expensive. Even high performance switches like the Catalyst 8500 only support four priority queues per port [18]. Another drawback of the CoS bit is the potential starvation of lower priority traffic when there are a significant portion of higher priority traffic [10].

3. Conceptual approach to STEP

In this section, we describe the basic philosophy behind the STEP protocol and its potential for provisioning enhanced performance, quality, and services.

3.1. STEP philosophy

In the STP and most of its variants, at any point of time, only one ST is used. The use of this ST is facilitated by blocked ports in various combinations in each of the Ethernet switches resulting in low utilization and inefficient usage of bandwidth, especially in gigabit Ethernet. Although many protocols have proposed the enhancement of the basic STP, they still use only a single ST for one flow at any point of time in any segment of the network. These protocols take relatively longer to recover from faults and also have no support for balancing load across the network.

The primary motivation behind the design of STEP is to allow the flexibility of using more than one ST while a flow is *en-route* to its destination. This flexibility allows the usage of more ports per switches. However, to avoid the formation of cycles in the network, certain restrictions are imposed. STEP does not propose a new algorithm to form the ST. Instead, STEP proposes a new robust way to manage multiple STs to include features for resilience, load balancing, and QoS.

The basic methodology for implementing the STEP philosophy is to identify multiple STs and number them sequentially to form an ordered list. The VLAN ids [3] can be used as the sequences for the STs. Frames of a flow start using one ST and if necessary, can be switched over to the next ST (none of the other variants of ST allow this flexibility) in sequence. This procedure can be repeated until the frame reaches the ST which has the highest id in the sequence, as seen in Fig. 2. At no point in time is a frame allowed to switch from a ST with a higher id to a ST with a lower id. A flow is switched, or *elevated*, from one ST to another whenever there is a link failure, or load imbalance. Note that in rare cases, all flows may reach the ST with the highest id in the sequence. This unlikely event happens when there are a large number of failures without any recovery. The handling of such rare events are discussed in Section 3.4. Since a VLAN has a one-to-one mapping to a ST, these terms are used interchangeably.

3.2. Loop free guarantee

As mentioned earlier, switching a frame from one ST to another is allowed *only* from a lower numbered ST to a higher one; the reverse cross-over is not permitted. Therefore, infinite loops cannot occur in STEP because of this monotonic increase, or elevation, property. In other words, STEP does not lead frames into an infinite loop when switching between STs. Initially, all traffic in STEP starts on the first ST. All the time no problems occur, the frames remain on the first ST. Since any single ST is loop free by nature, these frames will not encounter any loop and will be arrived at the destination. If a problem occurs with a link on the path of the first ST, STEP switches, or elevates, the frames to the second ST. Since STEP prohibits frames to be switched, or demoted, from a higher ranked tree to a lower ranked tree, a flow cannot be switched to the first ST. Therefore, it remains on this second ST. Once again, the same argument applies. The second ST is viewed as a single ST because frame demotion is not allowed; thus, frames traversing on this ST will not tangle in any infinite loop. By induction, the loop free property is guaranteed for higher order STs providing STEP switches to different STs.

STEP forms multiple STs by creating a set of single independent STs by running RSTP for each ST. The independency of the ST does not imply completely link disjoint. In fact, the physical topology dictates how much overlapping occurs. The implication of inde-

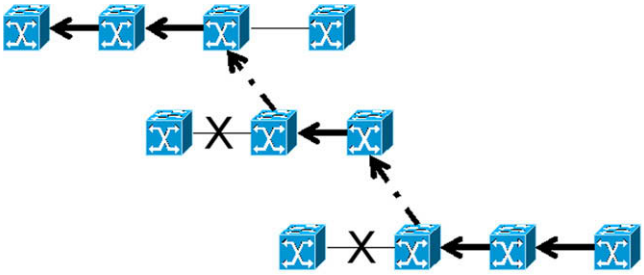

Fig. 2. A schematic of a three-layered Spanning Tree. Prior to a failed link (indicated by a cross), STEP elevates the traffic to the next highest Spanning Tree in the sequence. Traffic traverses from right to left.

pendent STs is that there is no co-relation between any two STs. The creation of the next ST in the sequence does not require any knowledge of the previously created ST, except the id of the regional root. Each of these STs is individually guaranteed to be loop free because RSTP blocks all the redundant links. It is possible that a frame might repeat the link they have traversed once before, but will not become stuck in an infinite loop as it will not repeat any path supervised by the first ST. The loop free property is guaranteed because the second ST is independent from the first ST, and it is also guaranteed individually by the RSTP loop free property.

For example, let there be three STs on four nodes: ST1 (A–B–C–D), ST2 (A–C–B–D), ST3 (D–A–B–C) as shown in Fig. 3a. To go from A to D, initially, ST1 is used; therefore, the path is A–B–C–D. If link C–D breaks, the traffic will be elevated to ST2 at node C; therefore, the path is now ABCBD as shown in Fig. 3b. We have a local loop B–C–B, but it is only transient. Later, if link B–D breaks, at B, the traffic will be elevated to ST3 so that the new path is ABCBAD as shown in Fig. 3c. The local loop occurs at ABCBA. Even though the frames revisit the nodes B and A creating a loop, the loop is only temporary so that the traffic can be elevated to the next tree, and thus exiting the loop. In addition, the local loops do not affect or create problems for the backward address learning process. Since the addresses are learned per VLAN; and each VLAN is associated with a ST, the switching does not create the ping-pong effect when forwarding frames. Each VLAN only knows its own learned addresses on the original port. Therefore, it will not see the local loops.

3.3. Provisioning QoS using STEP: traffic policing

In this section, it is shown how STEP can be used in MEN for traffic policing based on the idea of packet color marking. Implementation details are also presented to show the differences between STEP and MSTP.

Fundamentally, the VLAN ids are used as markers for the frames. Initially, frames of all priorities are tagged with the same VLAN id at the ingress switch indicating in-profile traffic conforming to the service level agreement. Therefore, traffic of different classes start on the same ranked ST, namely ST1, reserving the bandwidth for the higher priority traffic. The frame's source address, incoming port, or CoS bits can be used to map the frame to its negotiated class of service. When a frame is to be sent out on a port that has reached its threshold (meaning the service level agreement has been compromised, or out-of-profile), it is elevated to the next ST in sequence, as illustrated in Fig. 4.

Similar to the packet coloring scheme, when a packet uses up the tokens in the green bucket, it must be re-colored as red to indicate an out-of-profile packet. Each class of service in STEP has a corresponding set of STs. The higher the priority, the more STs through which a traffic class can be elevated (or in packet coloring scheme terminology: *mark down*). Each ST can be viewed as a token bucket and the link threshold is the service level agreement, but without the expense of implementing one as it is built in. In STEP, since there are more token buckets being used than regular packet coloring scheme, there are multiple levels at which a frame is marked down before being dropped. After each markdown, the switch elevates the frame onto a different ST and potentially traversing a new path toward the destination. Essentially, the congested part of the network is avoided, which increases the link utilization, and marks the frame as out of profile.

Each class of service is constrained to a certain number of STs. The first ST is shared by all classes. The sharing decreases as the spanning order increases. Thus, the contention lessens as the frames ascend the ladder of STs as shown in Fig. 4. For example, if there are three classes of service and six STs, then one possible assignment is as follows: the lowest priority class is assigned to STs 1 and 2; the next priority class runs on STs 1 through 4; and the highest priority class uses all six STs. All three classes contend for the first two ST's. ST3 and ST4 have two classes competing; and ST5 and ST6 are exclusively for the high priority class. When a frame belonging to a traffic class reaches its last allowed ST and the threshold on the outgoing link is reached, it is automatically dropped. Therefore, traffic shaping is less expensive with STEP than with a comprehensive token bucket mechanism. In addition, higher priority traffic is better served because of the reduction in contention due to differentiation of service.

3.4. Implementation issues

A key requirement for STEP is backward compatibility with current protocols. As a consequence, the MSTP protocol, 802.1 s were leveraged to implement the functionality and operations needed by STEP. Since MSTP is backward compatible with RSTP and STP, STEP can interoperate with these and MSTP. Thus, STEP retains the advantages of MSTP while providing enhanced features in terms of resiliency, capacity, and load balancing. Although a network will not benefit from new STEP capabilities until all switches understand STEP, this enables existing switches to support STEP via an incremental firmware upgrade.

The decision as to whether a frame should be elevated to the next ST is performed on a per frame basis. The reassignment of a frame to the next ST occurs in the same time period as a write to the frame header. As each frame arrives at a switch, the outgoing link associated with the current frame is examined to determine the current network condition, as shown in Fig. 5. Thus, conditions such as failure, load imbalance, and unsatisfied QoS are detected locally by each switch, remaining faithful to the spirit of Ethernet. Therefore, when a problematic link is detected, only a subset of the end-to-end path is involved in rewriting the header for the affected flows. Consequently, it is possible for a flow to be present on multiple STs at a given time.

To optimize the benefit of having multiple paths, the root for each ST is chosen to be unique, if possible. In other words, to the

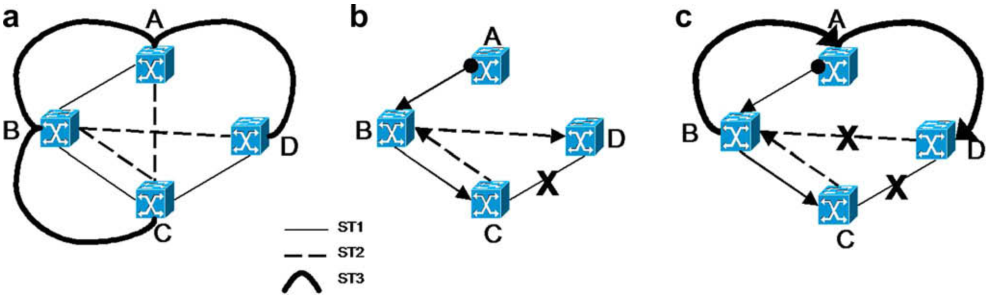

Fig. 3. An illustration of STEP operation in the face of link failures shown across the three topology diagrams.

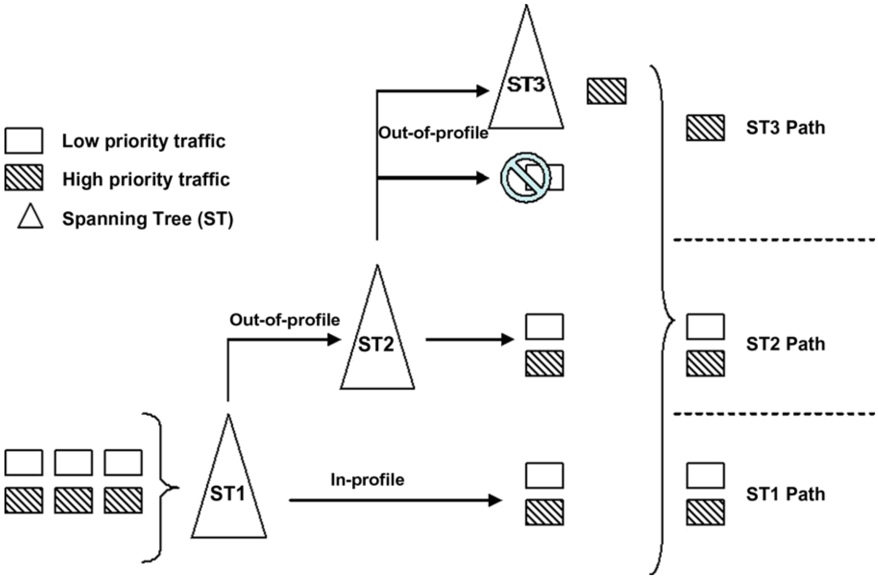

Fig. 4. STEP scheme for traffic policing.

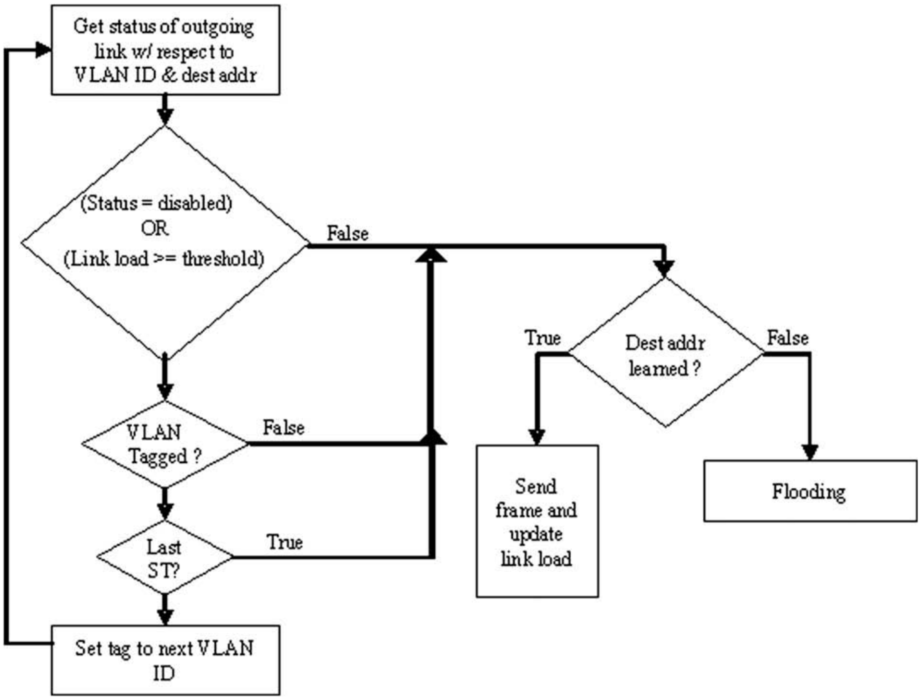
Get status of outgoing link w/ respect to VLAN ID & dest addr
(Status = disabled) OR (Link load >= threshold)
False
VLAN Tagged ?
False
Last ST?
True
Set tag to next VLAN ID
Dest addr learned ?
True
False
Send frame and update link load
Flooding

Fig. 5. Pseudocode for STEP.

extent possible each instance of the ST avoids sharing the regional root. Since the STP uses the shortest path to the root approach, having unique roots increases the chances of constructing disjoint trees. STEP does not propose a new algorithm to create these optimal STs. The performance of STEP will benefit from a well selected set of STs. STEP can leverage previous research to create optimal STs, such as [5,7,9] and [10].

STEP relies on the return traffic to notify the source that the ST used to forward the frame to a specific destination has been elevated. At each switch, there is a MAC-to-VID table that stores the MAC address and its associated VLAN ID. For every frame arriving at a switch, if the VID of the frame is anything other than the default, its source MAC address and VID will be recorded in the MAC-to-VID table. For every outgoing frame originating from the switch, a check is performed by the switch to see if the destination MAC address has an entry in the MAC-to-VID table. If an entry exists, the outgoing frame will use the VID from the MAC-to-VID table; otherwise, the default VID for the default ST is used.

The primary performance enhancement of STEP is the avoidance of the lengthy reconvergence procedure. Therefore, the reconvergence behavior of MSTP is adapted in the following ways:

1. When a switch detects a fault or a link recovery on one of its ports, the Spanning Tree Algorithm (STA) no longer initiates the port state/role re-selection.

2. The STA no longer flushes entries in the filtering database and forwarding tables. The switch acts as if nothing has happened and the traffic is switched to the new ST for a "soft reconvergence".
3. When a link recovers, instead of setting the recovered port to blocking state and performing the reconvergence, the switch reinstates the original role of the port per ST.

After a prolonged period of operation, it is possible that a significant proportion of the traffic is flowing on the last available (highest elevation) Spanning Tree due to multiple failures without any corresponding recoveries. As STEP is prohibited from switching traffic to a lower order ST, in the worst case scenario STEP performance degrades to that of standard RSTP. Each switch monitors for this condition by keeping track locally: of any failure resulting in flows being elevated the next ST and the load on the highest ST. If the load exceeds the predetermined threshold, the switch will broadcast a topology reconvergence on the affected tree. If switches receive at least two of such messages from distinct switches reconvergence will be triggered. Switches are then permitted to enter a self-reconfiguring state by reelecting state/port role, flushing the filtering database and the forwarding tables as before. Therefore, STEP remains faithful to the decentralized nature of Ethernet.

The original intention of having VLAN tags is for isolating traffic. As STEP uses VLAN tags as ids for STs, the original objective of VLANs is preserved. Instead of mapping a VLAN id to a traffic group, STEP can be implemented to map a set of VLAN ids that represent the STs to a traffic group. The VLAN partition is implementation dependent. The shortage of VLAN ids can be an issue, but there are proposals to perform VLAN stacking or Q-in-Q [14,15]. This technique increases the number of VLAN tag from 2^{12} to 2^{24}. Viking [5], a related work, also uses VLAN tag as the identification for multiple STs.

4. Simulation design

The OPNET [11] simulator tool was chosen because of its comprehensive implementation of Ethernet. OPNET includes implementations of RSTP, MSTP, and VLAN which are crucial to the evaluation of STEP.

STEP has been evaluated on two topologies: a topology representative of MANs [16] and a 6×6 grid topology, as seen in Figs. 6 and 7, respectively. A grid topology which inherently contains high degree nodes is included to show the impact of STEP on various topologies. Providing multiple alternative paths exist, the network will yield the benefits of STEP.

In the MAN topology of Fig. 6, RSTP has only a single ST configured on each side of the router. The initial RSTP's Spanning Tree is shown in Fig. 29 (Appendix). The root of the ST is located at the switch **core6**. By contrast, MSTP and STEP have four STs configured: the common root is at **core6**; the regional root for **vlan10 (ST1)** and **vlan40 (ST4)** is at **core1**; and the regional root for **vlan20 (ST2)** and **vlan30 (ST3)** is at **core2**. The ST configuration can be viewed in Fig. 29 through Fig. 32 (Appendix). Each VLAN represents a single ST and, similar to RSTP, the ST stops at the router.

Likewise, RSTP has one ST operating in the 6×6 grid topology in Fig. 7. The root of the tree is located at **node_14** which is the center of the topology. Conversely, MSTP and STEP is configured with six STs. The common root is at **node_33**. The regional roots for **ST1** through **ST6** are in the following order: **node_30**, **node_7**, **node_2, node_15**, **node_22**, **node_29**. As specified in the 802.1D standard, there are a maximum of seven hops. In order to form a stable ST for a topology of this size, the "hop count" parameter is increased.

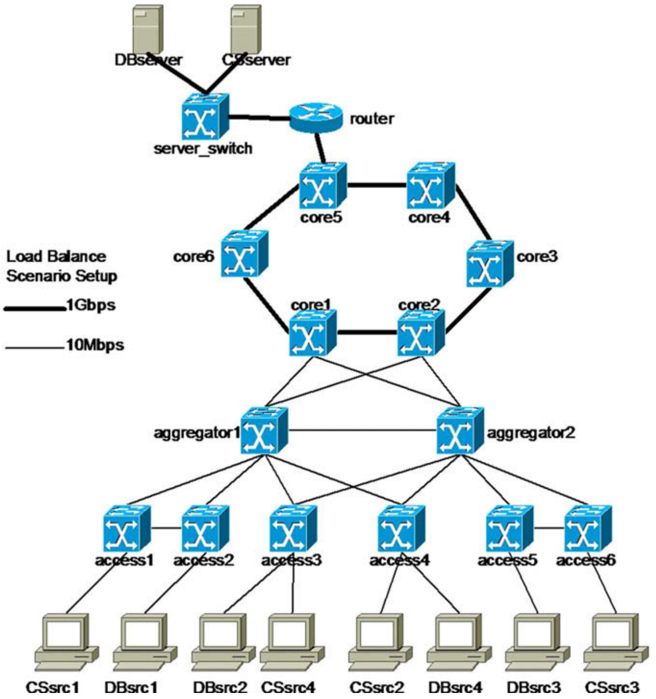

Fig. 6. A representative Metro Area Network (MAN) topology.

4.1. Metro Area Network Topology

Using the MAN topology in Fig. 6, resilience, load balancing, and service differentiation are evaluated. The description and specific parameters used are included. The notation ↔ indicates the link between two objects. For example, node_27↔node_28 means the link between node_27 and node_28.

4.1.1. Failure scenarios

There are six traffic flows from CSsrc{1, 2, 3} and DBsrc{1, 2, 3} to DBserver where each flow is a video conferencing session that starts after 100 s has elapsed, thus allowing the standard ST initialization to complete. Although irrelevant to the traffic profile, the notation "CS", "DB", and "src" stands for Content Server, Database, and source, respectively. Similarly, the notation G*n*, S*n*, B*n* in Fig. 7 stand for Gold, Silver, and Bronze. Again, the notation is irrelevant to the traffic profile and priorities until discussed Section 8. All links have a capacity of 1 Gbps. The simulation runs for a duration of 240 s. The link failures and link recovery are scheduled as follows:

- 120 s: aggregator1↔core1 fails.
- 140 s: aggregator1↔core2 fails.
- 180 s: aggregator2↔core1 fails.
- 220 s: aggregator1↔core1 recovers.

The results of the simulation experiment for RSTP, MSTP, and STEP are presented in the next section. Cumulative throughput was selected as the metric for comparing the resilience, as the difference in throughput clearly illustrates the performance loss for each of the protocols.

For MSTP, each traffic source is assigned to a ST in a round robin fashion from left to right in Fig. 6 as shown in Table 1. The traffic remains on the assigned ST throughout the simulation. However, STEP has the same traffic profiles as MSTP except that all sources initially begin with **vlan10 (ST1)**.

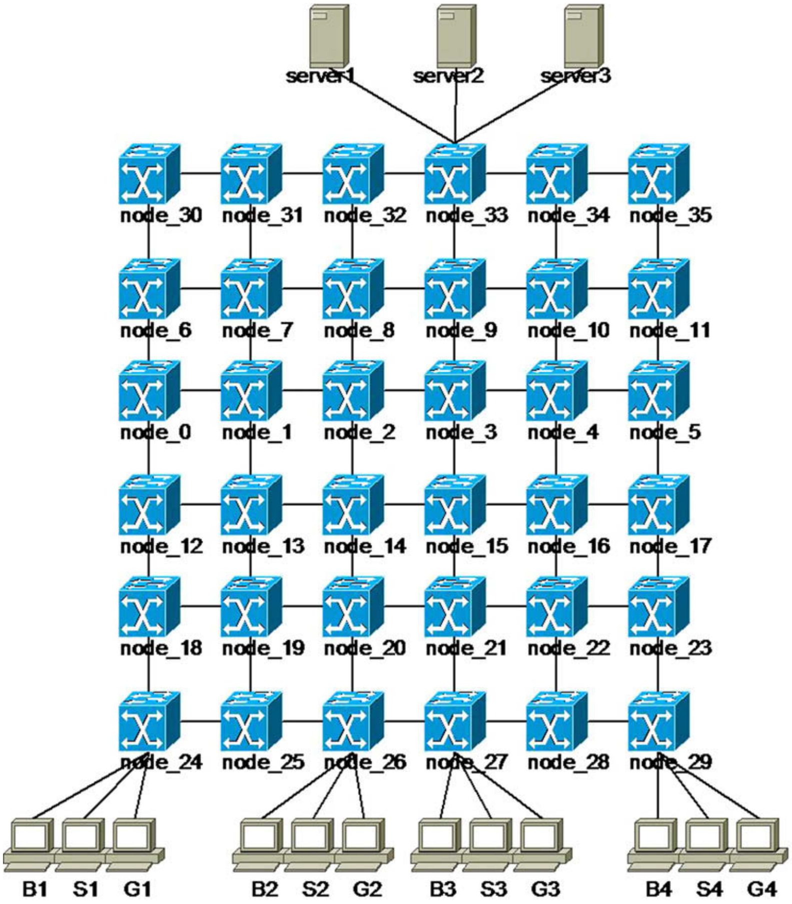

Fig. 7. A 6 × 6 grid topology.

Table 1
MSTP traffic mapping.

Traffic source	VLAN	Spanning Tree
CSsrc1	10	1
CSsrc2	40	4
CSsrc3	20	2
DBsrc1	20	2
DBsrc2	30	3
DBsrc3	10	1

4.1.2. Load imbalanced scenarios

To evaluate load balancing, the MAN topology from Fig. 6 was used but with two additional sources. CSsrc4 and DBsrc4 are added to **access3** switch and **access4** switch, respectively, Hence, there are now eight traffic flows from CSsrc{1,2,3,4} and DBsrc{1,2,3,4} to DBserver, where each flow is a video conferencing session starting at 100 s. The simulation runs for 240 s. However, there are no link failures in this experiment.

The link capacities are shown in Fig. 6. Our original OPNET simulations used explicitly modeled Gbps datagram traffic, however OPNET proved unable to process these traffic settings, citing insufficient memory. Therefore, the traffic was scaled back to 10 Mbps in order for OPNET to process the simulation. However, results for 1 and 10 Gbps links will be presented using extrapolation with the corresponding standard deviation and standard error.

The traffic load is distributed evenly for MSTP, with each ST having two sources. CSsrc1 and DBsrc3 traverse on ST1; DBsrc1 and CSsrc3 traverse on ST2; DBsrc2 and CSsrc4 traverse on ST3; and CSsrc2 and DBsrc4 traverse on ST4. In contrast, all of the traffic starts on the same initial ST for the STEP experiment.

4.2. Grid topology

Similar to the MAN topology, resilience, load balancing, and service differentiation are evaluated for the grid topology. The resilience test in this topology is more rigorous in that it includes both node failures and link failures. Whenever a node failure occurs, all of the links attached to the node also fail.

4.2.1. Failures scenarios

There are four flows to each of the three servers in Fig. 7. Each flow is a video conferencing session, again starting at 100 s. All links have a capacity of 100 Mbps. This capacity is sufficient to transport the traffic without causing any congestion. The simulation runs for a duration of 180 s. There are a total of 26 failed links and six failed nodes. The link failures and node failures are scheduled as follows:

- 110 s: node_7 fails.
- 110 s: node_8 ↔ node_9 fails.
- 120 s: node_10 fails.
- 130 s: node_13 fails.
- 140 s: node_14 fails.
- 140 s: node_27 ↔ node_28 fails.
- 150 s: node_16 fails.
- 150: node_20 ↔ node_21 fails.
- 160 s: node_23 fails.
- 160 s: node_32 ↔ node_33 fails.

In the MSTP experiment, the flows are grouped by the destination to put into the corresponding tree. For example, since S1, S2, S3, and S4 are going to **server1**, they are transported on the same ST. Again, all of the traffic starts on the same initial ST for the STEP experiment.

4.2.2. Load imbalanced scenarios

The load balancing experiment is similar to the configuration of the above resilience experiment, except that all links are now 10 Mbps. As before, the reason is due to simulation efficiency and resources. Since the bottleneck for the RSTP experiment are the links on the path from **node_33** to **node_21**, for fairness these links were upgraded to 100 Mbps for all the protocols. The simulation runs for a duration of 170 s, and no link failures were schedule in this experiment.

5. Enhanced Ethernet resilience

As alluded to earlier, resilience is of particular importance for carriers and this is one area for which Ethernet is well recognized as being very weak. STEP was specifically formulated to address the inherent weakness of Ethernet resilience. Results are presented in this section in which RSTP, MSTP, and STEP are evaluated for their resilience in the face of link failures and recoveries. The results of the MAN topology from Fig. 6 are presented first and described in detail to illustrate the behavior of STEP. The results from the grid topology are then explained to demonstrate the impact of STEP on a more dense topology.

5.1. Performance in Metro Area Network Topology

This subsection reports the performance of each individual protocol in the face of failures. A graph superimposing the results allows for comparison between the protocols.

5.1.1. RSTP

The throughput for RSTP as observed by the receiving host during the link failures is depicted in Fig. 8. As expected, when a link fails, RSTP reconverges and a dip in the throughput is witnessed before the new link assumes responsibility. Fig. 8 shows the effect of failures in the network at different times. The first dip accounts for the link failure at 120 s; and it takes 10 s for RSTP to reconverge. Following the reconvergence, the link **aggregator1** ↔ **core2** is unused, which explains why no dip is observed for the link failure at 140 s. The second dip accounts for the link failure at 180 s; while the third dip is the result of the link recovery at 220 s. Similarly, the second reconvergence takes 10 s and the third reconvergence takes 7 s.

The first reconvergence directly affects the path of any sources connected to access{1,2,3,4} except for sources connected to access{5,6}. This is shown when comparing Fig. 29 to 33; specifically, the link **aggregator1** ↔ **core1** is replaced by **aggregator1** ↔ **aggregator2**, thus traffic flowing through access{1,2,3,4} is affected. Meanwhile, traffic from access{5,6} remains on the same path. However, after the second reconvergence, the new ST affects all traffic. This can be observed in Fig. 8 by the depth of the throughput dips at 120 and 180 s. Following the recovery event, the reconvergence results in the subtree connected to access{1,2,3,4} reverting back to the original tree, while the subtree that is connected to access{5,6} alters to a new structure. Since reverting to the original structure reconverges faster, the performance hit results only from the sources connected to access {5,6}.

5.1.2. MSTP

Fig. 9 illustrates the impact of failures in the MSTP network. The first dip accounts for the failure at 120 s. The second dip occurs at 140 s on account of **aggregator1** ↔ **core2** failure. Unlike RSTP, MSTP utilizes more links; therefore, it also suffers a performance hit on the second failure. However, the performance hits for RSTP are much more severe than for MSTP. It is confirmed in Fig. 9 that the dips are not as deep as in Fig. 8. The third and fourth dips are the result of **aggregator2** ↔ **core1** failure and the recovery of **aggregator1** ↔ **core1**, respectively. On average, each reconvergence takes 7 s.

5.1.3. STEP

The throughput on the intermediate links between the core switches and the aggregator switches can be seen in Fig. 10. Unlike Figs. 8 and 9, Fig. 10 shows a snapshot of the traffic on the intermediate links while in transit to the destination. This is referred to as "partial" throughput as opposed to the "total" throughput that is collected at the end host where all traffic converges. The total throughput of STEP is shown in Fig. 11. In this figure, the granularity of the data collected is lower to illustrate the handoff between different STs, as shown by the drops in the throughput where the next link assumes responsibility at the same time the previous link fails. These drops do not affect the overall throughput received by the end host. The uninterrupted service is evidenced in Fig. 11. As prescribed by the monotonically increasing property of STEP, the traffic is initially sent on vlan10 (**aggregator1** ↔ **core1** link as shown in Fig. 29). Following a failure, the link for vlan20 (**core2** ↔ **aggregator1** link as shown in Fig. 30) takes over; and when that link fails, the link for vlan30 (**aggregator1** ↔ **aggregator2** link as shown in Fig. 31) takes over. When

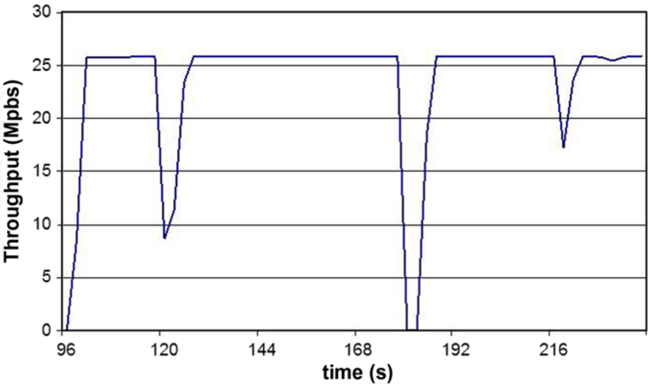

Fig. 8. The total throughput as observed by the receiving host during the link failures for RSTP.

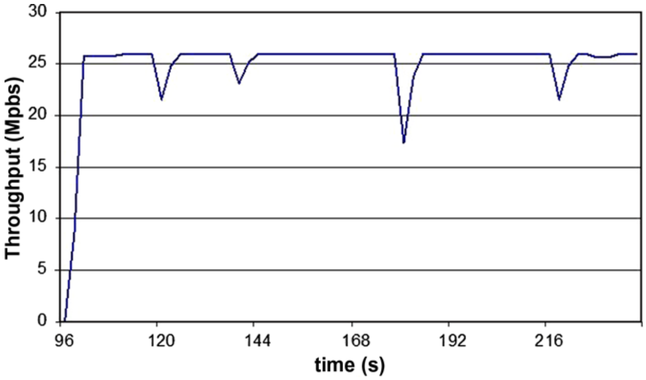

Fig. 9. The total throughput as observed by the receiving host during the link failures for MSTP.

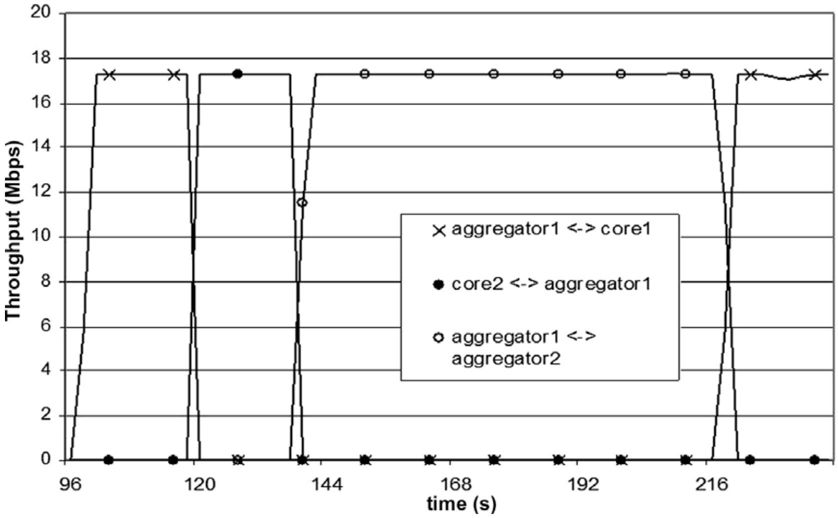

Fig. 10. The partial throughput as observed on various intermediate links in the topology as links fail and STEP re-routes traffic resiliently to maintain constant throughput.

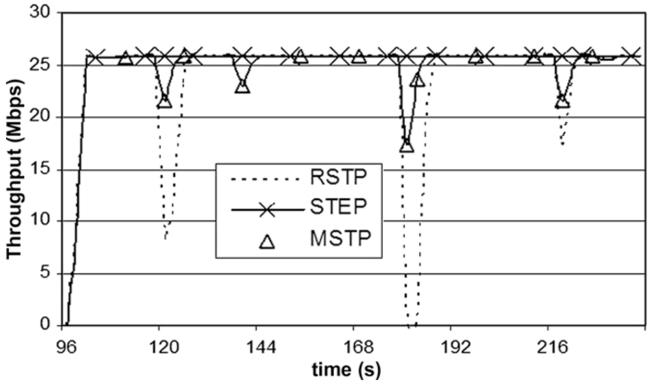

Fig. 11. In contrast with RSTP and MSTP, STEP maintained a constant throughput to the receiving host despite link failures and recoveries.

aggregator1↔core1 recovers, the recently arrived frames do not have to crossover so that the rate of **aggregator1↔core1** picks up again showing the dynamic nature of STEP.

5.1.4. Comparison of RSTP, MSTP, and STEP and extrapolation results

For this MAN topology resilience scenario, the comparative performance of RSTP, MSTP, and STEP can be visualized by superimposing the cumulative throughput graphs of RSTP, MSTP, and STEP as seen in Fig. 11. While all three protocols reach the same maximum throughput during the normal operational periods, it was shown that during the faulty periods, STEP was able to maintain a sustained throughput.

Despite incurring one or more link failures, STEP has been designed to minimize the frequent execution of the ST reconvergence algorithm. As seen in Fig. 12, the reconvergence time for STEP is zero at each link failure or recovery event. If one or more failed links recover before STEP exhausts the available VLANs (monotonic increase), then it is possible that no reconvergence is ever required despite links failing. Therefore, STEP is able to operate continuously without interruption to the service. This is a clear advantage of STEP over RSTP and MSTP. To measure the throughput graph, the lost percentage (the area of the dipped region) of RSTP and MSTP is compared against STEP. The results show that RSTP loses 7.3% of the total received traffic compared to STEP; and MSTP loses 1.69% of the total received traffic compared to STEP. In addition, STEP provides uninterrupted operations for real time services.

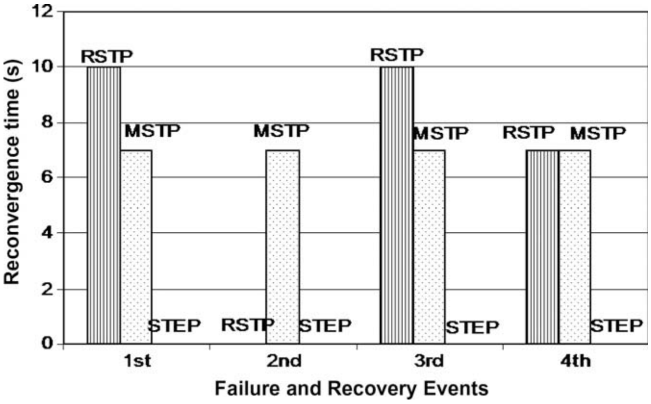

Fig. 12. The reconvergence time for the three protocols at each event.

These losses are significant in the MEN setting where the link capacity is in the gigabits range, as shown in Table 2, the losses per second in the 1 Gb network and the 10 Gb network. These data are projected by calculating the offset between STEP and RSTP maximum throughput and STEP and MSTP maximum throughput. Where the link capacity is 1 Gbps, in the face of failure, RSTP loses 313.1 Mbps on average compared to STEP; and MSTP loses 22.6 Mbps compared to STEP. Both loss rates have a standard deviation of 0.03586 Mbps. Meanwhile, the effect of the loss is exacerbated 10 fold in a 10 Gigabits environment. During the network down time, RSTP loses 3131.634 Mbps while MSTP loses 226.26 Mbps.

5.2. Performance in a Grid Network

In this scenario, Fig. 13 shows the cumulative throughput collected at the three servers. Both RSTP and MSTP begin to incur a performance hit after 130 s, whereas STEP maintains constant

Table 2
The lost throughput in Mbps averaged out during the network down time that could be recovered had we deployed STEP.

	1 Gb	10 Gb
RSTP (Mbps)	313.1	3131.64
MSTP (Mbps)	22.6	226.26

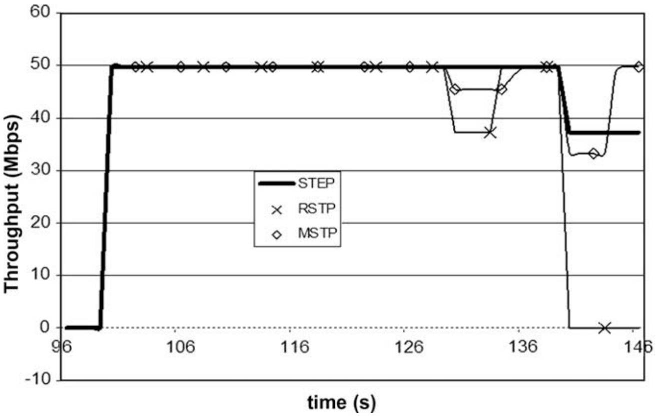

Fig. 13. The cumulative throughput of illustrates re-covergences for RSTP and MSTP, while STEP elevates some flows to the last Spanning Tree.

throughput until 140 s. In contrast to RSTP and MSTP, STEP does not reconverge for every single failure or recovery. Therefore, after numerous failures without any recovery, a proportion of the traffic flows are transported on the last ST. Evidence for this can be seen in Fig. 13 where STEP's graph line does not return to the maximum rate. Until 145 s, RSTP loses 7.69% of the total traffic compared with STEP, and MSTP loses 1.69% of the total traffic compared with STEP. After 160 s, the majority of the traffic in the STEP experiment is transported on the last ST. As a consequence of this the throughput for STEP drops to 24.9 Mbps. As mentioned in Section 3.4, this situation can be detected and simply averted by triggering a reconvergence. However, if one or more of the links recover, then reconvergence may not be necessary.

6. Ethernet switch load balancing

By being able to distribute the traffic, or load balance, across various links in a network, it is possible to increase the capacity and utility of the network. However, none of the existing Ethernet protocols allow the carrier to control load balancing dynamically across all the links in the network. The STEP algorithm will facilitate load balancing across all links in the metro Ethernet. The carriers will thus have an option for balancing load across the network, as well as fine-grained control of the load on individual links. This will be an attractive feature for the carriers as they can exploit maximal throughput, and thereby capacity from the network. Similar to the resilience simulation, the results of the topology from Fig. 6 are presented first and in detail showing the behavior of STEP. Then the results from the grid topology will be shown from the overall perspective. The traffic load generated saturate the links to their maximum capacity creating a highly congested scenario.

6.1. Performance in Metro Area Network

The load balancing performance for MAN topology is examined in this section starting with RSTP. In order to see the inefficiency of RSTP and MSTP, the utilization of the intermediate links between the core switches and the aggregator switches are shown. As can be appreciated in Fig. 6, the "left side" of the network refers to the links connected to **aggregator1**, and the "right side" of the network refers to the links connected to **aggregator2.**

6.1.1. RSTP

In this section, the traffic on intermediate links in the network was measured to show the inefficient link utilization of RSTP due to the inability to balance network load. The metric used in the graphs for demonstrating the efficacy of the respective protocols is again the cumulative throughput.

At **aggregator1** switch, there are three links that can potentially carry the traffic into the core network. However, in order to prevent loops, RSTP blocks two of those links. As shown in Fig. 14, the **aggregator1** ↔ **core1** link is loaded to its maximum capacity. Despite approximately 25.8 Mbps arriving at **aggregator1**, the output from the switch is only 10 Mbps. RSTP cannot use the other two links to transport the remaining traffic as they are blocked which forces **aggregator1** to drop the remaining traffic. In Fig. 15, 8.6 Mbps arrives at **aggregator2**and, since the **core1** ↔ **aggregator2** link has the capacity, no frames are dropped. However, if an overload situation occurs, excess traffic will be dropped because the **aggregator2** ↔ **core2** link is blocked.

6.1.2. MSTP

For MSTP, at **aggregator1**, 4.3 Mbps arrives for ST1, another 4.3 Mbps comes from ST2, and 8.6 Mbps is for ST3. Since ST3 root is at **core2** and the link **aggregator1** ↔ **core2** blocks ST3, the 8.6 Mbps must travel via **aggregator2**. At **aggregator2**, there are 4.3 Mbps for ST1, 4.3 Mbps for ST2, 8.6 Mbps for ST3 that comes from **aggregator1**, and 8.6 Mbps for ST4. As ST2 and ST3 share the same link, the capacity on the link **aggregator2** ↔ **core2** is exhausted causing frames to be dropped. The link **aggregator2** ↔ **core1** only carries ST1 traffic which is 4.3 Mbps, as shown in Fig. 17. The traffic for ST4 is sent via **aggregator1** because the root for ST4 is at **core1** and only link **aggregator1** ↔ **aggregator2** allows traffic for ST4. The link **aggregator1** ↔ **core1** carries the combined traffic of ST1 and ST4 (arriving from **aggregator2**), thus

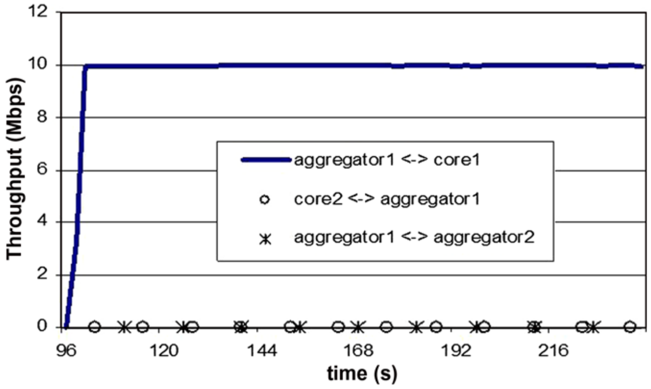

Fig. 14. The throughput as observed on links in the left side of the access network topology shows underutilization.

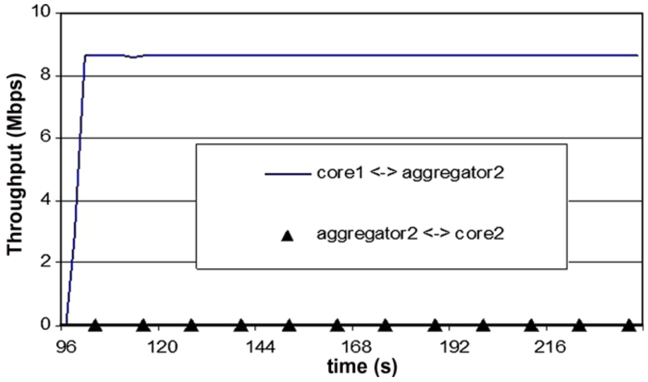

Fig. 15. The throughput as observed on links in the right side of the access network topology also shows underutilization.

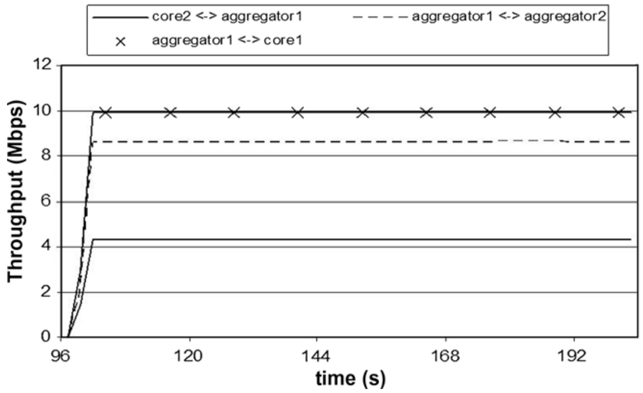

Fig. 16. The throughput as observed on various links on the left side of the access network topology.

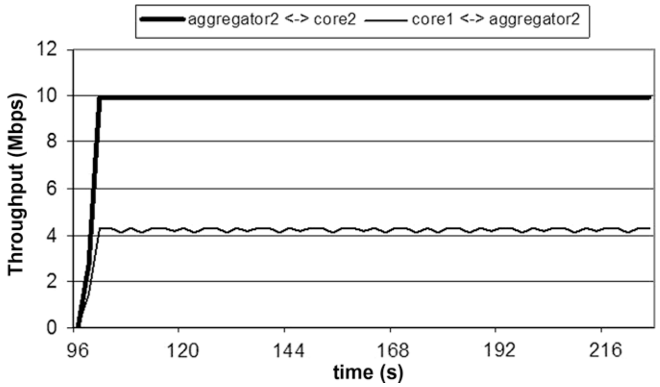

Fig. 17. The throughput as observed on various links on the right side of the access network topology.

is maxed out at 10 Mbps. The link **aggregator1↔core2** only carries traffic for ST2, and hence uses only 4.3 Mbps. The link **aggregator1↔aggregator2** directs ST3 traffic to **core2** as explained earlier, thus using 8.6 Mbps. This behavior is captured in Fig. 16.

Although this shows that it is possible to distribute the load in MSTP, it is not efficient. Even if the access ports are reconfigured to distribute the load across the network, it applies only to a specific situation. Due to the dynamic and unpredictable nature of packet switched traffic, there is no single static configuration that works for all. Unlike STEP, MSTP cannot be responsive to traffic conditions.

6.1.3. STEP

As with the RSTP experiment, the traffic on intermediate links in the network was measured to show the resulting link utilization. In this instance of the experiment, the link utilization threshold for load balancing was set at 80%. This means that for a link capacity of 10 Mbps, the switch will permit at most 8 Mbps on that link before it will try to switch the traffic to the next ST, unless it is the last ST to which the traffic can be elevated. In this case, the last ST is ST4. The 20% reservation is used as a buffer on the link to protect against peaked traffic and control messages, such as BPDU. Since the traffic enforcement mechanism is configured to check every half second and each source streams at 4.3 Mbps, the reserved 20% (or 2 Mbps on a 10 Mbps link) on each link is sufficient to protect against traffic bursts and control messages.

Initially, STEP starts all traffic in ST1 with each source sending 4.3 Mbps. There are 24.6 Mbps arriving at **aggregator1** on ST1. The other 1.2 Mbps (0.6 Mbps from **access3** and another 0.6 Mbps from **access4**) is sent to **aggregator2** on ST2, because the load balance threshold for the link is 80%. The 1.2 Mbps is now on ST2. In addition, there are 8.6 Mbps arriving at **aggregator2** on ST1. Of the 24.6 Mbps arrived at a **ggregator1**, 8 Mbps is sent to **aggregator1↔core1** link on ST1, 8 Mbps is sent to **aggregator1↔core2** link on ST2, and 8 Mbps is sent on the **agrregator1↔aggregator2** on ST3 toward **aggregator2**. The remaining 0.6 Mbps is elevated to ST4 and is sent via link **aggregator1↔core1**. Since the **aggregator1↔core1** link is shared by the last ST, it is allowed to transport more than the 80% threshold. On the right side, **aggregator2** sends 8 Mbps to **aggregator2↔core1** link on ST1 and elevates the remaining 0.6 Mbps to ST2. Now, aggregator2 sends the combined 0.6 + 1.2 Mbps on ST2 to **aggregator2↔core2** link. The 8 Mbps arriving to **aggregator2** from **aggregator1** on ST3 needs to be sent out on link **aggregator2↔core2**, and this link is shared by ST2 and ST3. However, the **aggregator2↔core2** link is capped at 8 Mbps, thus, 1.8 Mbps traffic must be elevated to ST4 and sent back toward the root of ST4 at **core1**. Therefore, the link **aggregator1↔core1** receives additional traffic which totals 1.8 + 8 + 0.6 = 10.4 Mbps. Since the link is only able to transport 10 Mbps, some frames are dropped. This behavior is illustrated in Fig. 18 and 19.

6.1.4. Comparison of RSTP, MSTP, and STEP and extrapolation results

For this experiment, the comparative performance of RSTP, MSTP, and STEP can be visualized by superimposing the cumulative throughput graphs in Fig. 20.

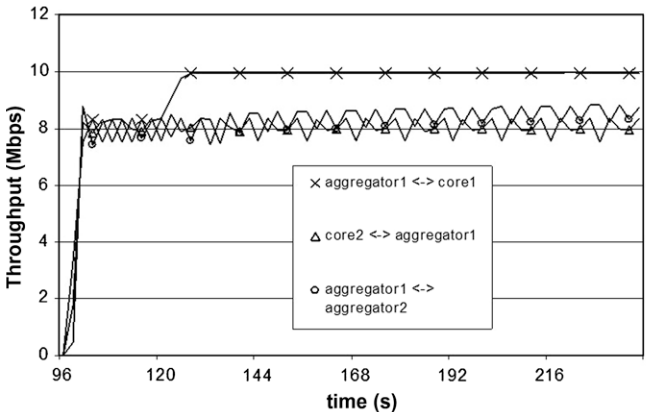

Fig. 18. The throughput as observed on links on the left side of the topology showing STEP improving and balancing link utilization.

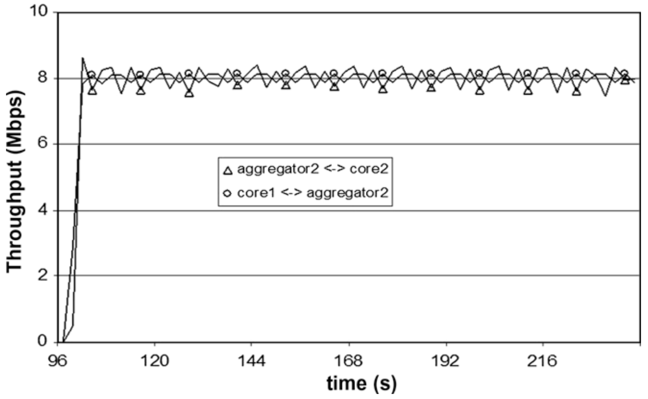

Fig. 19. The throughput as observed on links on the right side of the topology showing STEP improving and balancing link utilization.

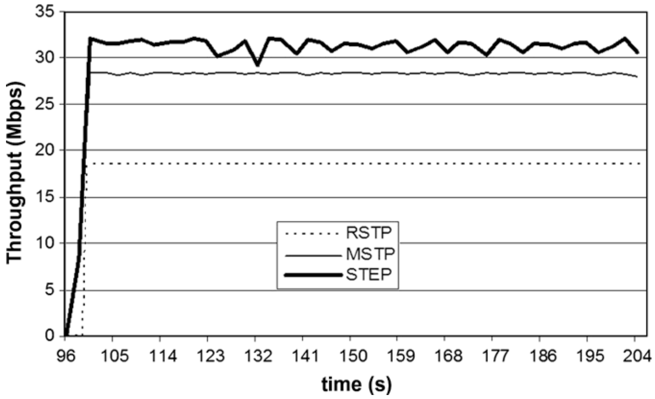

Fig. 20. The cumulative throughput of RSTP, MSTP, and STEP for MAN topology.

As the bottleneck for RSTP is at **aggregator1** where only 10 Mbps can be sent out, by combining the traffic from **aggregator2** the end host receives approximately only 17–18 Mbps. Although MSTP is able somewhat to redistribute the load, it is difficult to find an optimum assignment of the loads for a balanced network. Heuristic assignments provide limited assistance in this regard. By contrast, without any static pre-configuration STEP dynamically redistributes the traffic if the current link is congested, thus able to accommodate flexibly increased incoming traffic. The fluctuation effect as observed in the throughput for STEP in Fig. 18 through Fig. 20 is the result of STEP stabilizing around the link utilization threshold. In our simulation, the current link utilization is measured once per second, and naturally these fluctuations can be smoothed further by selecting a smaller time period. Fig. 20 shows the overlaid throughput of three protocols at the receiver. Comparing against the traffic throughput of STEP, RSTP loses 37%; MSTP loses 12.76%.

The extrapolation results for load balancing are shown in Table 3 for 1Gbps link capacity and 10 Gbps link capacity. The former scenario has each source sending 420 Mbps of streaming voice, and the latter scenario has each source sending 4.2 Gbps of streaming voice. The extrapolated results are calculated from the offset between RSTP projection of STEP and MSTP projection of STEP. The enhanced STEP performance gains 1.17 Gbps or 163.47 Gigabits total in 140 s over RSTP and 0.218 Gbps or 30.62 Gigabits total in 140 s over MSTP with a standard deviation of 0.138 Gbps. Similarly, when the link capacity is 10 Gbps, STEP increases the throughput by 11.74 Gbps for a total of 1822 Gigbits in 140 s over RSTP and 2.21 Gbps for a total of 309.79 Gigibits in 140 s over MSTP with the standard deviation of 1.41 Gbps. In Gigabits network, STEP improves the performance greatly.

6.2. Performance in the grid topology

In the grid topology, heavy congestion forces the switches to drop frames and, consequently, none of the protocols achieved the maximum throughput of 51.6 Mbps. Fig. 21 shows that STEP is able to achieve higher throughput than both RSTP and MSTP. STEP delivers 8.8% and 9.2% more of the total traffic than RSTP and MSTP, respectively. It is interesting to note that in this case, the performance of RSTP and MSTP is almost equivalent. Even with multiple STs, MSTP still drops the same amount of traffic as RSTP. This is due to inefficient construction of the ST [5,7–9].

Table 3
The gained throughput by STEP over RSTP and MSTP.

	1 Gb	10 Gb
RSTP (Gbps)	1.16	11.75
MSTP (Gbps)	0.21	2.21

7. Evaluation of service differentiation

The metro topology in Fig. 6 is used in this section to demonstrate how STEP can support service differentiation. During the heavy congestion period frames will inevitably be dropped, but some service classes require higher performance than others. By differentiating the services, STEP is better positioned to meet the requirements of the higher traffic class. The traffic in two scenarios below have two priorities and four STs. The lower priority traffic uses only ST1 and ST2 while the higher priority traffic uses ST1 through ST4.

In this scenario, the simulation configuration is similar to the load imbalanced scenario in Section 4.1.2. The traffic is heavily congested in the aggregation part of the network causing **aggregator1** and **aggregator2** switches to drop frames. Despite the total combined throughput for both classes of service in MSTP and STEP being the same (approximately 27.2 Mbps), MSTP equally divides the bandwidth between the two classes of traffic as shown in Fig. 23. As a result, the higher priority traffic is deprived of the required bandwidth of 16 Mbps. Similarly, RSTP distributes the available bandwidth equally giving each class a throughput of 8.92 Mbps, as shown in Fig. 23. However, the combined throughput for RSTP is much less than MSTP or STEP, because with only one ST present no alternative path is available when the link becomes saturated. Unlike STEP, RSTP cannot balance the network

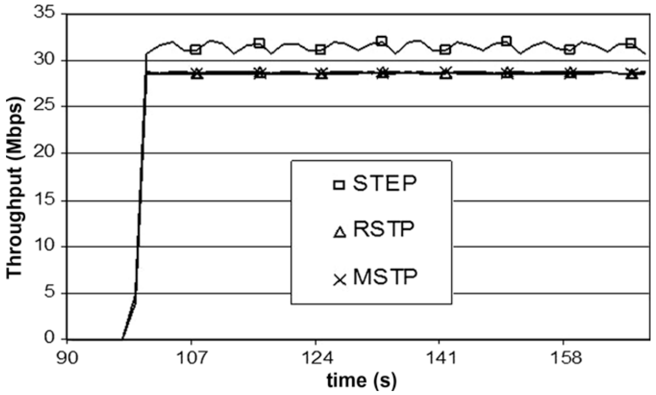

Fig. 21. The cumulative throughput of RSTP, MSTP, and STEP for grid topology.

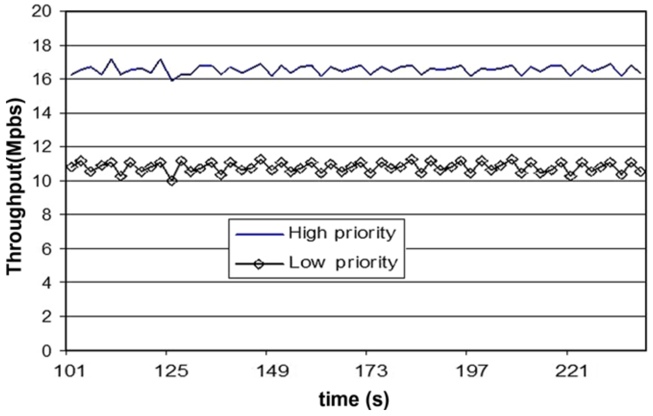

Fig. 22. STEP distributes bandwidth according to class of service.

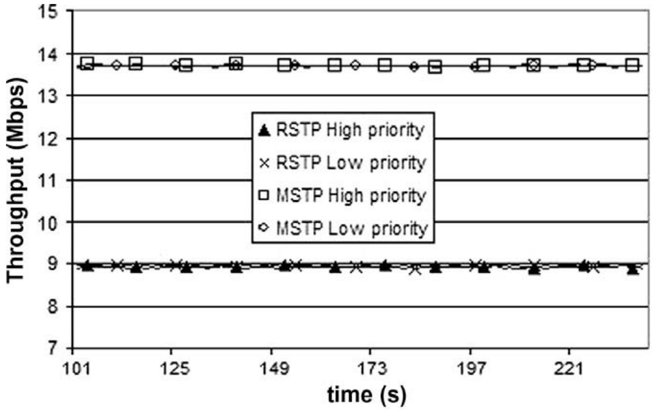

Fig. 23. MSTP and RSTP divides bandwidth equally for all service classes.

load in order to increase the throughput from congested links (see Fig. 23).

The throughput from Fig. 22 shows that the higher priority traffic is still able to receive the required 16 Mbps while the lower priority can only achieve 11.2 Mbps, which is 70% of the higher priority traffic. STEP clearly differentiates the two classes of traffic by reserving more bandwidth to support higher priority traffic, while MSTP and RSTP ignore the priorities of the traffic.

8. Impact of Spanning Tree allocation

It was discovered that above a certain threshold for a given topology, no performance advantage is gained by adding STs. This is due to all of the links being shared, therefore, subsequently allocated STs do not reveal any new links for alternate paths. As alluded to earlier, each ST creation avoids shared links as much as possible by picking a unique root.

The following scenarios evaluate the impact of ST allocation per service class. Since ST allocation is topology dependent, a strategy to find the optimal performance is presented. In this set of scenarios, there are three priorities and six STs that correspond to six VLANs. The performance is evaluated on a 36-node grid topology as seen in Fig. 7. From Fig. 7, **server1** receives the Best Effort (BE) lowest priority traffic. The medium priority traffic, Silver, is collected at **server2**, while the highest priority traffic, Gold, goes to **server3**. Each flow is a video stream of approximately 4.3 Mbps for a total of 16 Mbps for each traffic class. Each priority receives 4 flows for a total of 12 flows:

- Best Effort (BE): B1, B2, B3, and B4.
- Silver: S1, S2, S3, and S4.
- Gold: G1, G2, G3, and G4.

Since STEP manages the service differentiation by allocating a number of STs to a service class, an investigation into the threshold for the ST allocation is reported in this section. There are four scenarios that are named according to how the STs are allocated to the traffic classes, as shown in Table 4. For example, the 2–4–6 STs scenario allocates the first two STs to the BE class, first four STs to the Silver class, and all six STs to the Gold class. The performances of all scenarios are shown in Figs. 24–27.

To discover the threshold for ST allocation, the graph lines of each service class of adjacent pairs of ST allocation were contrasted. For instance, when we compare 1–4–6 to 2–4–6 of BE traffic class while keeping the Silver and Gold class allocations the same, it reveals that there is no difference in BE throughput. This implies that the threshold does not lie between allocating 1–2 STs. However, when contrasting 2–4–6 scenario to 2–5–6 scenario of the Silver traffic class while keeping the BE and Gold class allocation constant, the throughput increases significantly. Therefore,

Table 4
Spanning Trees allocation to service class.

Scenarios	Best effort	Silver	Gold
1–4–6	1 ST	4 STs	6 STs
2–4–6	2 STs	4 STs	6 STs
2–5–6	2 STs	5 STs	6 STs
3–4–6	3 STs	4 STs	6 STs

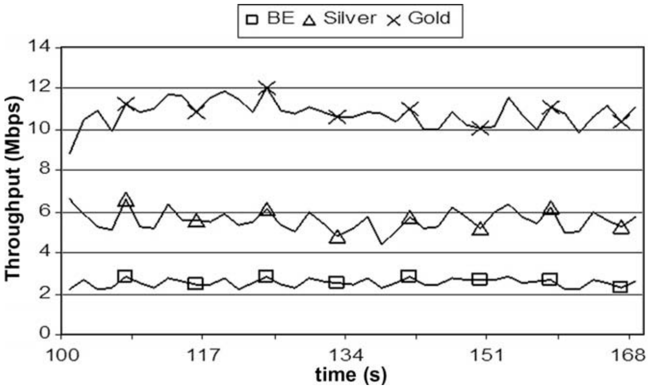

Fig. 24. 2–4–6 STs.

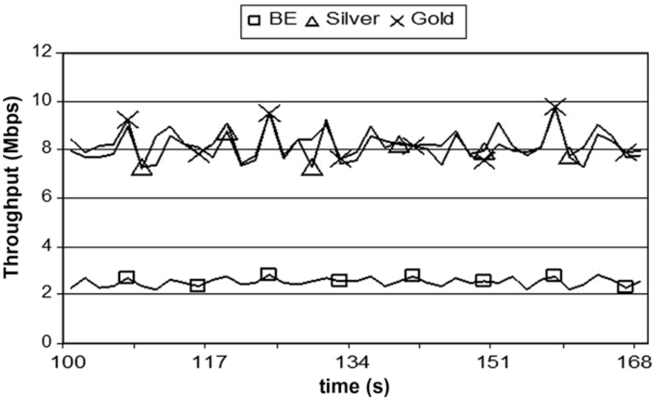

Fig. 25. 2–5–6 STs.

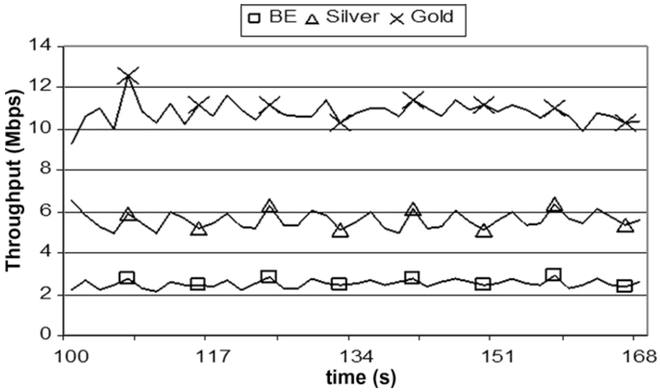

Fig. 26. 1–4–6 STs.

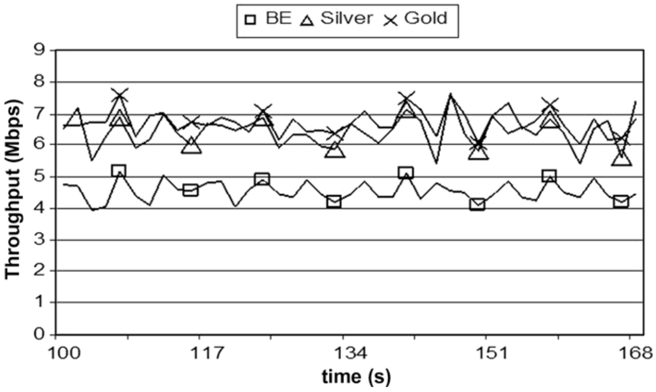

Fig. 27. 3–4–6 STs.

the threshold is determined to lie between allocating 4–5 STs. Continuing this strategy, the ST allocation for STEP for this topology can be derived. The threshold for the BE traffic class is placed between 2 and 3 Spanning Trees and the threshold for the Silver class is between 4 and 5 STs.

9. Out of order problem

Since STEP allows a flow to traverse multiple STs simultaneously, it is conceivable that frames belonging to the same flow may take different paths. Therefore, the potential exists for end hosts to receive frames out of order. If the out of order frames take a significant long time traveling on the suboptimal alternate path, they could trigger the time out mechanism that forces TCP to retransmit, then it has the potential to diminish TCP performance greatly.

In recognition of this scenario, a set of simulations were designed to test and quantify the effect of out of order frames. In this scenario, there are six sources uploading to a server. Each source uploads a single file of size 200 MB. The topology used is that of Fig. 6, with each node having 64 KB receiver buffer. Since a single RSTP or MSTP flow stays strictly on the same path (in normal condition), there are no outstanding out of order frames that hamper the performance of TCP. They are used as the base controlled traffic for STEP to compare with. Fig. 28 illustrates that the STEP simulation completes the file upload before both RSTP and MSTP. This demonstrates that the out-of-order-packet issue with STEP does not impact on the performance of TCP. The TCP retransmission mechanism was not triggered by any out of order frames. It is interesting to note here that RSTP performs better than MSTP. If MSTP concentrates the load onto a single ST, it behaves just like RSTP. However when the load spreads equally among the STs, the throughput deteriorates as the result of fairness in TCP.

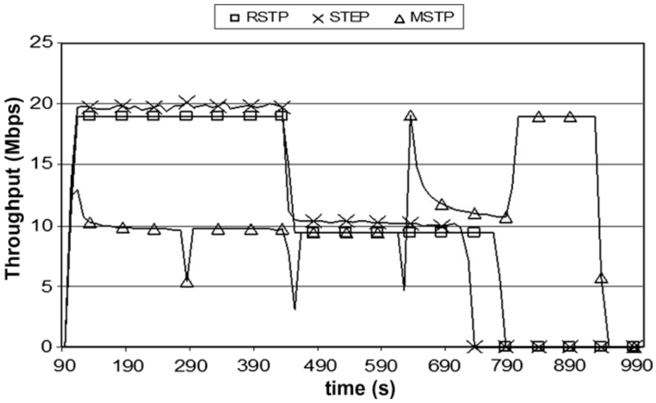

Fig. 28. TCP throughput for RSTP, MSTP, and STEP.

10. Related works

In an effort to provide QoS to MEN, the Metro Ethernet Forum (MEF) group defined traffic management requirements that include bandwidth profiling at the edge and inside the MEN [19]. At the MEN ingress, the customer class of service is mapped to the carrier class of service to be used inside the network. A set of parameters to control the traffic includes Committed Information Rate (CIR), Committed Burst Size (CBS), Excess Information Rate (EIR), Excess Burst Size (EBS), Coupling Flag (CF), and Color Mode (CM). The parameters are input into a bandwidth profile algorithm that verifies the conformance of the traffic. The algorithm is based on the packet coloring concept that uses token buckets. In addition, the specification also defines the frame delay performance, frame jitter performance, and frame loss ratio.

Viking is a Multiple Spanning Tree architecture proposed by Sharma et al. [5]. Viking precomputes multiple STs that can satisfy the required QoS metrics so that it can change to a backup ST in the event of a failure. The paths are computed based on the weight that is assigned to each link. A path aggregation algorithm is then used to merge the paths into the ST. Viking uses a client–server model that needs to be informed by the end hosts to update the server on the condition of network before the STs are periodically recomputed.

Ethereal [6] is an approach to flow reservation similar to Int-Serv, offering a real time connection-oriented architecture supporting best effort and assured service traffic at the link layer. When an application makes a request for a connection, it sends QoS parameters, the destination IP address, and the destination IP port number. The request propagates through the switches and is assigned a unique connection id. If the request survives to the destination and satisfies the QoS requirements throughout the path, the destination acknowledges the source, and all the switches commit to this connection.

SmartBridge [8] and STAR [9] are two approaches to improve upon the STP. They both find an alternate route that is shorter than the corresponding path on the ST. SmartBridge requires full knowledge of the topology, whereas STAR is an overlay approach requiring STAR-aware switches to be the super nodes of the topology. STAR calculates the shortest path from a super node to the next using the distance vector.

Lim et al. [10] address the underutilization of the standard Spanning Tree. They recognize also that the simple priority queuing of 802.1 potentially starves low priority traffic when the high priority traffic dominates a significant fraction of the traffic. In their scheme, they construct a ST for different multimedia traffic flows based on the defined category. Each category is defined by the tuple ⟨traffic type, VLAN⟩. On the other hand, non-multimedia traffic flows use the ST that is built just for it. Each flow remains in the designated ST with no crossing over permitted.

Another approach to load balancing is Tree-Based Turn-Prohibition (TBTP) [7]. TBTP constructs a less restrictive ST by blocking a small number of pairs of links around nodes, called a turn, so that all cycles in a network can be broken. The benefit of TBTP increases proportionally to the degree of the nodes. As MEN access networks have a low node degree, TBTP is of marginal benefit. Since TBTP relies on the standard STP to reconverge before it can re-compute its routing, the recovery time is in the order of seconds.

Instead of taking the ST approach, Rbridges [20] and LSOM [21] run link state protocol over the topology. Both protocols broadcast addresses to obtain the global view of the topology. Rbridges proposes a link state protocol similar to that of IS–IS. In addition to routing frames, Rbridges also optimize IP by having Rbridges aiding the ARP functionality. Similarly, LSOM uses the Dijkstra algorithm to calculate shortest path, but LSOM only apply to the backbone switches where the MAC addresses are stable and fewer to learn.

11. Conclusion

In this paper, a new concept of STEP is proposed for switching packets in the MEN. Simulation assumptions and designs were presented for evaluating STEP directly against the incumbent protocols of RSTP and MSTP. Simulations results were analyzed and presented that demonstrate the potential benefits that STEP can offer to the carriers. In addition, the implementation of STEP has a low complexity overhead and can leverage MSTP support already commonly available in Ethernet chipsets.

The primary focus of this work has been on the resiliency, load balancing, and service differentiation aspects of STEP. Results obtained through the numerous simulation experiments using OPNET revealed the following potentials of STEP:

- STEP is very resilient to failures. It requires no reconvergence in face of multiple link failures. The throughput provided by STEP is significantly higher than that of RSTP and MSTP.
- STEP accommodates the seamless re-integration and adoption when failed links recover, thereby obviating a significant proportion of reconvergences. This advantage thus increases the MTBF of switches.
- STEP provides load balancing of Ethernet frames throughout the metro access network, which is a new Ethernet layer feature. This gives carriers a new level of control over Ethernet traffic that they previously never had.
- STEP offers inexpensive and effective service differentiation and traffic policing for enabling Quality of Service in MEN.

Appendix A

See Figs. 29–35.

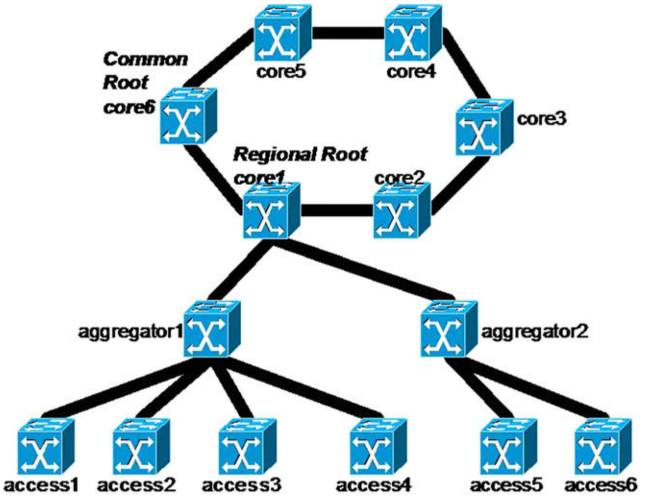

Fig. 29. ST1 configuration for MSTP and STEP and the initial ST configuration for RSTP before any failure.

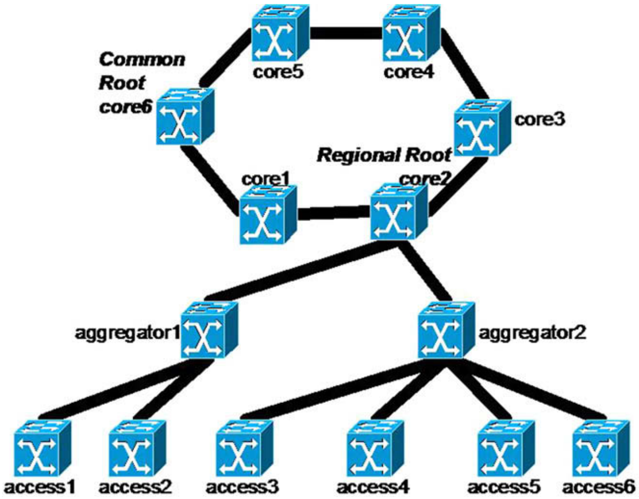

Fig. 30. ST2 configuration for MSTP and STEP.

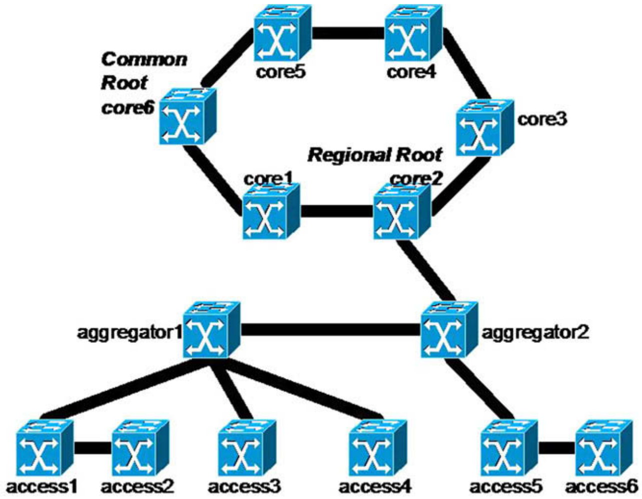

Fig. 31. ST3 configuration for MSTP and STEP.

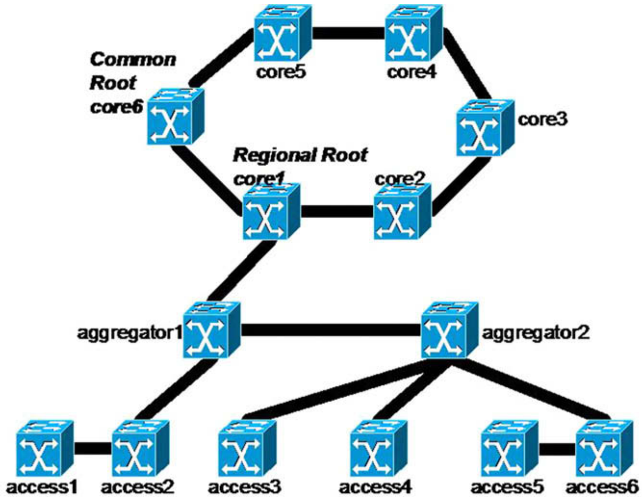

Fig. 32. ST4 configuration for MSTP and STEP.

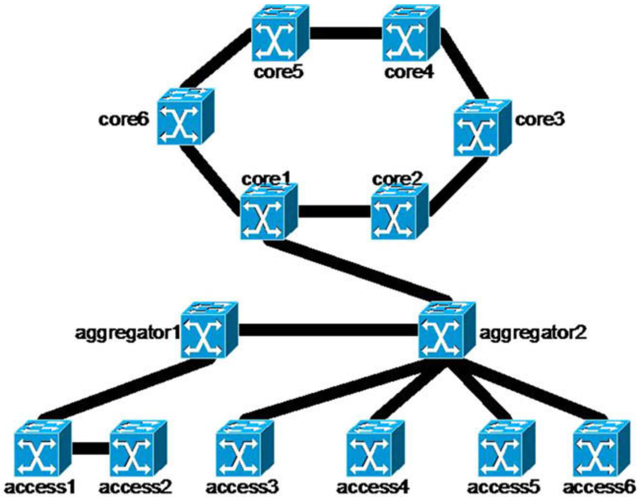

Fig. 33. RSTP reconverged after failure at 120 s.

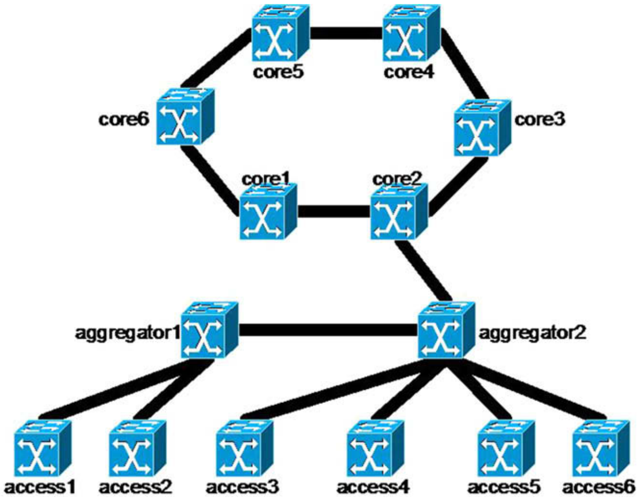

Fig. 34. RSTP reconverged after failure at 180 s.

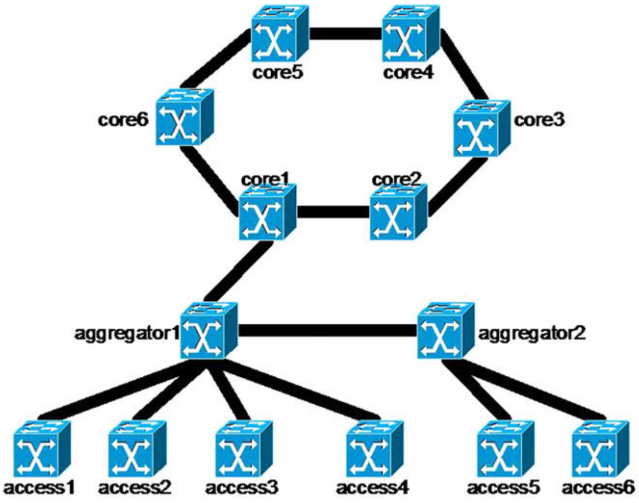

Fig. 35. RSTP reconverged after the link recovery at 220 s.

References

[1] IEEE Information Technology, Part 3: Media Access Control (MAC) bridges, ISO/IEC 15802-3, ANSI/IEEE Std 802.1D, 1998.

[2] IEEE Standard for Local and Metropolitan Area Networks – Amendment 2: Rapid Reconfiguration Amendment to IEEE Std 802.1D, 1998 Edition. IEEE Std 802.1w-2001.

[3] IEEE Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks, IEEE Std 802.1Q-1998.

[4] IEEE Standards for local and metropolitan area networks Virtual Bridged Local Area Networks – Amendment 3: Multiple Spanning Trees Amendment to IEEE Std 802.1Q™, 1998 Edition. IEEE Std 802.1s-2002.

[5] S. Sharma, K. Gopalan, S. Nanda, T. Chiueh, Viking: a multi-spanning-tree Ethernet architecture for metropolitan area and cluster networks, in: Proceedings of IEEE INFOCOM 2004.

[6] S. Varadarajan, T. Chiueh, Automatic fault detection recovery in real time switched Ethernet networks, in: Proceedings of IEEE INFOCOM 1999.

[7] F. De Pellegrini, D. Starobinski, M.G. Karpovsky, L.B. Levitin, Scalable cycle-breaking algorithms for gigabit Ethernet backbones, in: Proceedings IEEE INFOCOM 2004.

[8] T.L. Rodeheffer, C.A. Thekkath, D.C. Anderson, SmartBridge: A scalable bridge architecture, in: Proceedings ACM SIGCOMM 2000.

[9] K. Lui, W.C. Lee, K. Nahrstedt, STAR: A transparent spanning tree bridge protocol with alternate outing, ACM SIGCOMM Computer Communications Review 32 (2002) 3. July.

[10] Y. Lim, H. Yu, S. Das, S.S. Lee, M. Gerla, QoS-aware multiple spanning tree mechanism over a bridged LAN environment, in: Proceedings IEEE GLOBECOM 2003.

[11] OPNET simulator, Available from: <http://www.opnet.com>.

[12] MEF, Metro Ethernet Networks – a technical overview, Available from: <http://www.metroethernetforum.org>.

[13] IEEE Std 802.3z-1998, Gigabit Ethernet, Available from: <http://www.ieee802.org/3/z/index.html>.

[14] Nortel Networks, Service Delivery Technologies for Metro Ethernet Networks, Nortel Networks Whitepaper, Sept. 19, 2003, Available from: <http://www.nortel.com/solutions/optical/collateral/nn-105600-0919-03.pdf>.

[15] Riverstone Networks, Scalability of Ethernet Services Networks, Available from: <http://www.riverstonenet.com/solutions/ethernet_scalability.shtml>.

[16] G. Holland, Carrier Class Metro Networking: the high availability features of Riverstone's RS Metro Routers, Riverstone Networks White Paper #135.

[17] MEF, MEF 5: traffic management specification: phase 1, Available from: <http://www.metroethernetforum.org/TechSpec.htm>.

[18] CISCO, Overview of the Catalyst 8500 Campus Switch Router, Available from: <http://www.cisco.com/univercd/cc/td/doc/product/l3sw/8540/rel_12_0/w5_6f/softcnfg/1cfg8500.pdf>.

[19] MEF, MEF 5: traffic management specification: phase 1, Available from: <http://www.metroethernetforum.org/TechSpec.htm>.

[20] R. Perlman, Rbridges: transparent routing, in: Proceedings IEEE INFOCOM 2004.

[21] R. Garcia, J. Duato, F. Silla, LSOM: a link state protocol over MAC addresses for metro backbones using optical Ethernet switches, IEEE Network Computing & Applications (2003).

[22] C. DeSanti, C. Carlson, R. Nixon, Transmission of IPv6, IPv4, and Address Resolution Protocol (ARP) packets over fibre channel, RFC 4338, Available from: <http://tools.ietf.org/html/rfc4338>.

[23] InfiniBand Trade Association, Available from: <http://www.infinibandta.org>.