A Scalable Hybrid Approach to Switching in Metro Ethernet Networks

Minh Huynh and Prasant Mohapatra
Computer Science Department
University of California at Davis
Davis, CA. USA
{huynh, prasant}@cs.ucdavis.edu

***Abstract*—The most common technology in Local Area Networks is the Ethernet protocol. The continuing evolution of Ethernet has propelled it into the scope of Metropolitan Area Networks. Even though Ethernet is fast and simple, the Spanning Tree in Ethernet is inefficient in terms of network utilization and load balancing. In this work, we compare the performance of Spanning Tree and Link State algorithms in the context of layer 2 switching. In addition, we introduce a hybrid scheme that is customized for Metro Ethernet Networks. The results show that the hybrid scheme increases utilization and reduces the congestion ratio and delay. The performance gained as compared to RSTP, link state, and MSTP are 20.9%, 9.4%, and 11.4%, respectively. In addition, the hybrid scheme is more scalable than using pure link state.**

Keywords-Link state, Metro Ethernet Network,Spanning Tree, routing.

I. INTRODUCTION

Ethernet, the predominant technology in Local Area Networks, has been known for its cost-effectiveness and wide-scale familiarity. The recent standardization of the Gigabit Ethernet [11] protocol has propelled it into consideration for Metropolitan Area Networks (MAN). Metropolitan Ethernet Networks (MENs) [10] are comprised of a metro core network and several access networks. All the access networks connect to the core at one or two aggregation Ethernet switches. The customers' networks are connected to an 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 both in the core as well as the access networks.

The current Ethernet solutions deploy the Spanning Tree Protocol and its variants [1][2][4] to manage the topology autonomously. However, they inefficiently manage the resources in the topology [5][7][8][9][13]. In this work, we compare the performance of the Spanning Tree Protocol and the Link State Protocol using the following metrics: average diameter, congestion ratio, and utilization. We introduce a hybrid approach that is designed for Metro Area Networks using the concepts of Spanning Trees and link state routing to forward frames.

The hybrid approach is able to increase the resource utilization in layer 2 that is traditionally managed by the Spanning Tree protocol. In addition, it reduces the congestion ratio and delay more than the link state protocol and the Spanning Tree protocol. The performance gained by the hybrid scheme over RSTP, link state, and MSTP are 20.9%, 9.4%, and 11.4%, respectively.

The organization of the paper is as follows. The Spanning Tree protocol and the link state protocol will be briefly overviewed. The next section presents a comparison study between the link state protocol and the Spanning Tree protocol as the motivation for the hybrid scheme. It is followed by the introduction of the hybrid approach and its evaluation. Finally, related works are presented before the conclusion of the paper.

II. BACKGROUND

This section presents how the Spanning Tree protocol and the link state protocol operate. The advantages and drawbacks of each protocol motivated the hybrid approach.

A. *Spanning Tree Protocol*

Traditionally, Ethernet-based networks use the Spanning Tree Protocol (STP), standardized in IEEE 802.1d [1], for switching frames in a network. STP constructs a shortest path to the root tree that is overlaid on top of the mesh-oriented Ethernet networks. Primarily, the Spanning Tree (ST) is used to avoid the formation of cycles or loops in the network. STP prevents loops in the network by blocking redundant links. Therefore, the load is concentrated on a single link, making it at risk during failures, and no load balancing mechanism is provided. 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 recompute the ST, which can take 30 to 60 seconds. At any one time, only one Spanning Tree dictates the network.

Although STP has been used for most Ethernet networks, it has several shortcomings in the context of MEN. These shortcomings are enumerated as follows:

1. Low Utilization: Spanning trees restrict the number of ports being used. In high-capacity Ethernet networks, this restriction translates to a very low utilization of the network.

0742-1303/07 $25.00 © 2007 IEEE
DOI 10.1109/LCN.2007.88

2. Bottleneck Links: Traffic on STP concentrates on links surrounding the root.

3. No Load Balance: STP does not have any mechanisms to balance load across the network or to provide backup path.

4. Not Shortest Path: Traffic tends to go to the root resulting in longer path to reach the destination.

An improvement to STP is the Rapid Spanning Tree Protocol RSTP [17] specified in IEEE 802.1w. RSTP reduces the number of port states from five in STP to three: discarding, learning, and forwarding. Through a faster aging time and a rapid transition to the forwarding state, RSTP is able to reduce the convergence time to between 1 and 3 seconds. 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 Spanning Tree 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 [18] is defined in IEEE 802.1s. MSTP uses a common Spanning Tree 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 Spanning Tree. Since MSTP runs pure RSTP as the underlying protocol, it inherits the drawbacks of RSTP. However, a failure in MSTP can be isolated to a single 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 Spanning Tree.

B. *Link State Protool*

Typically at the network layer, the link state protocol routes packets based on the concept of greedy algorithm. Before the link state protocol can begin, it requires the global knowledge of the topology and the all the link costs. An example of a link state protocol is the Dijkstra's algorithm; it computes the least cost path from one node to all other nodes in the topology. After the kth iteration, the least cost paths are known to k destinations. In each iteration, the algorithm uses the least cost path that has not been used yet out of all the known paths of that iteration to find any new least cost path. This procedure repeats until all the nodes have been considered. Despite its popularity, the link state protocol has drawbacks such as:

- Global knowledge of topology is required
- MAC table explosion since the algorithm needs to know the locations of all destinations.
- No backup or alternative path is provided

III. Comparative Performance

The general characteristics of the two protocols are compared in this section using a series of metrics designed to show the shortcomings or advantages of each protocol.

A. *Metrics of Comparison*

First, diameter is used to measure the shortest distance between the two farthest nodes. This distance is expressed in terms of the cost of the path that is between these two nodes. The average diameter of the routing protocol reveals the efficiency of the path in term of least cost calculation. This metric is partially responsible for the propagation delay when routing packet between end points.

Utilization is another metric to evaluate a routing protocol. At the macro level, global utilization tells how well the load is balanced across the network. Intuitively, global utilization is defined as the number of links used over the total number of physical links available. By contrast, at the micro level, the degree of individual nodes is examined. The degree is the number of outgoing or incoming link into a node. The degree of a routing protocol, or port utilization, translates to the ability of the protocol to facilitate the physical links into its routing paths. In the context of layer 2 switching, port utilization is the number of ports that will be used in forwarding traffic. Therefore, the higher the degree of a protocol, the better a node can manage its resources. As a result, the average degree (or local utilization) shows the capability of the protocol to locally load balance the traffic and resource allocation.

The efficiency of a protocol can be measured through the congestion factor. Let the link congestion be defined as the average percentage of a link's bandwidth that is used. The congestion ratio is defined as the link with the highest congestion over the average congestion of the topology.

B. *Spanning Tree Protocol*

In the Spanning Tree Protocol, the diameter of the protocol is the height of the tree because the distance from the root to the lowest leaf node is the shortest distance between two furthest nodes. The placement of the root node affects the efficiency of the routing topology. In the best case scenario where the root is at the center of an $n \text{ x } n$ grid topology, the diameter is $O(diameter) = \log(n)$. Otherwise, if the root is located at the corner of the grid, $O(diameter) = 2(n-1)$.

The global utilization is l_u / l_t where l_u is the number of links used by the protocol and l_t is the total number of links in the topology. By taking the limit as n approaches infinity, we will see that the utilization converges to ½. This means that as the topology gets larger and larger, the Spanning Tree's global utilization cannot be worse than ½. On the other hand, the average degree converges to 2 meaning, at best, a very large topology can have at most 2 ports utilized per node on average.

$$\text{Utilization} = \frac{l_u}{l_t} = \frac{n^2-1}{2n(n-1)} = \frac{n+1}{2n} \approx \lim_{n\to\infty} \frac{n+1}{2n} = \frac{1}{2}$$

The average degree is as followed:

$$\frac{\sum_{i=1}^{nxn} d_i}{total\_nodes} = \frac{2(n^2-1)}{n^2} \approx \lim_{n\to\infty} \frac{2(n^2-1)}{n^2} = 2$$

where d_i is the degree of the routing protocol (port utilization) per node.

C. Link State Protocol

Each node within the Link State Protocol determines its path to all destinations. Therefore depending on the location of the node, the routing diameter ranges from $O(diameter) = \log(n)$ to $O(diameter) = 2(n-1)$.

In a setting where all links in the grid have equal weight, the global utilization is 1. Each source uses all of its adjacent neighbors as the next hop because each node has its own routing path. All links in the network are utilized.

The average degree is as followed:

$$\frac{\sum_{i=1}^{nxn} d_i}{total\_nodes} = \frac{2[2(n^2-1)]}{n^2} \approx \lim_{n\to\infty} \frac{2[2(n^2-1)]}{n^2} = 4$$

D. Comparisons

Table 1 shows that link state is more efficient at allocating resources to the topology in terms of the global utilization and local utilization (or average degree) metrics. The utilization measures how well the protocol can perform load balancing on the topology. High utilization leads to better load balancing. In terms of the diameter, both schemes show the same variance.

TABLE I. COMPARISON BETWEEN SPANNING TREE AND LINK STATE

	Link State	Spanning Tree
Global Utilization	1	1/2
Local Utilization	4	2
Diameter	$\log(n) \to 2(n-1)$	$\log(n) \to 2(n-1)$

E. Validation

A concrete topology is analyzed in this section in order to validate the formulation in Section III. The evaluations are on *n*x*n* grids where n = 3, 4… 10. The following metrics are used: average diameter, average degree, global utilization, and congestion ratio.

Figure 1 shows the average diameter for both protocols including the upper bound and the lower bound. The average diameter affects the path length that is a factor in the delay. Overall, link state protocol provides shorter routing path than the Spanning Tree Protocol. Both protocols grow linearly as the size of the topology increases.

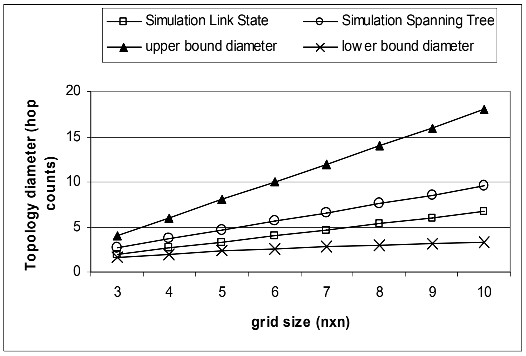

Figure 1. The diameter for grid topologies

For utilization in general, the architecture of the link state protocol enhance its utilization more than the Spanning Tree. Intuitively, link state uses a set of routing tables where each element of the set is a specific routing table for a node in the topology. Each routing table constructs a spanning tree where the root is the source node. On the other hand, Spanning Tree Protocol uses only one routing table from the same set of routing tables that link state uses. The shortest paths from this source to all nodes are stored in this table. Similarly, the Spanning Tree calculates shortest paths from the root to all nodes. Since each routing table is used for a different node, the set of links to make up the shortest path from a single source to all nodes is different from table to table. Therefore, even though there are individual links that are used in multiple tables, each table differs from one another as a whole or partially disjoint.

In the grid topology, each node has 4 physical ports except for the edge nodes. Figure 2 shows that the link state's average degree is closer to the physical restriction of maximum four ports per node, resulting in better utilization. However, Spanning Tree is limited to a degree of 2 which is only half of the number of physical ports. As grid size increases, link state gets closer to the ideal port utilization, and link state's performance is twice that of Spanning Tree.

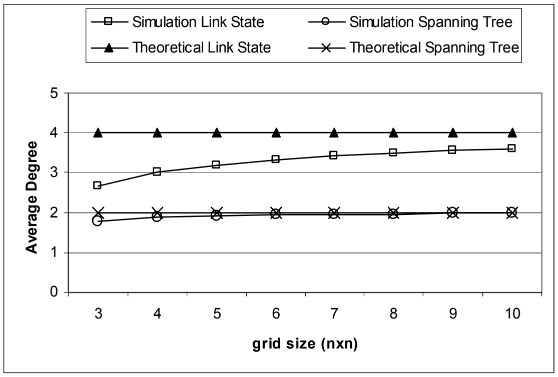

Figure 2. The average degree or local utlization in grid topology

The improvement of the link state protocol is because each route table uses a different root as oppose to Spanning Tree that uses the same root. Let $T_{LS} = \{t_1, t_2 \ldots t_{nxn}\}$ be the set of routing tables for the link state protocol where $t_i = \{l_1, l_2 \ldots l_m\}$ is a routing table for node i consisting of link $l_1, l_2 \ldots l_m$. Let $T_{ST} = \{t_R\}$ be the set of routing table for the Spanning Tree. Suppose each node in the Spanning Tree approach uses t_R for routing where $t_R \in T_{LS}$ and each node in link state approach uses the correspond t_i where i = 1, 2 … nxn. Then the set $\{t_i - t_R\}$ is the set of links that is not used by the Spanning Tree. The union of $\{t_i - t_R\}\ \forall\ t_i \in T_{LS}$ is the additional utilization that link state has over Spanning Tree.

Since the link state protocol uses different root for each route table, based on the shortest path approach, all directly connected links to a root will be used in the routing table for that root. This is true if all links have the same positive cost. Since each node is the root for its own routing table, all links eventually get used as shown in Figure 3 for the link state utilization. Therefore, the link state protocol achieves 100% utilization. On the other hand, Spanning Tree uses only one routing table and applies it to all nodes. Intuitively, Spanning Tree needs n -1 links to connect all the nodes to guarantee loop freedom. Since each node has at least 2 ports on the grid, there are at least 2n links. Therefore, Spanning Tree only utilizes half of the physical link as shown in Figure 3. and the calculation in Section IIIB.

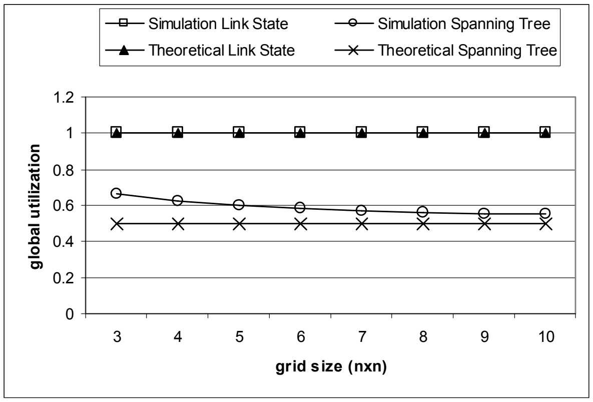

Figure 3. The global utilization in grid topology

The congestion ratio of link state and Spanning Tree are shown in Figure 4. The congestion ratio is defined as the highest congestion over the average congestion. Instinctively, a congestion ratio of 1 indicates that the load is distributed evenly onto each link. As the ratio becomes greater than 1, the load becomes unbalanced. The congestion ratio will never be less than 1. The larger the value away from 1, the more skew the traffic load is inclined toward the node with the highest congestion. In the Spanning Tree case, this node is the root.

As shown in Figure 4, link state and Spanning Tree behave similarly for small size topologies such as 3x3 and 4x4. However, link state's congestion ratio reduces for larger size topologies. On the contrary, Spanning Tree's congestion ratio continues to increase. For small topologies, there are few links so that in order to form a connected graph, both protocols use almost all of the links. Therefore, the number of links that the link state and Spanning Tree use is approximately close. There is little that the link state protocol can take advantage of for load balancing. For example, the 3x3 topology yields 8 links for Spanning Tree and 12 links for link state; while the 4x4 topology yields 15 links for Spanning Tree and 24 links for link state. Thus, the link congestion which is defined as the number of paths going through the link is close between Spanning Tree and link state. As a result, both protocols have similar congestion ratio for small topology size.

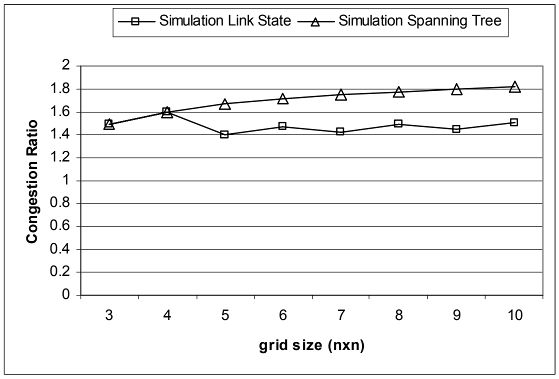

Figure 4. The congestion ratio in grid topology

IV. THE HYBRID APPROACH

The evaluation in Section IIIE shows that in the dense mesh topology, i.e. the metro access network, the link state protocol is more suitable. However, the metro core topology in most cases is a ring [14] which is simpler. The advantages of using the Spanning Tree Protocol to manage a network are low cost and simple to manage. Therefore, it is more suitable for the core in that the core is a haul network where we only need to move trunks of traffic from one end to the other. The ring topology is simple enough for used with STP while link state is more complex than necessary to set up. However, to cope with the drawbacks of STP, we need to redesign some enhancements to customize for Metro Area Network.

A. Metro Core

In the metro core, a protocol that uses multiple Spanning Trees performs frames switching is deployed except that it is customized for the ring structure. Therefore, it still has the advantages of STP such as simplicity, cost effective, rapid provisioning, and flexibility. However, it will be enhanced for resilience to have sub-second re-convergence.

Since a typical metro core structure is a ring, it is more costly than necessary to use Virtual Private LAN Services (VPLS) [16] to manage the core. To create a loop free environment, VPLS deploys the split-horizontal technique that requires a full mesh topology. In the split-horizontal approach, a node would not forward a frame that it had received from one node to any other node. Essentially, each node broadcast to all via direct connections. However, in a ring structure, the metro core does not have this capability; and therefore, it would require additional physical links prior to VPLS deployment.

Similar to Resilient Packet Ring (RPR) [15], however, the hybrid approach is simpler and provides more flexibility to organize classes of traffic by creating multiple domains of forwarding plane called logical ring. Each logical ring can hold a number of virtual LAN or VLAN intended for traffic isolation.

A Multiple Spanning Tree Protocol will manage the metro core. First, a logical ring topology is formed from the physical ring topology by having each node to have a primary port and a secondary port per logical ring. A node will be elected as the root. The election process uses the priority of the node. Each logical ring is managed by an instant of a Spanning Tree where multiple VLANs can be mapped to. Additional logical rings can be added to the metro core, and each logical ring is managed by a different instant of a Spanning Tree. Different combination of primary and secondary port form different logical rings.

For normal operations, data traffic initially will be forwarded on the primary port. Each node blocks all traffic on its secondary port except for control traffic to avoid creating a loop as shown in Figure 5. Forwarding and address learning proceed as in the original STP. The root sends a HELLO control frame at a regular interval from its primary port to be received on the secondary port. The HELLO frame checks the status of the current ring.

There are two ways a fault is detected: through a generated fault detection message sent by the non-root node that detects the fault or by missing polled frame. If a non-root node detects the fault, it sends a fault detection message via its good port toward the root as shown in Figure 5. When the root receives this frame, it goes into the failed state and unblocked the secondary port. Then it flushes the forwarding database and notifies the other nodes to flush theirs and open their secondary port to data traffic. Frames forwarding and address learning are then proceed as before. The second way to detect a failure is through HELLO frame timed out. If the HELLO frame time out before reaching the secondary port of the root node, the root assumes that the ring has problem and goes into the failed state forcing the other nodes to reconverge. The HELLO frame is sent every millisecond. If the ring is fine, this HELLO frame should arrive at the root on the secondary port so that the root can reset the time out. However, if the frame is timed out, the root goes into the failed state and performs the following:

1. Unblock the secondary port
2. Flush forwarding database
3. Force other nodes to flush their forwarding database

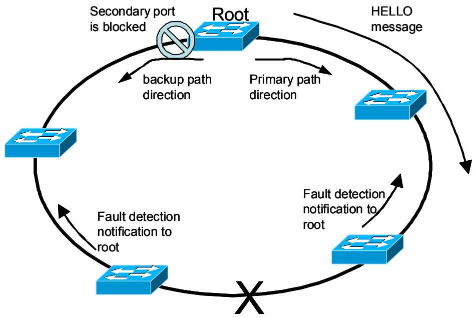

Figure 5. Hybrid protocol operation at the core with fault detection

Once timed out, the root continues to poll on the primary port. If the downlink is restored, the root blocks its secondary port, reopen the primary port, and send flush forwarding database to all other nodes. The non-root nodes forward traffic on the recovered port only when they have received the flush-database message. This is to ensure that no loop is introduced between the time that the link recovers to the time that the root acknowledges the recovery.

Since the secondary port wastes resources in standby mode it can be used as the primary on a different Spanning Tree to maximize the utilization. For example, on node1, port1 is configured as primary, and port2 is configured as secondary for Spanning Tree 1. Port2 idles while the primary path forwarding traffic. To regain the resources, Spanning Tree 2 is configured to have port2 as its primary and port*x* as its secondary where *x* is any available port beside 2.

B. Interface between the core and the access

The operation of the Spanning Tree Protocol is limited to the metro core only, and it is transparent to the link state protocol that manages the access network. Therefore, the metro core tunnels all traffic going through it. The egress switches of the access network that interface with the core switches encapsulate all outgoing frames using the layer 2 header with an additional field for the address of the next hop node before sending the frame into the metro core. The next hop node is next node in the link state protocol in order for the frame to continue on its path to the destination. Because of encapsulation, the control frames of the link state protocol do not interfere with the metro core. Like wise, the control frames for the Spanning Tree stay within the metro core boundary.

C. Metro Access

The typical structure of the metro access network is a mesh topology [14]. Since the link state protocol is better suited for mesh topologies [13], the hybrid approach uses the link state protocol to manage the metro access network. Compared to the distance vector, link state converges faster [13]. Therefore, in the MAN scale, it is more efficient to deploy the link state protocol. As shown in Section III.E, the link state protocol provides higher utilization than the Spanning Tree protocol in both global utilization and local utilization. Hence, link state effectively takes advantages of multiple paths and does not waste resources by letting the link in idle backup mode. Deploying the link state protocol assure that frames travel on shortest path toward the destination.

Compared to VPLS in term of scalability and cost, link state is preferred in the metro access network. The split-horizontal technique used by VPLS has poor scalability. It would need n^2 links for deployment where n is the number of nodes in the access network.

Initially, the access network needs to know the complete map of the topology. To obtain the global map, each node broadcast itself to its direct neighbors through HELLO frames. If 3 consecutive HELLO frames are missed from a neighbor node, then a node assumes that the connectivity with the missing neighbor is down. Once a node knows its neighbors on all the active links, it floods the following information onto the networks:

1. self node id
2. list of direct neighbors with the associated cost

After the broadcast phase, each node computes the shortest path using the link states protocol such as Dijkstra single source to all destination shortest path. If any node detects a change in the topology, the broadcast phase is initiated again for new path computation. During the recomputation, some nodes might have stale routing information that potentially create a temporary loop. However, the hop count parameter prevents the infinite looping of the frames.

V. THE HYBRID SCHEME EVALUATION

The evaluation will compare the performance of four protocols: RSTP, Link State, Hybrid, and MSTP using the metrics mentioned earlier.

A. Simulation Setup

The OPNET [19] simulator tool was chosen because of its comprehensive implementation of Ethernet including the implementation of RSTP, MSTP, and VLAN. The hybrid scheme will be evaluated on a topology representative of the Metro Area Network [14], as shown in Figure 6. The nodes in the core network are **core**{1,2,3,4,5,6}. While the nodes in access networks are **access**{1,2,3,4,5,6} and **aggregator**{1,2}.

RSTP has only a single Spanning Tree (ST) configured on each side of the router. The root of the Spanning Tree is located at the switch **core6**. By contrast, MSTP has 4 Spanning Trees configured: the common root is at **core6** but the regional root for ST 1 and ST 4 is at **core1**, the regional root for ST 2 and ST 3 is at **core2**. In this analysis, only 1 logical ring is considered in the metro core.

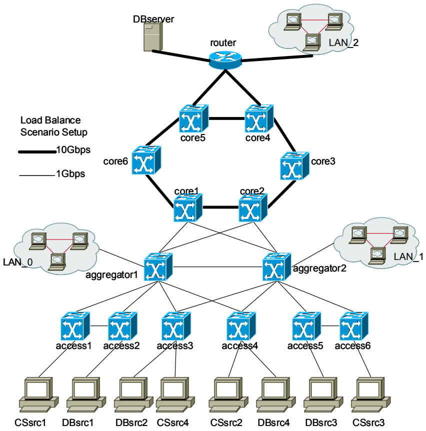

Figure 6. A representative of the metro topology.

The main traffic flows are from CSsrc{1,2,3,4} and DBsrc{1,2,3,4} to DBserver with each flow is a video conferencing sessions that starts after 100s has elapsed, thus allowing the initialization to complete for all protocols. Each source sends 25 flows producing 106.25Mbps per source. The background traffic from LAN_0 to LAN_2 and LAN_1 to LAN_2 send a total of 3Gbps. Its main purpose is to congest the bottleneck links at the aggregator section of the network. These are the link interconnect **aggregator1**, **aggregator2**, **core1**, and **core2**. The link capacities are shown in Figure 6. The remaining background traffic are randomly selected between LAN_0, LAN_1, CSsrc{1,2,3,4}, and DBsrc{1,2,3,4}. Each random traffic flow represents a voice quality application demand.

B. Utilization

Utilization is presented first to illustrate the load balancing capability of different protocols. As mentioned earlier, load balancing is a critical feature to desire for in Metro Ethernet Network. To illustrate how each protocol handles this capability, the utilization histogram is presented in Figure 7 along with the actual values in TABLE II. Since the background traffic from LAN_0 to LAN_2 and LAN_1 to LAN_2 each takes up 1.5Gbps, they fill up two links to the maximum capacity. Hence, all four protocols have 4.54% of their topology 100% congested. Similar to the analytical section, the Spanning Tree protocols leave many links unused. As expected, RSTP do not have any load balancing mechanism, thus, it uses about half of the links in the topology to transport traffic. MSTP is able to create multiple Spanning Trees and distributes traffic on them; therefore, its utilization is higher than RSTP but more than a third of the link resources idle still. In contrast, link state and hybrid leave few links idling and the majority of the links are occupied. They distribute the traffics well enough that most links are occupied around 25%. Since link state has the whole view of the metro topology, its routing utilize resources more efficient than hybrid. However, as the topology becomes larger, link state encounters scalability problems such as address distribution, longer convergence time, and stale topology information. The delay to get the topology information from the two farthest nodes creates an instable routing table. On the other hand, the hybrid scheme only needs to converge within an access network. In effect, routing information converges faster and more reliable. The interface nodes between the core and the access handle the inter-access routing. The trade off is shown in the utilization between pure link state and hybrid.

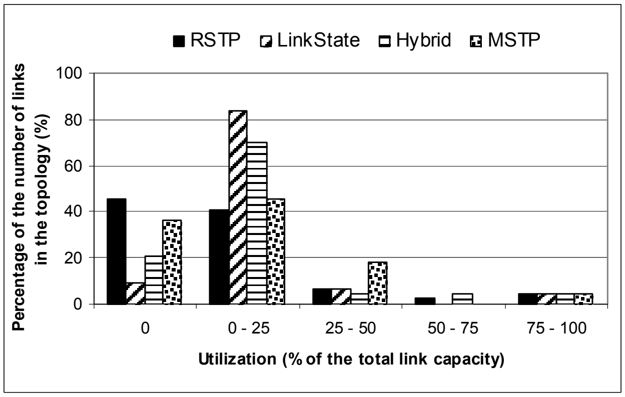

Figure 7. The utilization histogram for the 4 protocols

TABLE II. THE UTILIZTION HISTOGRAM

Link Capacity	RSTP	Link State	Hybrid	MSTP
0%	45.45%	9.09%	20.45%	36.36%

0% to 25%	40.90%	84.09%	70.45%	45.45%
25% to 50%	6.8%	6.8%	4.5%	18.18%
50% to 75%	2.7%	0%	4.5%	0%
75% to 100%	4.54%	4.54%	4.54%	4.54%

C. Path Length

The average diameter metric is measured through the path length of the routing protocol. The path length is expressed in term of the average number of hops as shown in Figure 8. Overall, both the link state and the hybrid scheme choose the shortest path. Therefore, the length for each of the paths is shorter than RSTP and MSTP. On the contrary, the Spanning Tree protocol and its variances create the path toward the root leaving some paths to the destination longer than the shortest path. The effect of this behavior is more apparent in MSTP where multiple Spanning Trees are created. The resulting paths take longer detour than necessary. Therefore, MSTP's average hops from all sources to all destinations are higher than RSTP.

The path length affects the propagation delay that contributes to the end-to-end delay. The end-to-end delay is shown in Figure 9. The significantly high end-to-end delay for RSTP and MSTP is not just the result of longer path length but also dues to the lower utilization. Since few links are utilized to transport traffic, when congestion occurs (as it did when LAN_0 and LAN_1 send to LAN_2), the currently active links are maxed out forcing the frames to queue up. As for link state, the control packets from all nodes to every other nodes across the topology add to the processing delay that contributes to the higher end-to-end delay. Although, the delay stabilizes at 0.5s, this delay is unacceptable for multimedia real time application [20]. In contrast, being more scalable than link state, the hybrid scheme is able to maintain the end-to-end delay at 0.08s.

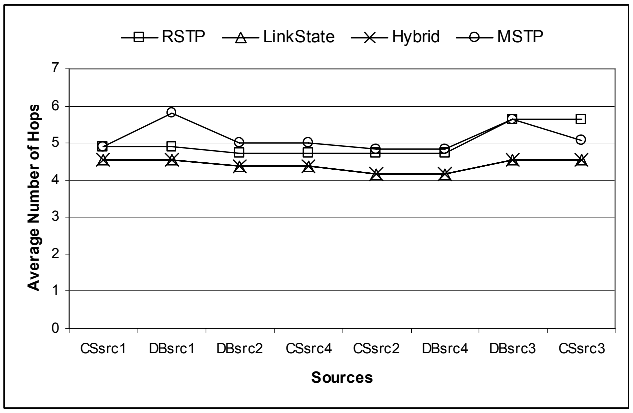

Figure 8. The average number of hops from different sources

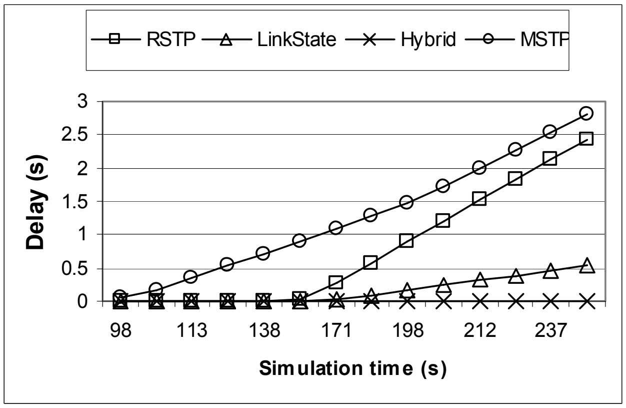

Figure 9. The end-to-end delay for the video conference application

D. Congestion Ratio

The congestion ratios for four protocols are presented in Figure 10. Although the link state protocol achieves higher utilization than the hybrid scheme, the hybrid has lower congestion ratio indicating better load distribution. The difference of 0.46 between the two schemes is small as compared to the performance gain over RSTP. This small gain is the result of the separation of the core from the access network in routing. There is little redundancy in the metro core that the link state protocol can take advantages of. However, once the metro core operates separately from the access networks, the congestion ration is reduced because the bottleneck links are moved away from the core.

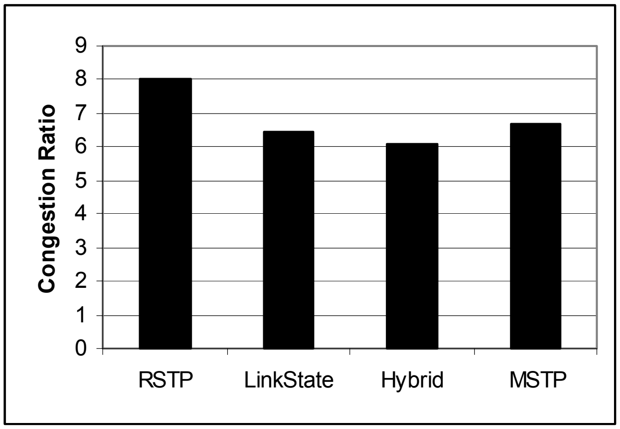

Figure 10. The Congestion Ratio among the 4 protocols

E. Throughput

Finally, the performance of the video conference application is presented to illustrate the result of the advantages of the hybrid scheme. The collected throughput received at the end hosts are shown in Figure 11. With the advantages discussed earlier, the hybrid scheme achieved the higher throughput. At 110s when the heavy background traffic from LAN_0 and LAN_1 to LAN_2 started, hybrid was able to keep the same consistent throughput. In contrast, the other three protocols were affected by the congestion; thus, their throughput reduced for the remaining of the simulation. RSTP takes the most severe performance hit. Compared to the hybrid scheme, hybrid performance gained over link state protocol, RSTP, and MSTP were 9.4%, 20.9%, and 11.4% respectively.

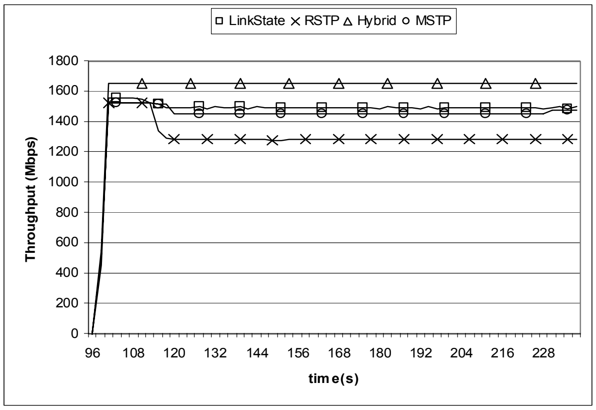

Figure 11. Throughtput receive from the main traffic flows

VI. RELATED WORKS

Viking is a Multiple Spanning Tree architecture [5] that precomputes multiple Spanning Trees so that it can change to a backup ST in the event of a failure. Viking's complexity lies in the computation of the *k* shortest primary paths and the *k* backup paths for each primary paths. A path aggregation algorithm is then run to merge the paths into the Spanning Tree.

Ethereal [6], a real time connection oriented architecture supporting best effort and assured service traffic at the link layer, proposes to use the propagation order Spanning Tree for faster re-convergence of the ST once a failure has been detected. SmartBridge [8] and STAR [9] are also two other approaches that find an alternate route that is shorter than the corresponding path on the Spanning Tree. SmartBridge requires the full knowledge of the topology. STAR is an overlay-based approach that calculates the shortest path from one overlay node to the next using the distance vector.

Another approach to load balancing is Tree-Based Turn-Prohibition (TBTP) [7]. TBTP constructs a less restrictive Spanning Tree by blocking a small number of pairs of links around nodes, called turn, so that all cycles in a network can be broken.

Instead of taking the Spanning Tree approach, Rbridges [12] and LSOM [13] run link state protocol over the topology. Both protocols broadcast addresses to obtain the global view of the topology. Rbridges proposes to use a link state protocol similar to the one in IS-IS. 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.

VII. CONCLUSION

In this paper, we compared the performance of link state protocol and the Spanning Tree protocol for a grid topology and a representative metro topology. Motivated by the advantages and shortcomings of both protocols, we proposed a hybrid approach for switching frames in layer 2 in the Metro Area Network. The hybrid approach was able to reduce the congestion ratio and delay as compared to the link state and Spanning Tree. Considered the size of Metro Area Network, the hybrid scheme scalability is suited better.

REFERENCES

[1] IEEE Media Access Control (MAC) Bridges, ISO/IEC 15802-3, ANSI/IEEE Std 802.1D, 1998.

[2] IEEE 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 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" Proceedings of IEEE INFOCOM 2004.

[6] S. Varadarajan, T. Chiueh "Automatic Fault Detection and Recovery in Real Time Switched Ethernet Networks" Proceedings of IEEE INFOCOM 1999.

[7] F. De Pellegrini, D. Starobinski, M. G. Karpovsky, and L. B. Levitin. "Scalable Cycle-Breaking Algorithms for Gigabit Ethernet Backbones" Proceedings IEEE INFOCOM 2004

[8] T. L. Rodeheffer, C. A. Thekkath, D. C. Anderson. "SmartBridge: A Scalable Bridge Architecture" Proceedings ACM SIGCOMM 2000

[9] K. Lui, W. C. Lee, K. Nahrstedt. "STAR: A Transparent Spanning Tree Bridge Protocol with Alternate Routing" ACM SIGCOMM Computer Communications Review Volume 32, Number 3: July 2002.

[10] MEF, "Metro Ethernet Networks – A Technical Overview" http://www.metroethernetforum.org

[11] IEEE Std 802.3z-1998, Gigabit Ethernet, http://www.ieee802.org/3/z/index.html

[12] R. Perlman "Rbridges: Transparent Routing" Proceedings IEEE INFOCOM 2004

[13] 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.

[14] G. Holland. "Carrier Class Metro Networking: The High Availability Features of Riverstone's RS Metro Routers" Riverstone Networks White Paper #135

[15] F. Davik, M. Yilmaz, S. Gjessing, N. Uzun, "IEEE 802.17 Resilient Packet Ring Tutorial" IEEE Communications Magazine. March 2004

[16] VPLS.ORG "Virtual Private LAN Services (VPLS) Technical Overview." http://vpls.org/vpls_technical_overview.shtml

[17] 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

[18] 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

[19] OPNET simulator. www.opnet.com

[20] Cisco. "Synopsis of Basic VoIP Concepts" – Catalyst 4224 Access Gateway Switch Software Configuration Guide.