Optimized QoS Protection of Ethernet Trees

Tibor Cinkler, András Kern and István Moldován
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Magyar tudósok krt. 2, Budapest, Hungary H-1117
email: {cinkler, kern, moldovan}@tmit.bme.hu

Abstract—Ethernet is being increasingly employed in metro networks. Ethernet gives routing schemes and protocols for building up trees, for instance STP, RSTP and MSTP. It also implies simple restoration mechanisms.

In this paper we propose an optimization framework, where the Ethernet MSTP trees are protected and QoS is guaranteed even after a failure. The optimization is based not only on the topology, but it also takes traffic conditions and QoS constraints into account.

The numerical results show that the proposed optimization significantly increases the throughput of the network. The best result can be achieved when preemption is assumed, i.e., when the best effort traffic may remain unprotected, but not the high priority one. This way high throughput can be realized at normal operation, while it still protects prioritized traffic in case of a failure.

Furthermore, protection mechanisms act faster than the standard restoration mechanism resulting shorter out-of-service times, and therefore higher availability.

Keywords: TE, Resilience, Protection, QoS, Optimization, Ethernet STP, MSTP

I. INTRODUCTION

Today, Ethernet is the dominant technology in the home and enterprise networks. It is also emerging as a cost efficient technology in the metro and regional networking arena. Therefore, the introduction of Ethernet technology both, in the first mile and in the aggregation network would provide a consistent networking technology which does not require any protocol translation, with all the above mentioned advantages in the SOHO network, in the local loop, in the aggregation network as well as in the regional network. However, to introduce Ethernet technology in the Metro environment it must fulfill carrier-grade service requirements. Carrier grade Ethernet Services should have the following features:

- Traffic control and user traffic separation
- Operation and Management features
- Quality of Service support
- High Availability: minimal down time, protection of critical paths
- Scalability not only in terms of bandwidth, but in terms of number of VLANs etc.
- Security

The 802.1Q [1] standard extends Ethernet to support QoS classes and also provides means for traffic separation introducing Virtual LANs (VLANs). Defining VLANs for users makes possible the user traffic separation. The VLAN tag holds the priority bits required for traffic prioritization often mentioned as 'p-bits', and a VLAN identifier. The 12 bit VLAN tag field allows only 4096 possible VLANs within a network. To obtain more flexible and more scalable use of VLAN-IDs, the IEEE is now standardizing stacked VLANs in draft standard 802.1ad [5]. This allows the tagging of Ethernet frames with two VLAN tags: the original VLAN tag, now called customer VLAN (C-VLAN) tag and a completely new service VLAN (S-VLAN) tag.

Thus, the 802.1Q Virtual LAN and the 802.1ad Provider Bridges standards give a solution for QoS, traffic separation and VLAN scalability. However the carrier grade traffic control and protection (resilience) requirements are not met with this solution. 802.1s MSTP [3] provides a way for traffic control by using multiple trees.

Using the means available we provide a management solution for carrier grade Ethernet services with QoS, protection and Traffic Engineering.

The paper is structured as follows: We give an overview of the technologies available and propose an new architecture that deals with both QoS and high availability in Section II. Section III gives a formal model for the architecture and an Integer Linear Program formulation. Here we propose alternative optimization scenarios as well. In Section IV these methods are evaluated by extensive simulations. Finally in Section V we conclude the paper.

II. TECHNOLOGICAL BACKGROUND

A robust, reliable network has to transfer traffic efficiently, provide redundancy, and be recovered quickly from faults. Carrier grade availability requires a network availability of 5 nines (99.999%) over one year. This means, that in 99,999% of 365 days the network is up and running. This gives 5.25 minutes per year for unplanned downtime. The table below shows the total downtime threshold for given availabilities:

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>AVAILABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>99.9</td>
</tr>
<tr>
<td>Downtime (minutes)</td>
<td>525.6</td>
</tr>
</tbody>
</table>

This work has been done within the research co-operation framework of three European 6FP IST projects: NoE e-PhotonOne (http://www.e-photon-one.org/), IP MUSE (http://www.ist-muse.org/) and IP NOBEL (http://www.ist-nobel.org/).
Up to 4 nines availability can be achieved by local supervision with 24 hours/day, 7 days/week without any resilience mechanism. 5 nines availability can only be achieved with resilience (protection or restoration). With restoration we can achieve failover times in range of minutes or tens of seconds, however lower restoration times can only be achieved by protection.

The end-to-end protection may involve different mechanisms and it is accomplished by many transports like Ethernet: Ethernet over DSL or SDH/SONET, IP/MPLS, etc. In a Layer 3 environment, routing protocols can track redundant routes to the destination, quickly identify a secondary path if the primary path fails, and share the traffic load across multiple paths. Unfortunately, routing protocols are not available in an Ethernet Network.

A. Resilience in Ethernet

The Spanning Tree Protocol [1] is the basic bridge protocol developed originally to avoid loops in the bridged network. The protocol builds a tree that spans all bridges in the access network and data is propagated along this tree. The links that are not part of the tree are blocked. In the case of a failure, the blocked links are activated providing a self-healing restoration mechanism for the bridged network. The main drawbacks of this algorithm are the long recovery times and the inability to use the backup links for forwarding until failures.

The Rapid Spanning Tree protocol [2][9] is based on STP, but it has major enhancements. It addresses the main problems of STP which reside in its slow speed, RSTP uses proposal agreement based handshaking mechanism to restore the connectivity in case of failures. After a port receives the agreement in reply of a proposal sent earlier, it immediately changes to forwarding state. This handshake propagates quickly toward the edges of the network and restores connectivity after a topology change.

RSTP lacks the fault-management capabilities that service providers get with SDH/SONET and is not designed for ring topologies, which continue to be widely deployed in metro networks. The technology is still not viable in a ring environment and often it is hard to deploy it in a mesh network.

That is why the IEEE is defining another standard, the 802.17 Resilient Packet Ring (RPR) [7] for enhancing resilience of Ethernet networks. Just as its name implies, RPR is designed for ring topologies carrying packets, but with the same resilience attributes as of a typical SDH/SONET ring.

The Ethernet Automatic Protection Switching (EAPS) [6] technology increases the availability and robustness of Ethernet rings. An Ethernet ring built using EAPS can have availability comparable to that provided by SDH/SONET rings, at a lower cost and with fewer constraints (e.g., on ring size). This technology works well in ring topologies for MANs or LANs. It converges often in less than 50 milliseconds. It does not limit the number of nodes in the ring, and the convergence time is independent of the number of nodes in the ring.

Link Aggregation or trunking (IEEE802.3ad) [4] is a method of combining physical network links into a single logical link for increased bandwidth. Several parallel physical links between two devices are grouped together to form a single logical link. Link Aggregation provides the following important benefits: Higher link availability - link aggregation prevents the failure of any single component link from leading to a disruption of the communications between the interconnected devices. Increased link capacity - the capacity of an aggregated link is higher than any of its individual links alone.

B. Spanning Tree Protocol convergence

The operation of the STP protocol relies only on timers, and in case of a failure the following timer values must be considered: 20 sec Hello (or failure detection) time, 15 sec Forward delay, 15 sec Learning time. The failure is detected by the loss of three consecutive Hello messages. Hello messages are sent by default every 5s, thus a failure is detected at most in 20s due to timers. Then, the bridge reacts to the topology change and starts flooding topology information using BPDU's (Bridge Protocol Data Units). The flooding procedure ends when the Forward delay timer expires. The default value of the Forward delay timer is 15s. The alternate port or ports, which have been activated, require a 15s learning time in order to learn the MAC addresses. After that the data stream starts to flow again. These steps lead to a potential delay of 50s in case of a failure, which is too high for carrier grade services.

RSTP introduces new port states in order to decrease the convergence time in case of a failure: the Backup Port and Alternate Port states. The Alternate Port offers an alternate path in the direction of the Root Bridge to that provided by the Root Port of the Bridge, whereas a Backup Port acts as a backup for the path in the direction of the leaves of the Spanning Tree. When a direct uplink fails, RSTP unblocks the highest priority Alternate Port and begins forwarding traffic without going through the Spanning Tree listening and learning states. In case of a topology change, RSTP uses a proposal-agreement based method instead of timer based operation. The recovery process involves propagating topology change information. Switches with better root path cost will send a proposed BPDU (proposal flag is set), which will be propagated back to the requesting switch. As soon as all the switches agree on the new topology (replying BPDU with Agreement flag set), the network is recovered. Thus, the recovery time in case of topology change depends on the topology and the processing time of RSTP BPDU's. Bridges may have different BPDU processing times. The IEEE 802.1w standard specifies only a maximum value of 1s for this parameter, however several switch implementations may have a processing time as low as a few milliseconds, resulting in restoration times in range of seconds.

The shortcomings of the spanning tree protocols have been identified in several papers too. STAIR [12] proposes a novel bridge protocol that attempts to find and forward frames over alternate paths that are shorter than their corresponding tree paths on the standard spanning tree, and makes use of the standard spanning tree for default forwarding. An automated way to configure the network, taking into consideration the QoS factors is described in [11], where a simple and effective
enhancement to the Multiple Spanning Tree protocol is proposed to achieve high degree of QoS by keeping in perspective the different characteristics of the various traffic types. Several resilience methods have been developed exploiting the new possibilities provided by the MSTP protocol. Viking [10] uses multiple spanning trees to select two different switching paths between end nodes, and in case of a failure it switches to the backup path changing to the other spanning tree instance by changing the VLAN tag.

C. Architectural overview

The proposed network architecture reflects the typical Ethernet based aggregation topologies, however we suppose that there is enough resilience capability in the network for protection. Just like in current networks, all traffic goes to the network edges for security and billing reasons. However, instead of PPoE tunnels we use VLANs for traffic separation and user identification.

The architecture adopted provides 4 QoS classes based on the 802.1p priorities. We suppose priority queuing in the bridges, therefore careful dimensioning is required to avoid starvation of lower priority classes. In order to provide the required QoS constraints and avoid the degradation of services, for each link predefined traffic percentages are admitted (Table II). Admission control is only done at network edges, and policing ensures that for each flow the traffic does not exceed the allowed percentages. In case of STP due to the tree topology, it is ensured that the traffic percentage ratios are kept since all traffic follows the same path towards the root. If the network is correctly dimensioned and there are no bottlenecks to the root, the QoS constraints are satisfied. MSTP can be used for traffic engineering (TE), and it can use different paths for different VLANs. In case of MSTP we must take care that two trees sharing the same link do not carry more prioritized traffic than allowed. Thus, not only dimensioning is crucial but optimal path selection is required as well.

The protection method is similar to the VLAN switching described in [10]. However, in our case the optimization step takes care of selecting the appropriate protection paths ensuring disjoint trees on resilient topologies.

D. MSTP based protection switching

The Multiple Spanning Tree Protocol (MSTP) [3] is a modified, improved version of RSTP. MSTP improves RSTP scalability by aggregating a number of VLAN-based spanning trees into distinct instances, and by running only one (rapid) spanning tree per instance. It introduces two major improvements: divides the network into regions and several VLAN-based tree instances may be present on these regions. A region represents a group of switches that have similar configuration. Inside an MST region, there may be several MST instances. The regions are connected through a central instance of the spanning tree protocol. The MSTP provides a maximum of 64 spanning tree instances in the region. These MST instances may have different roots and may follow different paths in the network. Each VLAN in the network may be part of exactly one Spanning Tree Instance. The desired trees can be achieved by setting appropriately the bridge and port priorities and port costs for the different MST instances.

MSTP can be used for traffic engineering (e.g. load balancing) but also to set up protection paths in the network. The basic idea behind the protection mechanism is to set up two disjoint trees between the root(s) and leave(s). For the protected demands we assign 2 VLANs, one for each tree. Initially, the traffic is tagged with the first VLAN tag, and the traffic follows the first tree instance. In case of a failure, the border nodes will be notified about the failure, and the traffic will be switched to the protection path by tagging the traffic with the second VLAN tag.

Of course, this solution requires intelligence at the border nodes, and the restoration time will highly depend on the failure detection and notification method. Several detection and notification methods are available:

- Link-layer failure detection with alarm indication: times can vary widely depending on the physical media and the Layer 2 encapsulation used. In case of failure Alarm Indication Signals are used to notify the edges. This approach is adopted by IETF in the draft standard 802.1ag. SNMP traps can also be used for notification as in [10].
- Depending on the upper layer technology, L2TP defect detection mechanisms, MPLS LSP tunnel connectivity verification or RSVP-TE hello protocol can provide fault notification.
- Bidirectional Forwarding Detection (BFD) [8]: BFD can provide fast failure detection times for all media types, encapsulations, topologies, and routing protocols. In the best-case scenario, it can provide fast failure detection similar to that found in SDH/SONET.

Although detection and notification play a very important role in protection switching, the specification of the method is out of scope of this paper.

When a failure is detected, the protected traffic will be rerouted to the protection path. The switching is triggered by the failure notification. The original tree will heal itself by the MSTP, and possibly the resulting new tree will not be disjoint. Thus, it is vital to ensure that no traffic is directed to this tree. However, it is enough to switch only the traffic affected by the failure.

In our scenario we assume that the traffic is only between leaves and roots, and a leaf can only communicate with another leaf through a root. This reflects the current situation when all traffic is flowing through the edge nodes. Thus, it is enough to switch the VLANs for the leaves on the branch affected by the failure.

The proposed protection method has the following advantages:

- Runs on any topology, i.e. it does not require ring
- The protection switching is very fast, it depends only on the speed of failure detection
- It can provide many types of protection: 1+1, 1:1, shared etc.

When combined with an efficient Connectivity Fault Management solution it can provide carrier grade service protection on any resilient network topology.
Depending on the protection method, we can protect:

- 1:1 all traffic,
- 1:1 QoS traffic,
- 1:1 only selected traffic, and
- all traffic with sharing the spare capacity among them (Shared Protection).

In case of Metro Ethernet shared protection or protection of some paths (e.g. QoS) is suitable. To select the optimal method it is needed to know the availability requirement of the network. High availability can be provided by the restoration mechanism of RSTP protocol when the network topology is resilient, i.e., the topology is two-connected and has enough capacity to guarantee two disjoint paths for each protected demand. However, higher availability with seamless transition and QoS guarantees can only be provided by protection.

In our paper we present an optimal solution for network design to provide QoS and high availability using MSTP based traffic engineering and protection switching. Good results can be achieved using RSTP for restoration, however, failover times will be in the range of seconds and there will be no QoS guarantees for rerouted traffic. Our solution provides high speed protection switching and QoS guarantees for protected traffic even in case of a failure. Further advantage of our algorithm is that it does not require ring topology; therefore, it can provide protection over any kind of resilient topology.

III. PROBLEM FORMULATION

Here we discuss the formal model of the architecture and propose a Traffic Engineering method to build up working and backup trees and assign VLANs to them in order to ensure the required level of QoS and availability. After an optimization process weights are assigned to the links of the network: equal low link weights to edges to be used by the considered tree while equal high weights to edges not to be used by the considered tree.

A. Formal Graph Model

We consider the network as a directed graph \( G(N, L, B) \) where \( N \) is the set of nodes (vertices) where the switches are placed, \( L \) is the set of links (edges) connecting the nodes, and let \( B \) be the bandwidth (capacity) of these links. The set \( N \) consists of three disjoint subsets: set of the access nodes \( AN \), the set of the edge nodes \( EN \) and the set of the bridge nodes \( BN \). We consider that the point to point links connecting the bridges are full-duplex, thus, each physical link is given as two anti-parallel graph edges having the same bandwidth.

The traffic of different users are differentiated at the Access Nodes by C-VLANs. The aggregated traffic of one or more C-VLANs with the same destination form a demand at the access node. Thus, each demand \( (o \in O) \) is assumed to originate from the access nodes \( s(o) \in AN \) and their destinations are the edge nodes \( d(o) \in EN \) that connect the access network to the backbone. The roots of the trees are also placed here. Without loss of generality, we made two assumptions. First, the demands are connected to each other only through the edge nodes. Direct connections between the access nodes can be enabled by private trees due to management decisions. Second, the demands are directed from the access to the edge nodes. We need this to deal with the tree constraints but we allocate symmetrical capacity \( (b^o) \) in both directions.

The demands belong to different traffic (or QoS) classes \( c \in C \) having different requirements. Since there are 3 bits in the Ethernet header for this purpose, up to \(|C| = 8\) traffic classes can be differentiated. For simplicity reasons we use only four classes as shown in Table II. We assign the highest weight \((w^1)\) in our case) to the highest priority traffic class, however, it will have the lowest volume, therefore the lowest ratio of link resources \((\beta^1 = 10\%) \) is assigned to it. Each demand \( o \in O \) belongs to one traffic class \( c \in C \), therefore, set \( O \) consists of \(|C|\) subsets \( O = O^1, O^2, \ldots, O^{|C|} \).

<p>| Table II THE TRAFFIC (QoS) CLASSES |</p>
<table>
<thead>
<tr>
<th>Number</th>
<th>Name</th>
<th>Resource Ratio</th>
<th>Priority Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Platinum</td>
<td>( \beta^1 = 10% )</td>
<td>( w^1 = 4 )</td>
</tr>
<tr>
<td>2</td>
<td>Gold</td>
<td>( \beta^2 = 20% )</td>
<td>( w^2 = 3 )</td>
</tr>
<tr>
<td>3</td>
<td>Silver</td>
<td>( \beta^3 = 30% )</td>
<td>( w^3 = 2 )</td>
</tr>
<tr>
<td>4</td>
<td>Best Effort</td>
<td>( \beta^4 = 40% )</td>
<td>( w^4 = 1 )</td>
</tr>
</tbody>
</table>

The trees in an STP or MSTP based Ethernet network define how to route traffic. There can be up to 64 trees \( t \in T \) \(|T| \leq 64\) in an MST region according to the IEEE 802.1s standard.

Also for simplicity reasons, the roots of the trees reside at the edge nodes. Of course it is possible to assign more than one tree to one edge node. This enables to realize either traffic engineering or protection (see Section II-D for details). All nodes that are neither roots (edge nodes) nor leaves (access nodes) are referred to as bridge nodes.

Originally, the demands (aggregated traffic of C-VLANs) are tagged with Service VLAN tags (S-VLANs) and carried to the edges. However, in this paper we define two S-VLANs instead of one and each C-VLAN can be assigned to both S-VLANs. This attributes the protection switching to a dynamically changeable VLAN assignment, i.e., if the working S-VLAN becomes unavailable then all C-VLANs are reassigned to the proper backup S-VLAN. S-VLANs from different access nodes having demands with the same destination form a tree. Thus, let \( t_w(o) \subset T \) and \( t_p(o) \subset T \) be the available working and protection trees respectively for demand \( o \in O \). We are optimizing S-VLANs; henceforth, we refer to the S-VLANs simply as VLANs.

B. ILP Formulation

Here we give the Integer Linear Programming (ILP) formulation of the optimal 1:1 protected tree design. First, the not yet presented parameters and variables are defined. Then the objective and the constraints are explained.

a) Variables: The input parameters are: the network topology \( G(N, L, B) \) the sets of the demands \( O \), the trees \( T \) and the demand classes \( C \).

The solution consists of the paths of the C-VLANs, and of the working and protection trees. All parts are described with binary indicator variables.
\( x_i^o \) is 1 if and only if demand \( o \in O \) uses link \( l \in L \) as a part of the working tree.

\( p_i^o \) is 1 if and only if demand \( o \in O \) uses link \( l \in L \) as a part of the protection tree.

\( y_t^i \) is 1 if and only if tree \( t \in T \) uses link \( l \in L \).

b) The objective (1): is to minimize the weighted sum of the used resources for all links in the network. The weight parameter \( \alpha \), where \( 0 < \alpha \leq 1 \) is used to minimize the number of links used by a tree (for larger \( \alpha \)) or to minimize the total capacity used (for smaller \( \alpha \)), while \( w^o \) is used to make even shorter paths for demands of higher priority.

\[
\min \sum_{(i,j) \in E} \left[ \alpha \sum_{l \in T} y_{ij}^l + \frac{1-\alpha}{B_{ij}} \sum_{s \in O, v \in T} w^o(x_{ij}^o + p_{ij}^o) b^o \right].
\]

Subject to:

c) Capacity Constraints (\( \forall (i, j) \in E \)): We have similar capacity constraints for each class (inequalities 2–4) guarantee that each class uses no more capacity than assigned to its class. However, the best effort (BE) traffic may use any resources that are not used by higher classes, therefore we sum up all the demands including BE traffic and force them to fit into the link capacity (5):

\[
\sum_{s \in O^1, v \in T} (x_{ij}^o + x_{ji}^o + p_{ij}^o + p_{ji}^o) b^o \leq \beta^1 \cdot B_{ij}
\]

\[
\sum_{s \in O^2, v \in T} (x_{ij}^o + x_{ji}^o + p_{ij}^o + p_{ji}^o) b^o \leq \beta^2 \cdot B_{ij}
\]

\[
\sum_{s \in O^3, v \in T} (x_{ij}^o + x_{ji}^o + p_{ij}^o + p_{ji}^o) b^o \leq \beta^3 \cdot B_{ij}
\]

\[
\sum_{s \in O^4, v \in T} (x_{ij}^o + x_{ji}^o + p_{ij}^o + p_{ji}^o) b^o \leq B_{ij}.
\]

d) Flow Conservation (\( \forall o \in O, \forall i \in N \)): The flow conservation constraints guarantee that a demand goes from its source to its destination, and if it enters any other node it has to leave it as well. They are given for both working (Equation 6) and backup (Equation 7) paths of the demands. Parameter \( \pi_o \) deals with demand-differentiated protection in such manner that \( \pi_o \) is set to 1 if and only if demand \( o \in O \) should be protected. Otherwise \( \pi_o = 0 \) meaning the equations with \( \pi_o = 0 \) are left out. The default value of \( \pi_o \) is 1, for \( \forall o \in O \). The constraints are:

\[
\sum_{j \in N^+} x_{ij}^o - \sum_{k \in N^+} x_{ki}^o = \begin{cases} 
1 & \text{if } i = s(o) \\
-1 & \text{if } i = t(o) \\
0 & \text{otherwise} \\
\end{cases} \\
\forall i \in N \text{ and } \forall o \in O
\]

and

\[
\sum_{j \in N^+} p_{ij}^o - \sum_{k \in N^+} p_{ki}^o = \begin{cases} 
\pi_o & \text{if } i = s(o) \\
-\pi_o & \text{if } i = t(o) \\
0 & \text{otherwise} \\
\end{cases} \\
\forall i \in N \text{ and } \forall o \in O
\]
- **Multiple Spanning Tree Protocol (MSTP):** Here we assume that multiple spanning tree instances (MSTIs) exist and exactly one MSTI is assigned to each edge node. One MSTI per edge node is enough since if more MSTIs run in the same node and all of them use the same default configuration the spanned trees will be the same.

- **Optimized Multiple Spanning Tree Protocol (MSTPopt):** In this case one tree is assigned to each edge node as for the MSTP case, however, the trees are optimized. MSTPopt is based on a simplified version of the ILP model presented above without protection: \( \pi_o \) is set to 0 for each demand. While STP and MSTP are based on topology only, MSTPopt takes into account the traffic and QoS requirements as well.

In order to provide protection for the two upgraded versions of MSTPopt we use two trees per edge node instead of one:

- **MSTPopt with 1:1 protection of all traffic (FullProt):** Here we use the traditional 1:1 protection that defines two link disjoint paths for each demand and dedicates the required bandwidth to both paths. Obviously, the two paths use the two disjoint trees. Although it wastes the network resources, we can apply it easily. The spare capacity can be decreased if not all the demands are protected, but only the distinguished higher priority demands.

- **MSTPopt with 1:1 protection of QoS traffic (QoSProt):** This method implements the spare capacity decrease idea that protects only several demands. For the QoS traffic 1:1 protection is defined, while the BE traffic can use the spare capacity allocated for the prioritized demands. However, in the case of a link failure the QoS traffic crowd out the BE from the backup link. Obviously, if the link transmitting the BE traffic fails there is no backup path defined, thus it will completely be lost.

### B. Investigated Topologies

![Fig. 1. Investigated Metro Network Topology](image)

The network topologies used in the simulations are based on the structure of a typical Metro Network. It usually consists of two significantly different parts (as Figure 1 shows): a core and several access parts. The first one transports the aggregated traffic between different access parts and the edge nodes and it is formed of one or more interconnected rings. The access parts aggregate the traffic originated from the access nodes to one or more core switches and usually have either tree or dual-homed structure. However, to provide protection two disjoint paths are needed that is possible only when dual-homing is used.

Obeying the previously discussed assumptions two topologies are defined. Both have the same core part consisting of four core and two edge bridges. The core bridges form a 1 Gbps ring and the edge bridges are connected to two of the four core ones with GbE links. The access parts follow the same dual-homing structure as Fig. 1 shows. Each bridge at the top of the access part is connected to two of the four core bridges. The difference between the two topologies is the number of such aggregation parts: the small topology has only one while the large topology has two of them, connected to bridges #3 and #4.

Demands are generated for each pair of access and edge nodes for each QoS class. The demand sizes follow the standard normal distribution where the mean value is the product of the global traffic level (F) and a QoS class traffic ratio parameter (Tr). This value expresses that the amount of the traffic belonging to the various QoS classes can have large differences (see Table III). A moderate variation is assumed (about 5–10% of the mean value).

The topology and the set of generated demands form together a problem class, in which 12 independent problem instances are generated. We solved each problem instance with all the proposed and reference methods. A solution of the problem consists of

1) the optimized trees,
2) the assignment of trees and demands, and
3) the allocated capacity on each link and other statistics.

The results presented are averaged of the 12 simulations.

### C. Evaluation Criteria

We have used 4 criteria for evaluation as follows.

- **Fair available throughput** means the maximal amount of transmitted traffic when the ratios of the traffic classes are fixed, i.e. increasing the rate of one demand does not deteriorate the rate of the other demands.

- **Spare capacity** (or backup capacity) is defined by the allocated capacity devoted to provide the fault tolerance.

- **Fault tolerance** is described by the bandwidth lost in case of a single link failure. Two statistics are defined: the average and the maximal loss relative to the throughput. Both are calculated for all the traffic and for the QoS traffic only. In this latter case the BE traffic is not considered.

- **Scalability**

They will be explained in more details in IV-D.1, IV-D.2, IV-D.3 and IV-D.4.

### TABLE III

**SELECTED TRAFFIC RATIOS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Platinum</th>
<th>Gold</th>
<th>Silver</th>
<th>Best Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic ratio Tr&lt;sup&gt;2&lt;/sup&gt;</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Related to the total traffic</td>
<td>3%</td>
<td>13%</td>
<td>30%</td>
<td>54%</td>
</tr>
</tbody>
</table>

**342**

Authorized licensed use limited to: BME OMIKK. Downloaded on November 21, 2008 at 04:52 from IEEE Xplore. Restrictions apply.
D. Performance evaluation

1) Fair available throughput: This criterion is determined using a binary search within a predefined range of global traffic levels (\( \Gamma \)). If the generated problem instance is solved, the variable \( \Gamma \) will be increased; otherwise it will be decreased. This evaluation method ensures the fairness, i.e., the ratios of the demands of different QoS classes remain the same.

![Graph](image)

Fig. 2. Available throughput achieved by 3 proposed and 2 reference methods for the two investigated topologies.

Figures 2(a) and 2(b) show that compared to the unprotected \( MSTP_{opt} \) introducing protection the throughput is significantly decreased as it was expected. However, \( FullProt \) protection scheme results in roughly the same throughput level as the plain "topology-driven" MSTP without protection, while if only the QoS traffic (only 47% of the total) is protected with 1:1 (\( QoSProt \)), there is about 40% throughput gain over MSTP.

These ratios come from the coincidence of effects of two factors. First, the 1:1 protection scheme roughly halves the available throughput; while the optimization of MSTP trees doubles the throughput!

The effect of the first factor is obvious and well investigated. However, why the optimization doubles the throughput is due to the use of the redundant links: if a bridge has two paths with the same path costs, with default settings the path offered by bridge with lower ID will be selected. Thus, all the access nodes will be connected to the same bridge in the access network. But if the traffic of access nodes is balanced equally between the paths the double throughput can be ensured.

![Graph](image)

Fig. 3. Total allocated working and spare capacities of the investigated protection schemes. The amounts of the spare capacities relative to the working ones (lighter and darker parts of the columns) are depicted.

3) Fault tolerance: We have shown above that the network spare capacity can be decreased if the QoS traffic is protected only. However, the cost of the higher utilization is the lower fault tolerance for the unprotected BE traffic. This criterion is described as average and maximal throughput loss compared to the whole throughput. Both losses are measured immediately after the link failure, so the restoration capability of the STP is not considered at this time. Nevertheless, if we rely only on STP, the QoS requirements are not surely guaranteed after a failure: the QoS constraints may be violated after tree reconstruction, since it is possible that more traffic gets routed on the links of the reconstructed tree than their QoS constraints allow. This results in a deteriorated QoS.

Table IV shows the statistics for these two losses on the large topology. We can see that \( STP \) performs always worse than the other methods: the whole offered traffic can be lost if a link placed close to the root of the tree fails. The \( MSTP \) based methods loose at most half of the throughput on the
investigated topologies thanks to the multiple trees that can use different links.

Optimization itself ($MST_{opt}$) results in a slightly lower loss in average; but in worst case it losses the same percentage of the total traffic. The reason is that it balances better the traffic between the links; however, the link failures near the root remain crucial.

The $QoSProt$ scheme is shown to introduce further increase of fault tolerance compared to the unprotected methods even in worst case. Moreover, it is important to emphasize that all the traffic lost in case of $QoSProt$ is best effort, i.e., there is no loss of QoS traffic at all! (See Table IV.)

Obviously, the $QoSProt$ performs worse than the $FullProt$, which does not lose any capacity in case of single link failure. However, $FullProt$ requires much more spare capacity: 130% of working capacity instead of about 20% measured in the case of $QoSProt$.

These results show that the $QoSProt$ is a reasonable solution for realizing efficient protection over Ethernet. It is important to notice that if some best effort traffic demands should be protected, for the selected BE demands a backup path is also determined.

4) Scalability: Since the optimization task is usually performed off-line, there is no strict time constraint for the method. Nevertheless, the running times of the methods are important from the point of view of scalability.

Theoretically, the trees we search for are in all cases Steiner trees, not spanning ones in graph theoretic sense. Therefore, all the optimization problems considered above are NP-complete. In practice, the simulations show that the tree optimization without protection requires about 20 minutes to give a solution in case of the large topology. Introducing the dedicated protection schemes the running time increases up to 3 hours on a Dual AMD Athlon MP 2000+. Besides, there was slight difference between the Full and the QoS protection schemes.

E. Lessons Learned

The simulations on the investigated topologies indicate that the protected and optimized spanning trees have roughly the same performance as the standard based $MSTP$ that supports slow restoration only. This is in consequence of the effects of applied protection and “traffic-driven” tree optimization method. Further throughput gain can be achieved if the QoS traffic is protected only. However, to provide the same throughput, the $FullProt$ and $QoSProt$ with $MST_{opt}$ consume much more network resources than plain $MSTP$ does. Besides, the main goal of applying optimized protection trees was demonstrated: lower ratio of the traffic is lost in case of a single link failure meanwhile higher throughput can be achieved.

Last, but not least, the concept of protecting only QoS traffic ($QoSprot$) is proved to be a rational tradeoff between the availability and network capacity consumption: the throughput can be increased about 30% compared to the 1:1 full protection at the price of having the chance of loosing the best effort traffic.

Finally, it is important to notice that all the optimization problems studied in this paper are NP-complete that urges developing heuristic algorithms. Furthermore, these latter methods will enable shared and other more sophisticated protection schemes.

V. Conclusion

To achieve quality of services within a network, the service provider has to plan and engineer his network. In this paper we have described a novel optimization framework for Ethernet network configuration based on MSTP aiming to provide carrier-grade services. Our framework makes use of the traffic engineering abilities enabled by the multiple spanning tree instances in order to maximize network utilization and also to provide end-to-end protection within the network. A further significance of our optimization framework is that we handle and differentiate multiple QoS classes within a single network. High availability is provided by protection switching of the traffic in case of failures, independently of the network topology.

The main advantages of the protection method over the restoration offered by the Spanning Tree Protocol are in the speed (range of milliseconds) and that QoS guarantees are retained even in case of failure.

We have presented two protection methods: one when all traffic is protected and one protecting only the high priority traffic with preemption. The results show that protecting all traffic wastes resources while “1:1”-protecting only the priority traffic is a more realistic scenario with acceptable capacity requirements.

REFERENCES

[5] IEEE 802.1ad/D4.0 DRAFT Amendment to IEEE Std 802.1Q -2003, February 8, 2005