# Hierarchical Restoration Scheme for Multiple Failures in GMPLS Networks

SuKyoung Lee*, Chul Kim and David Griffith
National Institute of Standards and Technology
100 Bureau drive, Stop 8920, Gaithersburg, MD, 20899, USA.

*Abstract*— **It is expected that GMPLS-based recovery could become a viable option for obtaining faster restoration than layer 3 rerouting. Even though dedicated restoration ensures restorability of connections, exclusive use of dedicated scheme would result in wasting network resources, especially in case of providing for multiple failures. A range of restoration schemes have been proposed that use the concept of sharing capacity to improve efficiency. However, the case of multiple simultaneous failures has not been considered. In this paper we propose a hierarchical scheme for handling multiple simultaneous failures, where hierarchical Shared Risk Link Groups (SRLGs) are applied. We also introduce Backup Group Multiplexing (BGM) into our hierarchical scheme to precipitate the restoration of multiple Label Switched Paths (LSPs) with failures all at once. Furthermore, the proposed scheme selects a backup path with enough resources to satisfy renegotiated Quality of Service (QoS) of each backup group, among $M$ backup paths. Our simulation results demonstrate that our scheme utilizes bandwidth more efficiently through multiplexing gain.**

## I. Introduction

With the migration of real-time and high-priority traffic to IP networks and the deployment of more and more critical applications, network survivability has become critical for future IP networks. Above all, fast recovery of service is an important aspect of current and future IP networks. To support this fast recovery, most recently, there have been many works mentioning restoration functionality in both MPLS [1] and GMPLS networks [2]. Different from legacy IP networks, these protocols establish LSPs, where packets with the same label follow the same path. This potentially allows GMPLS networks to pre-establish backup LSPs for working LSPs, and achieve better protection switching times than those in legacy IP networks.

These protocols, however, can still cause serious disruption of service while shared restoration is being run over the Internet. For $M:N$ mode, even though backup paths are pre-reserved, it takes sometime for a working path with failure to search for a backup path with available capacity if the other working paths with failures have been already occupying some backup paths from backup path pool. When the resources are in use for other low priority paths and the backup resources are needed, it also takes sometime to tear down the low priority connection, if backup paths are not pre-reserved. In such a case, the backup paths are configured in advance at the ingress, but are not signaled until the failure is reported. This is unacceptable for many applications that require a highly reliable service, and has motivated network providers to give serious consideration to the issues of not only network survivability but also restoration time.

Another major challenge of restoration in GMPLS networks is capacity. In order to achieve protection against failures, enough capacity must be provided for the interrupted traffic to be restored. Several capacity-efficient restoration schemes have been proposed [3],[4]. But, these schemes are restricted to single SRLG failure event.

Therefore, in this paper, we will propose a scheme to handle multiple simultaneous failures over SRLG disjoint resources. In the proposed scheme, in the event of multiple failures over SRLG disjoint resources, we apply different restoration scheme by defining a higher level SRLG for the SRLG disjoint resources as in [5], while the existing shared capacity scheme [3],[4] still works in the event of single failure. Especially in the event of multiple simultaneous failures, while some primary paths of high priority are supposed to use the pre-reserved shared capacity, for the rest of the paths with failures, restoration controller performs GMPLS-based $M:N$ protection mechanism without pre-reservation. In order to recover multiple failed LSPs promptly at once, we multiplex the LSPs into some backup groups, QoS of which are different from each other. The QoS of each backup group is renegotiated, since, usually, resource would be scarce to restore all the simultaneously failed LSPs.

The performance of the proposed scheme is demonstrated by simulation. Simulated tests are useful to verify that our scheme recovers multiple simultaneous LSPs promptly through BGM while improving resource utilization.

This work was supported, in part, by the National Institute of Standards and Technology (NIST), the Advanced R&E Activity (ARDA), the Laboratory for Telecommunications Sciences (LTS) MENTER project, the Defense Advanced Research Projects Agency (DARPA) Fault Tolerant Networks (FTN) program, and the National Communications System (NCS).

## II. Hierarchical Restoration

We have two hierarchies for SRLG where at the level of individual fibers or fiber bundles we use the concept of a physical SRLG (PSRLG), and at higher logical levels we use the concept of a logical SRLG (LSRLG). In optical networking literature, references to SRLGs typically involve objects at the PSRLG layer, so the diversity algorithms being proposed compute constraint-based paths which are PSRLG-disjoint. The LSRLG concept is defined to target some regions with high failure probability of multiple simultaneous failures over different PSRLGs. In our scheme, an LSRLG is a shared risk group that covers multiple nodes and links and can correspond to a geographical region as proposed in [5] or an administrative region controlled by a specific carrier. Grouping network elements into LSRLGs by region allows the network operator to account for increased risk of element failures due to natural events (e.g. earthquakes) or intentional disruptions (e.g. sabotage). Also over a region operated by a specific carrier, multiple simultaneous failures occur on various PSRLG-disjoint links/nodes due to serious outages by software upgrades in large-scale network infrastructures, or deficits in network management tools [6].

While many researchers [3],[4] have mainly focused on using PSRLGs to provision backup paths, we use a different restoration scheme at each of the two layers in the SRLG hierarchy. Even though the algorithms in [3],[4] benefit from sharing capacity among backup paths whose primary paths are PSRLG (e.g. link/node) disjoint, they assume that any failure event will affect only one PSRLG. Consider for example the network topology shown in Fig. 1, which is taken from [4]. When failures occur on both link groups $a$ and $c$, not all the primary paths carried on those links can be recovered because the protection resource sharing scheme proposed in [4] does not provide sufficient resources on link group $h$. If, however, we treat multiple PSRLG failures as a failure of a higher layer

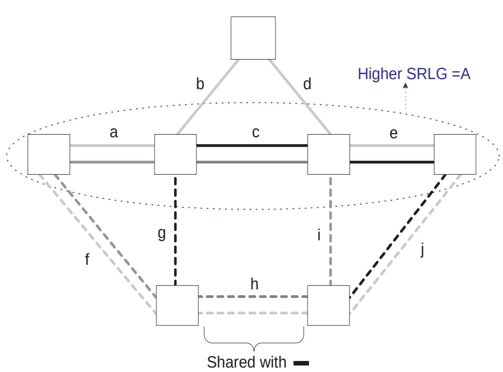


Fig. 1. Example Hierarchy

LSRLG $A$, shown in the figure, we can use a protection mechanism at the LSRLG layer to restore the traffic affected by the multiple failures on link groups $a$ and $c$. Generally, multiple PSRLGs for links/nodes are grouped into new LSRLG for a region in our model. Thus, LSRLG protection will be able to recover from multiple simultaneous failures over multiple PSRLG-disjoint resources that use shared protection mechanisms. To restore all the traffic affected by these multiple failures over PSRLGs, we use a $M : N$ restoration function at the LSRLG layer that we describe in detail in the next section.

## III. Backup Group-based Restoration over LSRLG

### A. Backup Grouping

Since the proposed $M : N$ restoration with BGM (we call this MN_BGM for the rest of the paper) operates over an LSRLG an allows recovery from multiple failures at the PSRLG layer, pre-reserving enough bandwidth to recover from all the multiple failures will lead to enormous bandwidth waste. Thus we assume that there is no pre-reservation on the backup paths. MN_BGM operates by ranking the failed working paths according to priority based on QoS and assigning them to the available pool of backup resources, granting the highest-priority access to the protection resource pool first.

In particular, when $N$ working paths that have shared backup bandwidth experience multiple PSRLG failure events, the traffic of the highest priority among the failed paths is allowed to use the shared bandwidth unconditionally, so that it effectively receives $1 : 1$ protection. We assume that this traffic requires high QoS, analogous to the Expedited Forwarding (EF) class of traffic in the Differentiated Services (DiffServ) framework [7]. For the rest of the traffic over the failed paths, the MN_BGM function is carried out according to the flowchart depicted in Fig. 2. This remaining traffic belongs to service classes that have lower QoS and, therefore, lower priority than the EF traffic that has been already been restored. This "second-tier" traffic can be viewed as belonging to the Assured Service (AS) or Best Effort (BE) classes. Under MN_BGM, the traffic of lower priority on the failed paths is multiplexed into several LSPs by aggregating service classes of individual failed paths into new service classes having common per-LSP QoS parameters. Thus, MN_BGM results in restoring the failed paths faster than existing $M : N$ restoration scheme in which restoration manager tries to search for individual backup path corresponding to each failed path, because MN_BGM restores traffic on a per-class basis rather than on a per-path basis. In the case of optical network, a several traffic flows of lower priority over the failed paths can also be multiplexed onto

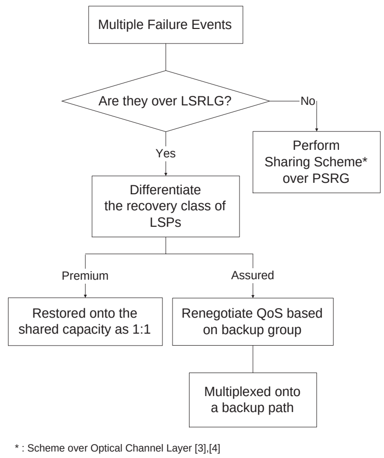


Fig. 2. Basic Flow in MN_BGM

higher-capacity wavelength [8]. Another advantage over original $M:N$ restoration is to get multiplexing gain by aggregating primary LSPs, that leads to accommodating more traffic on the backup path. This multiplexing gain makes MN_BGM scheme allocate less bandwidth than the requested demand (e.g. PDR: Peak Data Rate or CDR: Committed Data Rate) to meet the QoS.

In order to define restoration-related classes to support each class of failed working paths, the restoration manager must renegotiate QoS with each group of failed traffic streams. We model this by classifying the failed paths into the following four groups: Restoration Class-High (RC-H), RC-M (Medium), RC-L (Low), and RC-N (None). These groups and their characteristics are listed in Table I. To support the renegotiation of bandwidth resources, we introduce the concept of restoration bandwidth ratio ($RB$), which is the ratio of the bandwidth on the protection path to the bandwidth on the working path.

The RC-H class corresponds to Expedited Forwarding (EF) service in DiffServ. For all the traffic flows in this class, backup paths are pre-configured and bandwidth is also pre-reserved due to the precedence for the shared bandwidth on PSRLG layer. Thus, the backup paths are guaranteed to support the same level of QoS as the working paths, and $RB = 1$. The RC-M class can be regarded as AF1 service in DiffServ. While backup paths are pre-configured as part of $N$ paths, bandwidth is allocated on-demand. In addition to on-demand allocation, the bandwidth necessary for restoration (we call restoration bandwidth) is renegotiated or determined offline to satisfy constraints on the restoration bandwidth ratio, $RB \geq \varepsilon_m$ ($< 1$). The RC-L class is similar to AF2 service in DiffServ. All the restoration plans are same as in RC-M except for meeting different QoS $RB \geq \varepsilon_l$ ($< \varepsilon_m < 1$). The RC-N class corresponds to Best Effort (BE). No restoration is provided for traffic of this type.

TABLE I
RESTORATION SERVICE CLASS

| Service | Traffic | Backup path | Resource |
|---|---|---|---|
| RC-H | Interactive | Pre-configured | Pre-reserved |
| RC-M | Non-interactive | Pre-configured | None |
| RC-L | Non-interactive | Pre-configured | None |
| RC-N | Unspecified | None | None |

Interactive: real-time; Non-interactive: low-loss

For the RC-M and RC-L classes, it is reasonable to control the traffic flow after renegotiating the bandwidth, because there are no quantifiable delay requirements associated with the forwarding of AF packets [9]. Because traffic classification depends on each Internet Service Provider (ISP)'s policy, ISPs could provide different levels of restoration services with the higher level service at the higher cost when they control multiple failures in an area like an LSRLG. Moreover, ISPs can pre-negotiate offline the QoS after failure, for RC-M and RC-N classes.

The estimate of $RB$ can be computed on the assumption that failures occur on an working LSP at random. Considering there exist $N$ working LSPs under MN_BGM scheme, $N_f (\neq N)$ failed paths are classified into the above service classes as $G_h$, $G_m$, $G_l$, and $G_n$ each of which includes $N_h, N_m, N_l$, and $N_n$ failed paths. Thus, the set of $N_f$ failed LSPs, $\{P_1, P_2, \cdots, P_{N_f}\}$ is expressed as the set of groups, $\{G_h, G_m, G_l, G_n\}$. Generally, ISPs determine grouping policy in accordance with restoration service classes.

For a backup group with $N_f$ (= $N_h$, $N_m$, or $N_l$) paths, some notations are defined:

- $B_f$ is the total bandwidth affected by failure. $B_f = \sum_{i=1}^{N_f} B_f^i$, where $B_f^i$ is the bandwidth affected by failure over LSP $P_i$.
- $B_{min}$ is the minimum guaranteed bandwidth that users in a group requested ISP to offer. $B_{min} = \sum_{i=1}^{N_f} B_{min}^i$, where $B_{min}^i$ is the minimum guaranteed bandwidth for $P_i$.
- $B_{RN}$ is the bandwidth necessary to satisfy renegotiated QoS. $B_{RN} = \sum_{i=1}^{N_f} B_{RN}^i$, where $B_{RN}^i$ is the renegotiated bandwidth for $P_i$.

In accordance to each restoration class, we define the bandwidth necessary to restore a backup group, $B_r$ as

$$B_r = \begin{cases} B_f & \text{if } P_i \in G_h \\ B_{min} & \text{if } P_i \in G_m \\ B_{RN} & \text{if } P_i \in G_l \end{cases} \qquad (1)$$

Given that $B$ is the whole bandwidth on a backup path belonging to $M$ backup paths, the $RB$ for each backup group is calculated as

$$RB = min\left(\frac{B - B_g}{B_r}, 1\right), \qquad B_r \neq 0 \qquad (2)$$

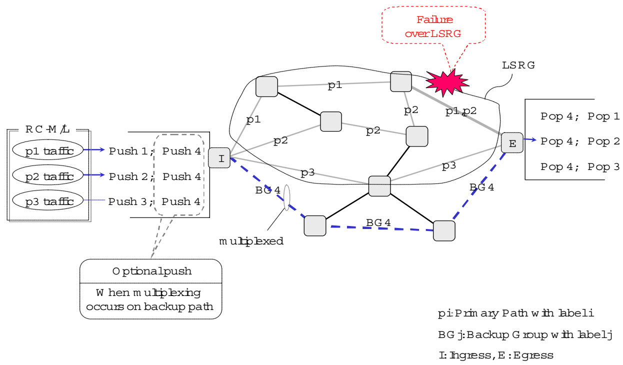


Fig. 3. Label Stacking for Backup Grouping

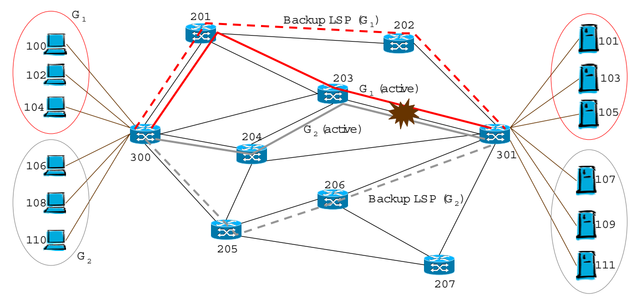


Fig. 4. Network scenario for MN_BGM scheme

where $B_g$ is the guaranteed bandwidth for premium and assured services that use the backup path as their primary path while restoration process is going on. Based on the customers' tolerable QoS degradation after failure, ISPs could not only determine $RB$ but also appropriate the bill for each group. In accordance with $RB$ for a restoration service group, restoration manager could find appropriate backup path to meet the $RB$.

### *B. Label Stacking for Backup Grouping*

We use label stacking [10] at edge nodes to multiplex some LSPs into a backup group. As a matter of fact, some other proposed mechanisms [11],[12] could be also used to multiplex working LSPs onto a backup LSP. While in [11],[12], multiple LSPs could be merged into a single Forwarding Adjacency (FA) LSP, handling by OSPF/ISIS might result in slow restoration. That is, grouping of working LSPs should go through layer 3 at ingress LSR. To avoid this overhead on layer 3, when working LSPs are multiplexed into a backup group, we can make use of the stacking capability of GMPLS. An inner label is used to help guide the traffic on primary LSPs with failures, to a backup LSP. While, in MN_BGM scheme, all the failed primary LSPs in a group have the same inner label, the original labels of the failed primary LSPs become outer labels. The label assignment process is shown in Fig. 3. But in optical network, instead of label stacking, there should be a change of the corresponding ports between GMPLS router and optical switch which are integrated in Optical cross-connect (OXC) [8].

## IV. Simulation and Performance Evaluation

This section evaluates two aspects of the performance of MN_BGM, with bandwidth utilization and restoration time. As a simulation tool, we used GMPLS Lightwave Agile Switching Simulator (GLASS) [13] developed by NIST, that is the extension of Scalable Simulation Framework (SSF) Net [14].

Using the network topology in Fig. 4, we generated TCP traffic as files whose size is exponentially distributed with various mean values, $10^5, 10^6, 10^7$, $10^8$, and $10^9$ bytes. File requests are assumed to arrive at hosts 100,102,104,106,108 and 110, according to a Poisson process with a rate of $0.5$. In the simulation test herein, the traffic of AF service exists (e.g. RC-M or RC-L) where the schemes are tested with two QoS classes, $G_1$ and $G_2$, that usually determine ISPs to classify RC. The CDR and PDR for $G_1$ and $G_2$ are 2.5 and 1 Mbps, and 2 and 0.8 Mbps, respectively.

Fig. 4 shows that our simulation model has 3 active working paths for each restoration service class between edge nodes, 301 and 302 (i.e. $N_f = 6$). The bandwidth of each active path is guaranteed as CDR not interfered with by other traffic in the network. In the simulation model, given that a failure occurs on link between 203 and 301 nodes, while the traffics belonging to $G_1$ are multiplexed onto a backup LSP (300-201-202-301), the traffics of $G_2$ are done so onto a backup LSP (300-204-203-301).

Fig. 5 shows the average bandwidth utilization on each backup path under different traffic amounts on the failed paths when $RB$ is 1. According to Fig. 5, the bandwidth allocated to each backup group is lightly utilized. From these results, even though MN_BGM scheme tries to restore failures through the degradation of original QoS ($RB < 1$), especially for AF services, it can be known that more traf-

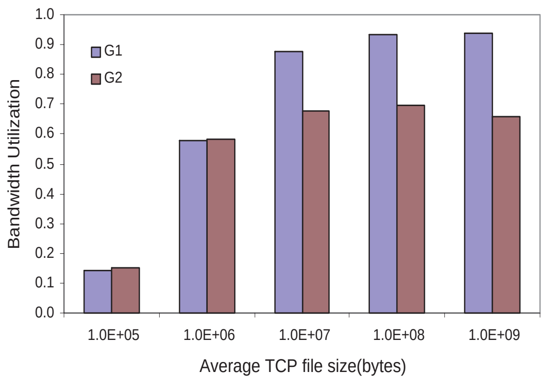


Fig. 5. Bandwidth utilization under various traffic amounts

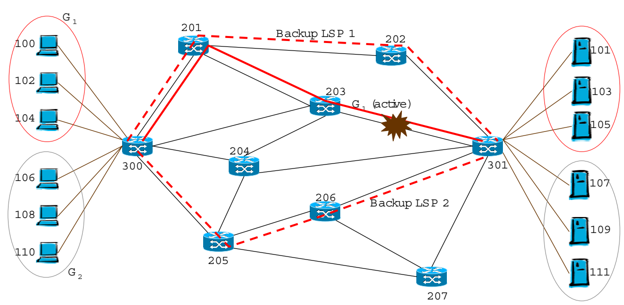


Fig. 6. Network scenario for $M : N$ restoration

fic can be accommodated on the backup path. That comes from multiplexing gain through backup group multiplexing. Therefore, ISPs can make customers send more traffic than the negotiated $RB$ for the AF services. For all the ranges of traffic amounts, the average loss rate did not exceed $4 \times 10^{-4}$. The fact indicates that ISPs could allow more traffic than the amount which $RB$ can accommodate on the backup paths, maintaining reasonable loss rate.

To compare the performance of MN_BGM with $M : N$ restoration, we have also tested original $M : N$ restoration for the LSPs of $G_1$. The simulation set-up is same as in MN_BGM except that only the TCP file with mean size of $10^7$ bytes was generated. We see that there is no backup group multiplexing in $M : N$ restoration scheme as can be seen in Fig. 6. As GMPLS signaling, we have used Constraint-based Routing-Label Distribution Protocol (CR-LDP) [15] in this test, which were also implemented by NIST. Comparing the restoration time, $M : N$ restoration is 2 times slower than MN_BGM scheme. Moreover, $M : N$ restoration could not restore the failed path between nodes, 104 and 105 due to the lack of backup paths. These results imply that MN_BGM scheme is more suitable for multiple failures.

TABLE II
BANDWIDTH UTILIZATION FOR $M : N$ RESTORATION AND MN_BGM

| $M : N$ Restoration | | |
|---|---|---|
| Backup path | Bandwidth Utilization | Loss Rate |
| $BP_1 \in G_1$ | 0.67 | no loss |
| $BP_2 \in G_1$ | 0.62 | no loss |
| $BP_1 \in G_2$ | 0.43 | 0.00069 |
| $BP_2 \in G_2$ | 0.39 | 0.00067 |
| $BP_3 \in G_2$ | 0.42 | 0.018 |
| MN_BGM | | |
| Backup path | Bandwidth Utilization | Loss Rate |
| $BGP \in G_1$ | 0.80 | no loss |
| $BGP \in G_2$ | 0.67 | 0.0005 |

$BP$: Backup Path, $BGP$: Backup Group Path

Table II shows the bandwidth utilization of both schemes. It can be seen from this table that MN_BGM produces better bandwidth utilization than $M : N$ restoration. It is because MN_BGM scheme makes better use of multiplexing gain. In other words, this scheme makes it possible to multiplex low rate traffics into a backup path to improve the utilization of the paths. In case of optical network, the low rate traffics are groomed into a lightpath. From this table, we could also know that the proposed scheme is capable of maintaining acceptable loss QoS, as well as improving bandwidth utilization. Therefore, MN_BGM would give ISPs useful information about $RB$, an appropriate selection of which cannot only improve resource utilization but also guarantee a certain loss quality.

## V. CONCLUSION

This paper describes a novel approach to resolve multiple failures. This approach facilitates resource-efficient service restoration as well as restoration of multiple failures. The proposed MN_BGM scheme restores the failed paths with the highest restoration level firstly by performing the existing sharing schemes which were developed over a network with only PSRLG such as optical network, while the remaining failed paths of lower priority are restored by backup grouping and QoS renegotiation functions in MN_BGM scheme. Therefore, when mutiple failures occur, the proposed hierarchical scheme can restore the multiple failed paths instead of loosing them at once, that is the usual phenomenon of the existing scheme handling only single failure. This scheme also enabled us to devise a bandwidth-efficient restoration for the traffic of lower priority, relying on BGM. In addition, the investigation of our simulation results has shown that the improvement in bandwidth utilization can be effectively realized by BGM while keeping loss QoS at a moderate level.

Finally, our ongoing work is to extend MN_BGM scheme with decent backup bandwidth provisioning mechanism. Instead of renegotiated QoS parameters, we are focusing on bandwidth provisioning via traffic measurements at ingress node of backup path to exploit multiplexing gain more effectively.

## REFERENCES


[1] V. Sharma, et al, "Framework for MPLS-based Recovery", *Internet Draft*, draft-ietf-mpls-recovery-frmwrk-03.txt, Feb. 2002.

[2] E. Mannie, et al, "Recovery (Protection and Restoration) Terminology for GMPLS", *Internet Draft*, draft-mannie-gmpls-recovery-terminology-00.txt, Feb. 2002.

[3] G. Li, D. Wan, C. Kalmanek, and R. Doverspike, "Efficient Distributed Path Selection for Shared Restoration Connections", *INFOCOM*, 2002.

[4] S. Datta, S. Sengupta, S. Biswas, and S. Datta, "Efficient Channel Reservation for Backup Paths in Optical Mesh Networks", *GLOBECOM*, 2001.

[5] D. Papadimitriou, et al, "Inference of Shared Risk Link Groups", *Internet Draft*, draft-many-inference-srlg-02.txt, Nov. 2001

[6] A. P. Snow, "Network Reliability: The Concurrent Challenges of Innovation, Competition, and Complexity", *IEEE Trans. on Reliability*, vol.50, no.1, Mar. 2001.

[7] P. Trimintzios, et al, "A management and control architecture for providing IP differentiated services in MPLS-based networks", *IEEE Commun. Magazine*, vol.39, no.5 , pp.80-88, May 2001.

[8] C. Assi, et al, "On the Merit of IP/MPLS Protection/Restoration in IP over WDM Networks", *GLOBECOM*, 2001.

[9] J. Heinanen, et al, "Assured Forwarding PHB Group", *RFC* 2597, Jun. 1999.

[10] E. Rosen, et al, "MPLS Label Stack Encoding", *RFC* 3032, Jan. 2001.

[11] A. Banerjee, et al, "Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements", *IEEE Commun. Magazine*, vol.39, no.1, pp.144-150, Jan. 2001.

[12] K. Kompella, et al, "LSP Hierarchy with Generalized MPLS TE", *Internet Draft*, draft-ietf-mpls-lsp-hierarchy-04.txt, Feb. 2002.

[13] NIST, "GMPLS Lightwave Agile Switching Simulator", *http://www.antd.nist.gov/proj/asia/glass*.

[14] "SSF Net: Scalable Simulation Framework", *http://www.ssfnet.org*.

[15] P. Ashwood-Smith, et al, "Generalized MPLS Signaling - CR-LDP Extensions", *Internet Drafat*, draft-ietf-mpls-generalized-cr-ldp-05.txt, Nov. 2001.