Regular Paper

Concept of Virtual Path Hopping to Improve QoS and Reliability of Connection-oriented Networks

Manodha Gamage,† Mitsuo Hayasaka† and Tetsuya Miki†

The QoS provided by today's best effort Internet is not good enough for real-time premium traffic. It is believed that the QoS guarantees of the Internet could be better provided by connection-oriented networks. IP/MPLS is one such technology and these connection-oriented networks are inherently more vulnerable to network failures. Network failures can broadly be classified into two types, ie., link/path and degraded failures. Degraded failures that account for about 50% of total failures are detected by control-plane timers. The control plane and the data plane of IP/MPLS networks are logically separated and therefore a failure in the control plane should not immediately terminate communications on the data plane. The Virtual Path Hopping (VPH) concept proposed in this study distinguishes these degraded failures from link/path failures and avoids the terminations of data communications on the data plane that are due to these degraded failures. It changes the traffic carrying a degraded VP to a new VP in a predetermined VP-pool before the control-plane timer expires due to degraded failure. This paper discusses the concept of VPH and its numerous advantages such as load balancing, traffic distribution, and the ability to support dual-failure situations. Computer simulations were conducted and the results are presented to support the concept of VPH.

1. Introduction

In recent years, the number of Internet users has grown enormously and they now prefer more capabilities in many dimensions such as transmission bandwidth, the number of hosts, reliability, and real-time multimedia applications. The emerging real-time multimedia applications that are categorized as Premium Traffic (PT) have the following main features. First, the durations of the sessions are very long in most applications such as remote lecturing and teleconferencing. Second, in most of these applications, such as telemedicine and e-commerce, the consequences of losing QoS guarantees can be very severe. The increment in the number of nodes and links due to expansions of networks may also increase the number of failures. These future applications are driving the Internet to move from a best-effort network to one that is more reliable and can assure a QoS guarantee. We believe that this can be better achieved by connection-oriented networks than connectionless IP networks. Connection-oriented high -peed packet networks such as IP/MPLS[1] would be widely used in the future as they improve QoS by reducing packet losses, delay jitter, and bandwidth variations. The drawback to these networks is their potential vulnerability to network failures. According to Autenrieth and Kirstäter[10], an e-commerce company with 99% availability (1% unavailability) would lose about $3.6 million annually due to network failures, and at current costs and volume of business, this may be even more. Therefore, instead of fast re-routing techniques, it is essential to find proactive techniques to minimize the occurrence of failures, because re-routings result in service outages.

According to Sharma and Hellstrand[3], the failures in these networks can be categorized into degraded and link/path failures. Degraded failures occur due to links in lower layers not being of suitable quality to guarantee data transmission. They do not immediately disconnect communications on the data plane due to the logical separation of control and data planes in IP/MPLS networks. They account for almost 50% of total failures and the focus of this paper is on these degraded failures that are usually identified by the timers on the control plane. The objective behind the novel idea of Virtual Path Hopping (VPH) proposed here is to overcome the terminations in data communications on the data plane due to these degraded failures, especially for PT in IP/MPLS networks that uses in-band control channels. It identifies degraded failures using timers on the con-

† Department of Electronics (Miki Laboratory), The National University of Electro-Communications

trol plane, before data-plane-communication sessions fail and the traffic carried VP that is degraded is changed to a new QoS guaranteed VP by way of a VP hop. Our results obtained from computer simulations reveal that this is a promising proactive technique to minimize the occurrence of failures on the data plane, especially for PT. MPLS with RSVP-TE is used here for all explanations as it is easy to understand the VPH concept with well defined protocols. Also, IETF has decided to promote RSVP-TE over CR-LDP[30]. Other major advantages of VPH are its dynamic distribution of traffic and its ability to handle dual-failure situations better.

The rest of this paper is organized as follows. In the next section, a description of the problem is given and the existing solutions are briefly analyzed. In Section 3, the proposed VPH concept is discussed in detail. Numerical analysis is discussed in Section 4. The performance of the VPH concept was evaluated through computer simulations and the results are presented in Section 5. Finally, Section 6 concludes the paper.

2. Problem Description and Existing Solutions

2.1 Network Failures

Failures in networks can be due to many reasons such as hardware and software malfunctions, link failures, routine maintenance, high congestion, protocol failures, restart of control-plane nodes and failures of control functions, and loss of adjacencies. Studies have shown that in a core network, about 10% of failures last for over 20 minutes, 40% last between 1–20 minutes, and 50% are very short-lived, being less than a minute[4]. According to Sharma and. Hellstrand[3], network failures with MPLS can mainly be classified into two types, i.e., link/path and degraded failures. A link/path failure means a situation where the actual connectivity of the links/paths between the ingress and egress is lost. Degraded failures occur because links in lower layers are not of suitable quality for data transmission and this paper is focused on explaining how these failures can be overcome. Most of the short-lived failures that account for about 50% of total failures can be due to degraded failures such as loss of adjacencies[4]. When adjacencies are lost and if control-plane timers expire, corresponding control-plane sessions are torn down causing terminations to data communications on the data plane. Also, RSVP teardowns after control-plane-timer expirations can be due to the loss of refresh messages or hello protocol failures, restarts or failures of control-plane nodes, and congestion[2),9]. The values for these control-plane timers (T) are usually determined when the control-plane session is established through negotiation with peers; they are usually in a range of 30–40 s, but they can be as long as 60–90 s[9]. The purpose of these timers is to reduce the convergence time after failures occur. Conventionally, the timers are reset whenever Protocol Data Units (PDUs) are received by peers.

As shown in **Fig. 1**, the control and data planes for MPLS are logically separated in recent router architectures and it uses in-band signaling, where control messages are sent over links that carry data. Logically separated control and data planes mean 'call setup request' is always accompanied by a 'connection request'[28]. The control plane performs functions such as setup, termination, and maintenance of the Label Switched Paths (LSP) on the data plane. In other words, there will be a corresponding control-plane session for each LSP on the data plane. Even though MPLS uses an in-band control channel, any problems on the control plane should not immediately terminate communications on the data plane due to their logical separation. If the control channel fails, the corresponding data plane is degraded and it can no longer guarantee QoS, irrespective of whether the control plane is logically separated from the data plane as in IP/MPLS or is completely separated and managed by the link management protocol[31] as proposed in GMPLS[32].

If such a degraded state continues beyond the threshold values of the control-plane timers (T), data-plane premium applications would decide to terminate their communications due to no guarantees of QoS. It is then necessary to do re-routing to recover the terminated data communications and these re-routings cause data losses in PT. In other words, degraded failures due to congestion, protocol and functional fail-

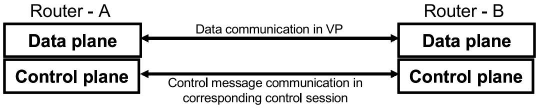


Fig. 1 The Control and Data planes of connection oriented networks.

ures on the control plane, teardowns of control-plane sessions due to loss of adjacencies and lack of refreshing, and restarts of control-plane nodes could lead the control-plane timers (T) to expire resulting in terminations in data-plane communications.

2.2 Existing Solutions and their Problems

The existing solutions to network failures in connection-oriented networks such as IP/MPLS can be broadly classified into three types, namely, local repair, path protection, and fast re-routing. The communication of signaling information in MPLS uses IP and therefore re-signaling an LSP due to failure will be time consuming. Furthermore, a signaling protocol such as RSVP-TE concentrates more on traffic engineering and is less favorable for local repairs. Also, network topologies are rarely fully meshed and local repairs might not succeed in MPLS and re-routing may need to be resolved at the ingress. In path protection, data is switched from a failed LSP to a backup LSP at the repair point, conventionally at the ingress. Fast re-routing will occur when the backup LSP can be pre-provisioned. As explained by Huang, et al.[12], path protection is more efficient than local repairs for connection-oriented networks. Some popular solutions to network failures in real-time applications are as follows. These include 1 + 1 protection, where the same data is simultaneously transmitted both in the active and backup paths (AP & BP) and the best path is selected at the receiver end. Another is 1 : 1 protection (extendible to m : n protection), where data is transmitted only via AP and BP is only used if a failure occurs. Therefore, when there are no failures in APs, the BPs can be used by some other not critical, best effort traffic. The 1 + 1 has very fast recovery times but is very inefficient with respect to the usage of bandwidth whereas 1 : 1 improves bandwidth efficiency at the expense of recovery time. Backup bandwidth (BBW) sharing is becoming increasingly popular due to improved bandwidth efficiencies as a single BP can be shared by many link-disjoint APs[8],[13],[26].

Almost all these proposals have considered single fault situations assuming that any failure can be repaired before the next failure occurs. Very few studies done on dual-failure scenarios have revealed that the current BBW sharing schemes with 100% restoration for single faults could on average recover about 60–70% of failures in dual-fault situations but it can be as low as 20%[25]. The expansion of networks and increased durations of applications require future networks to have 100% restorability even for dual faults. Also, network status rapidly varies over time and therefore very high BBW sharing of link-disjoint AP attributes at one time may not be possible at another time and therefore bandwidth utilization may not be as high as expected. Furthermore, none of the above solutions distinguishes between degraded and link/path failures. This is very inefficient as almost 50% of failures are degraded and most of these can be avoided on the data plane, if control-plane sessions can be recovered before data communications are terminated due to control-plane-timer expirations. Graceful restart and fault tolerance[21],[22] in control-plane protocols have been attempts to minimize control-plane failures terminating data-plane communications, but not all routers can preserve forwarding states across restart of the control plane. There would be scalability problems if all routers in the network were to preserve forwarding states and graceful restarts may only be possible for planned failures. Furthermore, even if routers could preserve such data during node failures, this data can still be lost, preventing a graceful restart. Furthermore, they are not suitable for PT as they might lose some data at failover.

3. Virtual Path Hopping (VPH) Concept

The main objective of the proposed VPH is to eliminate terminations in data communications on the data plane due to degraded failures that are detected by the control-plane timers (T) of connection-oriented IP/MPLS networks, thereby achieving very high availability. It is also focused on PT. VPH refers to "changing a degraded active VP (AVP) on the data plane to a new VP that can guarantee QoS before it fails due to the expiration of control-plane timers (T)". This is done by creating a new control-plane session to activate an unused VP in the VP-pool, which is explained below, and transferring the traffic from the degraded AVP, whose control-plane timer is about to expire. The term "re-routing" refers to "the change in data transmission from an AP to a BP, after failure on the data plane". Any restoration by re-routing after a failure would create an outage of service, whereas VPH has no service outages as it uses the make-before-break concept[23] when hopping VPs.

More about smooth VP hopping using a VPH_Timer and a Refresh_Timer without incurring any data losses will be discussed later, referring to Figs. 3, 4, and 5. In IP/MPLS, a VP is called an LSP. The ingress nodes of IP/MPLS will have to play a major role in the implementation of the VPH concept. When traffic arrives, the ingress, which is ideally a link-disjoint VP-pool that contains the candidate VPs that participate in VPH, is determined between the ingress and egress as shown in **Fig. 2**. This VP-pool should contain at least three link-disjoint VPs to make this concept effective and with the increment of the number of VPs in the VP-pool, its effectiveness will increase. The number of VPs in the VP-pool depends on several factors such as network load, the Service Level Agreements (SLA) of clients, failure analysis, and failure distributions in the network. It is not necessary to only restrict VPH to link-disjoint situations and block traffic, whenever there is not a minimum of three fully link-disjoint VPs for a VP-pool. Therefore, we suggest that it is preferable to find a VP-pool with fully link-disjoint VPs for the best VPH performance, but if this condition cannot be met, it is acceptable to create a VP-pool with VPs that are partially link-disjoint provided the QoS requirements of the traffic accepts this. This entirely depends on the requirements of the applications. However, two VPs of a VP-pool should never share the same tunnel. VPH would improve the performance of both fully link-disjoint and partially link-disjoint situations. Reliability of 100% might not be achieved, when all VPs are not fully link-disjoint. It is important to have this option as there may be certain applications that are better off by starting communication with a VP-pool with no fully link-disjoint VPs rather than blocking them. There have been many algorithms proposed in the literature[16]~[20] to find link-disjoint paths between an ingress and egress pair. It is beyond the scope of this paper to discuss these in great detail. Also, instead of only using Shortest Path First (SPF), to determine the VP-pool, it is necessary to use QoS routing along with SPF as proposed by Nikolopoulos, et al.[16] and Guo, et al.[18]. This will necessitate routers to advertise the available bandwidth, delay jitter, delay, packet loss rate, and QoS requirements of traffic.

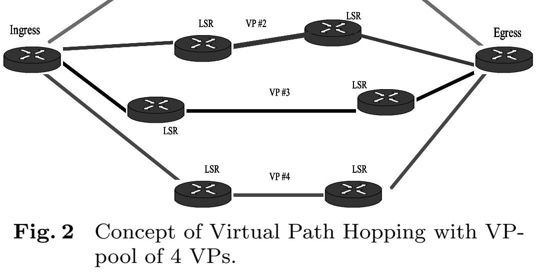


Fig. 2 Concept of Virtual Path Hopping with VP-pool of 4 VPs.

In this discussion, a VP-pool of n VPs has been considered. All VPs in the VP-pool should be ranked (from rank #1 to #n such that the most suitable VP is #1) taking factors such as the QoS they can provide, cost in terms of bandwidths, reliability, and the length of the VP into consideration. VPH does not reserve the resources of all VPs in the VP-pool as this is very inefficient. The ingress only does path computations when the VP-pool is formed and it does not do any label bindings. Resource reservation and the establishment of label binding for all VPs is done by exchanging PATH and RESV messages in RSVP-TE, just before that they are to be used by the intended traffic. Therefore, any best effort traffic (BET) that arrives after the VP-pool is formed can use any resources of unused VPs in that pool. This BET is preempted by PT, similar to conventional 1 : 1 protection, if there is a degraded failure in PT carried in an AVP. In other words, VPH reserves resources and uses only one VP at any given time and network resource utilization is not affected by its implementation.

At the beginning of the communication session, ingress will always start communication via rank #1 VP and it would then hop periodically to rank #2, rank #3 ...... rank #n, and then cyclically back to rank #1 and successive ranks assuming these VPs are neither used by other traffic nor have failed. If a VP is not free for VPH then the ingress will have to update its VP-pool with a new VP. The period of VPH is a vital parameter in this concept and for optimum results it should be less than the minimum threshold value of the timers used to detect control-plane failures as presented in the simulation results of one of our previous conference papers[14].

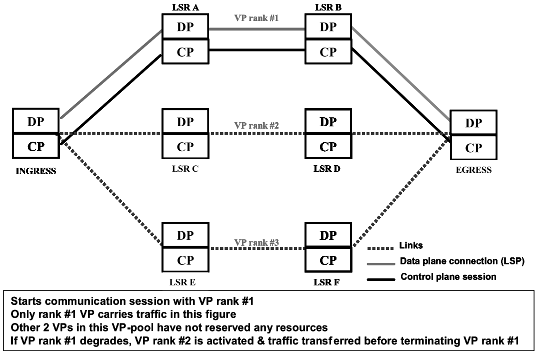
Once the ingress explicitly determine s the rank #1 VP, it first establishes the control-plane RSVP session for that VP and reserves network resources to guarantee the QoS of the data to be communicated. The control-plane-timer threshold values mentioned in Section II-A are usually determined by negotiating with neighboring peers, when the control-plane session is established. Therefore, these threshold values for the control-plane peers of the same session could vary. The period of VPH should be determined by the ingress after deciding all these threshold values and it should be less than the minimum threshold value for all peers in the session. For example, as outlined in Fig. 2, if the timer values for the three links of VP #1 are 40, 30 and 50 s, then the period of VPH should be less than 30 s (minimum of the three). It should be less than 30 s such that VPH can be performed before the 30-s control-plane timer expires. The timer-threshold-value information can be transmitted to the ingress with RESV messages in RSVP-TE. In this situation, the VPH concept will make sure it will change the AVP, before the control-plane timer expires and the RSVP session is torn down causing disconnection on the corresponding data plane. Obviously, the new VP is established and maintained by a new control-plane RSVP session. Since RSVP-TE always exchanges PATH and RESV messages between an ingress-egress pair to form a new control-plane session and a corresponding VP, it is clear that the sum of the round trip time (RTT) and processing time at each node is the transit time for a VP hop. Therefore, this transit time mainly depends on the distance or number of nodes between the ingress and egress. If necessary, it is possible to set priority as Key to all the PATH and RESV messages involved in VPH to reduce the transition time further. "Actively reserved bandwidth architecture"[7] can be used to rapidly allocate resources and minimize call rejections due to lack of resources. It is necessary to maintain a VP-pool table at the ingress, to implement VPH smoothly. Such a table should essentially include information such as VP ID, the VP rank and its status (i.e., active or inactive), apart from information on a conventional routing table for MPLS.

Whenever VPH is performed according to the above algorithm, there will be some overhead traffic added to the network. This added overhead can be minimized by minimizing the number of VP hops. Therefore instead of periodic VPH, one that is more efficient, where VP hops are triggered by a timer called a VPH_Timer, is evaluated in this study. This can be referred to as non-periodic VPH. Since this paper is focused on non-periodic VPH, VPH, after this, will refer to non-periodic VPH unless otherwise stated specifically. The VPH_Timer is reset to zero every time a VPH is done or a PDU is received. When RSVP-TE is used on the control plane, the RSVP Hello State Timer/Cleanup timer (T) of the control-plane peers is determined through negotiation with neighboring peers. Usually $T = k * R$; where R is the refresh period and k is an integer (according to IETF standards, by default, $k = 3$ or 4 and $R = 10$ s in most routers)[9),23)]. In the same way, the VPH_Timer values also can be determined by each control-plane peer that participates in VPH and these should be less than T. Whenever the VPH_Timer expires, T should be reset to zero and a Refresh_Timer should be started in the same way as the graceful restart in RSVP and LDP[21)]. This "Refresh" state is used to buy time to inform the ingress about the expiration of the VPH_Timer and to carry out the VP hop smoothly without loss of data. In other words, this "refresh" state helps to conduct the make-before-break concept without any data losses. Therefore, the data-plane communications of a degraded VP should not be terminated during the "refresh" state. The Refresh_Timer value should be determined based on the transit time for VPH, which mainly depends on the RTT between the ingress and egress, or more specifically;

$$\text{Refresh\_Timer} > \text{RTT} + \text{Process time at nodes.}$$

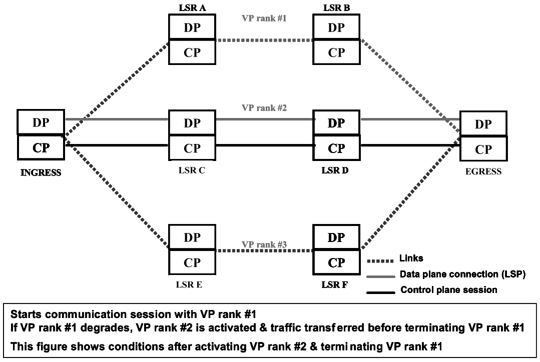
It is possible to use the Lost state of Hello messages to inform the ingress of the expiration of the VPH_Timer, as used in the graceful restart of RSVP-TE implemented by some widely used routers[24)]. The concept of VPH can further be explained using **Figs. 3**, **4**, and **5**. Figures 3 and 4 outline two different scenarios for a VP-pool with three VPs ranked as #1, #2, and #3. Figure 3 shows the beginning of a communication session, where VP #1 is active and traffic is transmitted through it. It also shows the LSP (top-line) and the corresponding control-plane session (bottom-line) of VP #1. According to this figure, VPs #2 and #3 (dashed lines) do not reserve any network resources. It is assumed that a degraded failure has occurred in active VP #1. Figure 5 is the time diagram for the use of make-before-break in VPH. According to the VPH concept, the VPH_Timer will always expire before the control-plane timer (T), as shown in Fig. 5. This is because the VPH_Timer < T and they start at the same time. The expiration of the VPH_Timer will trigger VPH and the ingress-egress pair will exchange PATH and RESV messages to activate the next best available VP (i.e., VP rank #2). Whenever the VPH_Timer expires, the Refresh_Timer is immediately started. The control-plane timer is not allowed to expire and tear down the control-plane session avoiding the termination of data communications on the data plane, when a degraded VP is in a "refresh" state. This is since the Refresh_Timer has been determined such that the value of the Refresh_Timer > (RTT + Process time at nodes) "refresh" state gives sufficient time to establish the new control-plane session to activate the new VP (VP #2) and transfer traffic as shown in Fig. 5. Once a smooth transfer has been completed, the Refresh_Timer is reset and the degraded VP is torn down, releasing its resources. Figure 4 shows that VP #2 is active and traffic goes through it after undertaking VPH. If VP #1 is repaired and healthy at the next expiration of the VPH_Timer, traffic is hopped to VP #1 as it is the most suitable in the VP-pool. If VP #1 is not available it is hopped to VP #3. This procedure of VPH is also summarized in the flow chart in **Fig. 6**.

If VPH is not done, as shown in **Fig. 7**, there



Starts communication session with VP rank #1
Only rank #1 VP carries traffic in this figure
Other 2 VPs in this VP-pool have not reserved any resources
If VP rank #1 degrades, VP rank #2 is activated & traffic transferred before terminating VP rank #1

Fig. 3 VP-pool with active VP rank #1 in VPH.



Starts communication session with VP rank #1
If VP rank #1 degrades, VP rank #2 is activated & traffic transferred before terminating VP rank #1
This figure shows conditions after activating VP rank #2 & terminating VP rank #1

Fig. 4 VP-pool with active VP rank #2 in VPH.

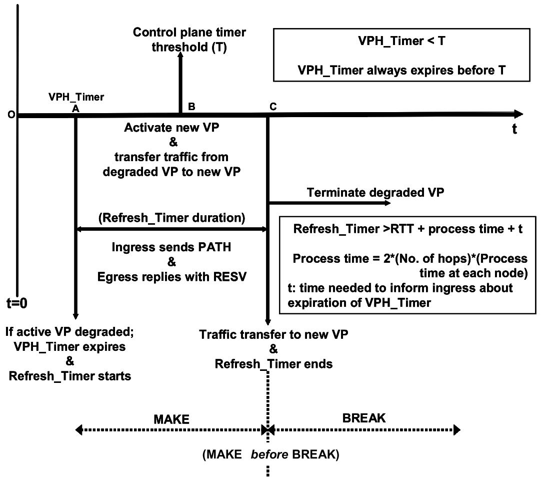
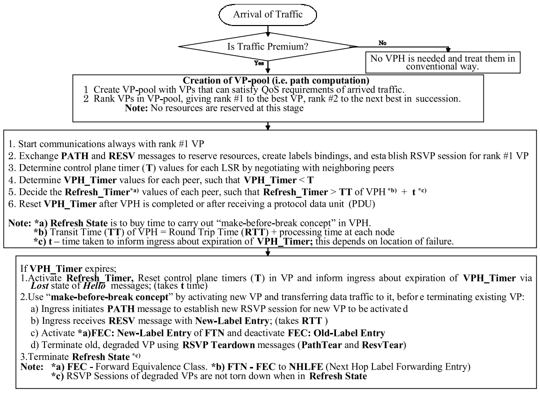

Control plane timer threshold (T)
VPH_Timer < T
VPH_Timer always expires before T
VPH_Timer
O
A
B
C
t
Activate new VP & transfer traffic from degraded VP to new VP
Terminate degraded VP
(Refresh_Timer duration)
Ingress sends PATH & Egress replies with RESV
Refresh_Timer >RTT + process time + t
Process time = 2*(No. of hops)*(Process time at each node)
t: time needed to inform ingress about expiration of VPH_Timer
t=0
If active VP degraded; VPH_Timer expires & Refresh_Timer starts
Traffic transfer to new VP & Refresh_Timer ends
MAKE
BREAK
(MAKE *before* BREAK)

Fig. 5 Time diagram for MAKE before BREAK.



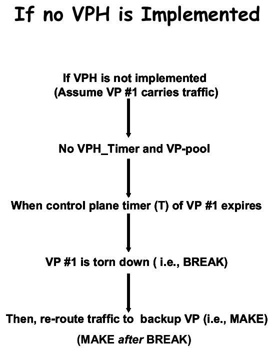
Creation of VP-pool (i.e. path computation)
1 Create VP-pool with VPs that can satisfy QoS requirements of arrived traffic.
2 Rank VPs in VP-pool, giving rank #1 to the best VP, rank #2 to the next best in succession.
Note: No resources are reserved at this stage

1. Start communications always with rank #1 VP
2. Exchange **PATH** and **RESV** messages to reserve resources, create labels bindings, and esta blish RSVP session for rank #1 VP
3. Determine control plane timer (**T**) values for each LSR by negotiating with neighboring peers
4. Determine **VPH_Timer** values for each peer, such that **VPH_Timer** < **T**
5. Decide the **Refresh_Timer***a) values of each peer, such that **Refresh_Timer** > **TT** of VPH *b) + **t** *c)
6. Reset **VPH_Timer** after VPH is completed or after receiving a protocol data unit (PDU)

Note: *a) Refresh State is to buy time to carry out "make-before-break concept" in VPH.
***b)** Transit Time (**TT**) of VPH = Round Trip Time (**RTT**) + processing time at each node
***c) t** – time taken to inform ingress about expiration of **VPH_Timer;** this depends on location of failure.

If **VPH_Timer** expires;
1.Activate **Refresh_Timer,** Reset control plane timers (**T**) in VP and inform ingress about expiration of **VPH_Timer** via ***Lost*** state of ***Hello*** messages; (takes **t** time)
2.Use "**make-before-break concept**" by activating new VP and transferring data traffic to it, befor e terminating existing VP:
a) Ingress initiates **PATH** message to establish new RSVP session for new VP to be activate d
b) Ingress receives **RESV** message with **New-Label Entry**; (takes **RTT**)
c) Activate ***a)FEC: New-Label Entry** of **FTN** and deactivate **FEC: Old-Label Entry**
d) Terminate old, degraded VP using **RSVP Teardown** messages (**PathTear** and **ResvTear**)
3.Terminate **Refresh State** *c)
Note: *a) FEC - Forward Equivalence Class. ***b) FTN - FEC** to **NHLFE** (Next Hop Label Forwarding Entry)
***c)** RSVP Sessions of degraded VPs are not torn down when in **Refresh State**

Fig. 6 Steps in virtual path hopping concept.



If VPH is not implemented (Assume VP #1 carries traffic)
No VPH_Timer and VP-pool
When control plane timer (T) of VP #1 expires
VP #1 is torn down (i.e., BREAK)
Then, re-route traffic to backup VP (i.e., MAKE)
(MAKE *after* BREAK)

Fig. 7 Without VPH implemented.

is no VPH_Timer to expire before the control-plane timer. Therefore, data-plane communications are terminated whenever the timer (T) to detect degraded failures has expired. There should then be re-routing to a BP to recover communications. In the event of re-routing, there will be data loss and VPH will eliminate re-routings on the data plane due to degraded failures that are identified by the control-plane timers. This will enable very high availability and reliability to be achieved for PT.

All this explanation of VPH has been for unidirectional LSP/VP. It is essential to have bi-directional VPs for interactive real-time communication and this can be done in three ways according to the existing protocols[27].

(1) Since basic MPLS LSP/VPs are unidirectional, it is possible to have two such VPs created separately (with separate control channels) for forward and backward directions. These are called asymmetric VPs as forward and backward directional VPs could follow two different physical paths. In this situation, the VPH concept can be separately applied to forward and backward directional VPs by having separate VP-pools. In other words a failure in a forward VP does not warrant a VPH in the corresponding backward directional VP or vice versa. Network administration tools could bind the two directional VPs and manage them as a single entity.

(2) The 'Upstream Label' object in the PATH message can be used and the procedures explained in GMPLS[32] can be followed to create bi-directional VPs by only exchanging two messages similar to unidirectional VPs. The VP-pool will then be bi-directional.

(3) A more secure solution is to follow the same procedure as in (2) above and use the RESV_CONF message in addition to PATH and RESV messages to confirm whether bi-directional VPs have been successfully formed.

Since (1), (2), and (3) are provided by existing protocols, the implementation of either of them in VPH is technically feasible and the most suitable depending on the network and the type of traffic could be selected.

VPH also has some other advantages. Since its main objective is to overcome terminations of data communications on the data plane due to degraded failures identified by the control-plane timers of connection-oriented networks that use in-band signaling, it is beyond the scope of this paper to evaluate all additional advantages in detail. However, some advantages such as providing backup paths to link/path failures, distributing traffic, and avoiding trap topologies are briefly described below. Only one VP in the VP-pool is used at any given time and if any of the remaining VPs are not used by other traffic one of them can be used as a BP for link/path failures. Therefore, it is possible to achieve fast re-routing without damaging the TE model, in the event of link/path failures. Also, as there is a VP-pool between each ingress and egress pair, it is possible to achieve 100% restoration even for dual-fault situations without disrupting the TE of the network. VPH can also dynamically distribute traffic well throughout the network. This evenly distributes the load throughout the network reducing the stress on links and nodes due to excessive congestion. This in turn will reduce the probability of failure in a node or link[15]. VPH can also avoid 'trap topologies', which cause problems with other existing BBW sharing techniques. A trap topology is a situation where it is impossible to find a link-disjoint BP for a certain AP at a given point of time. If this happens during a communication session, it will last throughout the session with the current setup, and this is very critical, especially for PT. Before a communication session begins in VPH, it determines the VP-pool mostly consisting of link-disjoint VPs, avoiding 'trap topologies'.

4. Numerical Analysis

The links in a network differ widely in their failure characteristics and a link-failure model should account for these[4]. In the general model considered here, it has been assumed that there are n different link-disjoint VPs in each VP-pool and each VP has H$_1$, $H_2 \ldots\ldots H_n$ links. The probability of failure is considered to be $p_{ij} = p^f_{ij} + p^d_{ij}$, where $i^{(1 \le i \le n)}$ is the number of paths, $j^{(1 \le j \le H_i)}$ is the number of links, p^f_{ij} is the probability of failure due to link/node failures, and p^d_{ij} is the probability of failure due to degraded failures. Since about 50% of the total failures are degraded, p^f_{ij} and p^d_{ij} would approximately be equal. The probability of having a failure in the nth VP is given by

$$1-\prod_{j=1}^{H_n}(1-(p_{nj}^{f}+p_{nj}^{d})) \qquad (1)$$

If VPH is not implemented and no failure occurs, any communication session can use the same VP throughout the entire session. The probability of path failure in such a session, $\mathrm{P_{No\_VPH}}$, is given by Eq. (1). If periodic VPH is implemented, on the other hand, where there are K VP hops during a session, the time average for the probability of path failure $\mathrm{P_{VPH}}$ during the period of communication is given by

$$1/K\sum_{i=1}^{K}\left[1-\prod_{j=1}^{H_i}(1-(p_{ij}^{f}+p_{ij}^{d}))\right]$$

for $K < n$ and

$$1/n\sum_{i=1}^{n}\left[1-\prod_{j=1}^{H_i}(1-(p_{ij}^{f}+p_{ij}^{d}))\right]$$

for $K \geq n$

If $K > n$, the VPH is done cyclically and therefore each path is used more than once as explained in the previous section. The VPH concept will eliminate almost all degraded type failures, making p_{ij}^{d} very small (almost zero). Therefore, $P_{VPH} < P_{No\_VPH}$. This demonstrates the improved reliability and availability of the network due to the VPH concept.

5. Performance Evaluation

The performance of VPH and its contributions to traffic engineering were evaluated through computer simulations. Different network topologies with nodes 10, 20, 40, 50, 60, and 90 were simulated for many combinations of failures. Random graphs were used to determine the network topologies and there is an example of such a network in **Fig. 8** (not all links have been shown for simplicity). The simulation results indicated similar patterns and therefore the results for the topologies in **Table 1** simulated for the failure combinations in **Table 2** are presented here. Over 30 connectivity orientations for topologies were simulated for the same failure combination. All these scenarios were simulated for a prolonged period of three years. Only the simulations done for non-periodic VPH and without VPH scenarios are presented here as the simulation results for periodic VPH have been discussed and presented in our previous conference papers[14),33)].

5.1 Simulation Model and Parameters

In all the simulations that were done, the following simple algorithm was used to determine the VP-pool. This algorithm was utilized because it is similar to the QoS routing algorithms followed by most MPLS-TE supported routers on the market today. First, links that did not have sufficient resources to support the required QoS were pruned off. Then, Dijkstra's[11)] shortest path algorithm was used on the remaining topology to find the best path. Once the best VP had been selected the links were pruned off and the same procedure was done on the rest of the network to determine the next best VPs in the VP-pool. A VP-pool of three VPs was considered in all simulations for the sake of simplicity. If it is not possible to find three link-disjoint paths, the least overlapping best VPs can be determined in a similar way to the algorithm in Nikolopoulos, et al.[16)]. For simplicity and better comparability, 10 ingress/egress pairs were chosen for each scenario, such that a link-disjoint VP-pool could be derived. Whenever a network failure occurred on the data plane, fast re-routing was done to restore it and an unused VP from the VP-pool was always used as a BP in all simulations. Therefore, the number of re-routings

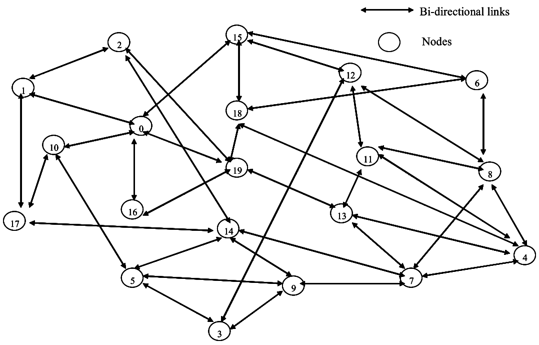


Fig. 8 Example random network with 20 nodes.

Table 1 Simulated network topologies.

Network	*No. of nodes*	*No. of bidirectional links*
A	20	68
B	90	270

Table 2 Different failure combination scenarios (simulated).

Combination	***Failures/link/month***	***Failures/node/month***
I	0.01	0.01
II	0.01	0.1
III	0.05	0.1
IV	0.05	0.01

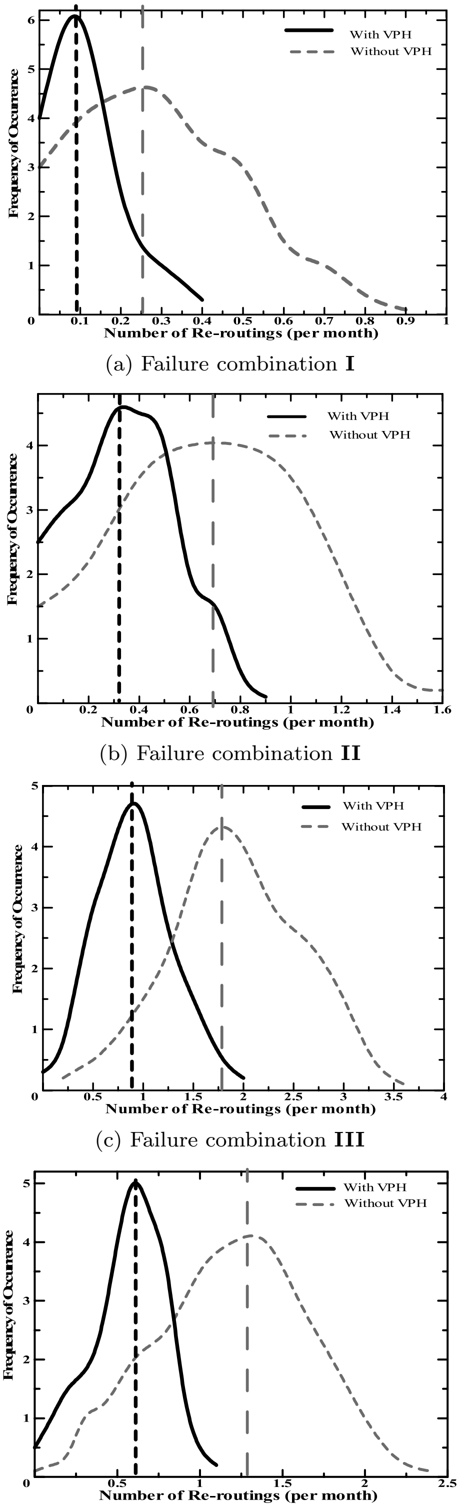
done will be the count for network failures and this was used as the measure to evaluate the performance of the VPH concept.

The arrival of traffic flow and the duration of communication sessions were randomly determined according to Poisson and exponential distributions, respectively. The bandwidth of the sessions was also randomly determined. Many different average values such as 3, 5, 10, 30, 60, and 300 s for traffic flow arrivals, 300, 600, 900, 1,800, and 3,600 s for session duration, and 1, 5, 10, and 20 Mbps for session bandwidth were simulated. We found that the improvement due to the VPH concept was not very sensitive to these average values and therefore the results presented here are for the 10 s, 1,800 s and 10 Mbps scenario. The number of sessions for any traffic flow arrival was randomly determined to be any value from 1 to 10.

Three major types of failures, i.e., hardware and software failures, failures due to burst losses and congestion, and failures detected by control-plane timers as previously discussed were considered in all simulations. The failure arrivals for all types were assumed to be distributed exponentially and their averages were determined case by case, based on the values in Table 1 and Table 2. According to the many simulations that were done, the parameters for the distributions of repair times for failed links were found to have very negligible effects on the performance of VPH. Therefore, the repair time after a link failure was assumed to be a constant. VPH_Timer values were randomly determined for each peer to be a multiple of 10 s in a range from 30–80 s.

5.2 Analysis of Simulated Results:

Figures 9 and **10** plot the frequency of occurrence versus the number of re-routings per month for topologies A and B. The vertical dashed lines in these graphs indicate the highest frequency of occurrence for re-routings and we can see that most occurrences are concentrated around these dashed lines as expected. According to these results, VPH always reduces this highest frequency of occurrence for re-routings by about 50% irrespective of the topology or the failure probability of nodes and links. This reflects a reduction in failures on the data plane and accounts for almost 50% of degraded failures, which were detected by control-plane timers so that they could be eliminated by VPH. Similar graphs were obtained for network topologies with 10, 40, 50, and 60 nodes



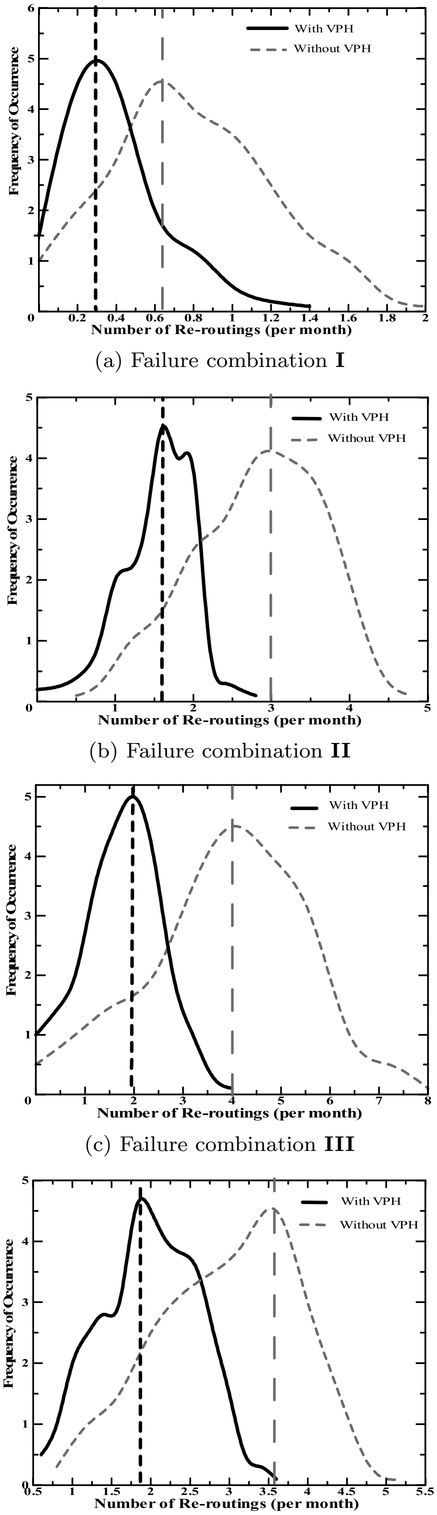
(a) Failure combination **I**

(b) Failure combination **II**

(c) Failure combination **III**

(d) Failure combination **IV**

Fig. 9 Frequency of occurrences vs. number of re-routings for Topology **A**.



(a) Failure combination **I**

(b) Failure combination **II**

(c) Failure combination **III**

(d) Failure combination **IV**

Fig. 10 Frequency of occurrence vs. number of re-routings for Topology **B**.

and all these graphs indicated similar 50% improvements irrespective of the network topology or failure probabilities.

Traffic distribution analysis was done by monitoring the traffic on all nodes and links of the network over a period of 24 hours. Simulations were conducted to compare the traffic distribution due to the non-periodic VPH concept with no VPH for many network topologies. As expected, we noted that the implementation of VPH improved the traffic distribution of the network irrespective of its size because it used a VP-pool between each ingress/egress pair. A better dynamic traffic distribution could be achieved for networks with higher failure combinations and that had prolonged communication sessions. The bar chart in **Fig. 11** plots the results for network topology A with a failure combination of III as in Table 2. It is very clear that the traffic that travels through eight nodes without VPH, distributes the same traffic among 14 nodes with VPH being implemented. Similarly **Fig. 12** shows the traffic distribution that can be achieved over many links by implementing the VPH concept. In the figure, the traffic carried by only eight links when no VPH is implemented is distributed among 18 links, with the implementation of VPH.

Without the special TE algorithm, routers would usually use the SPF algorithm and this would make some links overused while some others would not be used at all. In VPH, the VP-pool is always determined using not only SPF but also using QoS constraints and explicit routing. Therefore, implementing VPH provides a very good traffic distribution even when no special TE algorithms are used. The

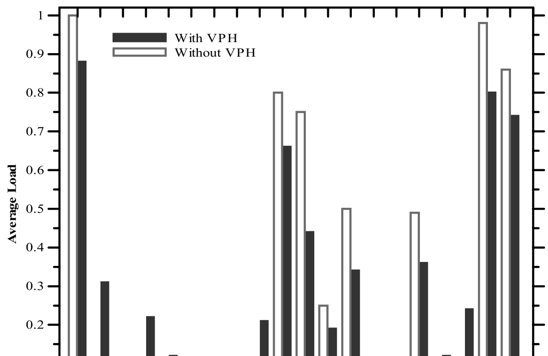


Fig. 11 Variations in average load with respect to number of nodes.

traffic distribution reduces the stress on the network and improves its robustness against failures. The changes in network-utilization efficiency due to the implementation of VPH were measured. **Figure 13** indicates the differences in utilization efficiency (with and without VPH) versus load. Here, the difference in utilization efficiency is defined as utilization efficiency with VPH — utilization efficiency without VPH. Using load values ranging from 0.1 to 0.9, the utilization efficiency of network resources was measured with and without VPH being implemented. As we expected the graph shows there is not much difference (less than ±5%) between them. This means that the implementation of VPH does not affect the utilization efficiency of network resources. This is because at any given time VPH uses only one VP and its resources, even though it determines the VP-pool.

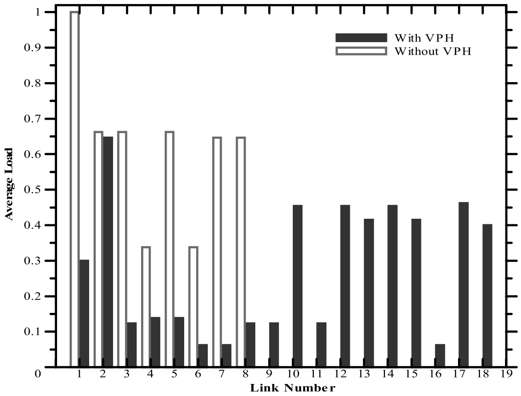


Fig. 12 Variations in average load with respect to number of links.

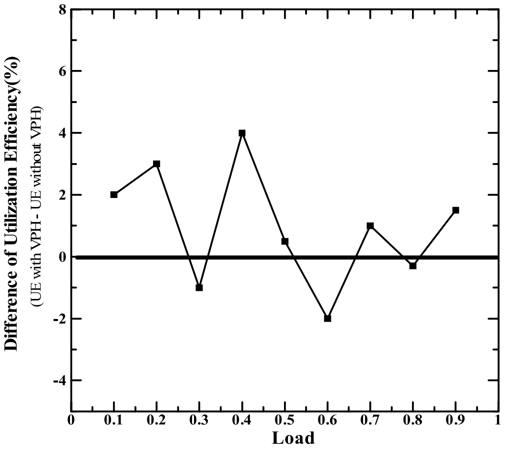


Fig. 13 Difference in utilization efficiency vs. load (for VPH).

6. Conclusions

VPH, a concept to protect data-plane communications from degraded failures detected by the control-plane timers of connection-oriented packet networks such as IP/MPLS, was extensively discussed in this article. According to the simulation results, almost all such failures, which account for about 50% of the total, could be eliminated as expected. As well as the concept of VPH, its numerous advantages such as efficient utilization of network resources through better traffic distribution, reduced stress on networks, very fast re-routing in the event of link/path failures, and robustness against dual failures in networks were also discussed. Periodic VPH could be somewhat inefficient as very frequent VPHs may add unfriendly overhead traffic to the network. Therefore, the focus of this paper was on a more efficient, non-periodic VPH and its evaluation. The average improvements to performance of the non-periodic VPH presented here and that of the periodic VPH presented in our previous conference papers [14),33)] were almost the same.

For the periodic VPH with a period of 30 s, the average number of VPHs per hour was 100, whereas for the non-periodic VPH it was about 25 VPHs per hour. Therefore, non-periodic VPH is very efficient compared to periodic VPH as it eliminates about 75% of the VP hops, without affecting performance. Therefore, it is possible to conclude that the non-periodic VPH proposed here is a very promising efficient and proactive technique to minimize the occurrence of failures in PT.

VPH will be very useful in the near future as the Internet will depend more on connection-oriented networks, which are inherently vulnerable to degraded failures. Therefore, it would be very interesting to investigate the possibilities of testing the VPH concept in a real network using real-time traffic in the future. This concept should also be evaluated in upper layers such as the application layer before future implementation. This paper was focused on connection-oriented packet networks such as IP/MPLS, where in-band signaling is used as previously discussed. In future work, we wish to evaluate the VPH concept in any connection-oriented networks such as GMPLS, where out-of-band signaling is also permitted.

References

1) Rosen, E., Viswanathan, A. and Callon, R.: Multi Protocol Label Switching Architecture, *IETF* RFC 3031 (Jan. 2001).

2) Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V. and Swallow, G.: RSVP-TE: Extensions to RSVP for LSP Tunnels, *IETF*, RFC 3209 (Dec. 2001).

3) Sharma, V. and Hellstrand, F.: Framework for Multi-Protocol Label Switching (MPLS)-based Recovery, *IETF*, RFC 3469 (Jan. 2003).

4) Iannaccone, G., Chuah, C., Mortier, R., Battacharyya, S. and Diot, C.: Analysis of link failures in a IP backbone, *Proc. Internet Measurement Workshop 2002* (2002). http://www.icir.org/vern/imw-2002/imw2002-papers/202.pdf

5) Wu, J., Montuno, D.Y., Mouftah, H.T., Wang, G. and Dasylva, A.C.: Improving the Reliability of the Label Distribution Protocol, *Proc. 26th Annual IEEE Conference on Local Computer Networks* (2001).

6) Brittain, P.: MPLS Traffic Engineering: A Choice of Signaling Protocols, White Paper of Data Connection Limited (Jan. 2000). http://www.dataconnection.com

7) Kuo, G. and Lai, C.T.: A new architecture for transmission of MPEG-4 video on MPLS networks, *ICC 2001 - IEEE International Conference on Communications*, No.1, pp.1556–1560 (June 2001).

8) Xiong, Y. and Mason, L.G.: Restoration strategies and spare capacity requirements in self-healing ATM networks, *IEEE/ACM Transactions on Networking*, No.1, pp.98–110 (Feb. 1999).

9) Braden, R., Zhang, L., Berson, S., Herzog, S. and Jamin, S.: Resource ReServation Protocol (RSVP), *IETF*, RFC 2205 (Sep. 1997).

10) Autenrieth, A. and Kirstäter, A.: Fault-Tolerance and Resilience Issues in IP-Based Networks, *2nd Int'l. Wksp. Design Reliable Commun. Net.* (*DRCN2000*), Germany (Apr. 2000).

11) Dijkstra's shortest path algorithm. http://en.wikipedia.org/wiki/Dijkstra%27s_algorithm

12) Huang, C., Sharma, V., Owens, K. and Makam, S.: Building Reliable MPLS Networks Using a Path Protection Mechanism, *IEEE Communications Magazine*, pp.156–162 (Mar. 2002).

13) Milind, L.L., Buddhikot, M., Chekuri, C. and Guo, K.: Routing Bandwidth Guaranteed Paths with Local Restoration in Label Switched Networks, *Proc. 10th IEEE International Conference on Network Protocols 2002* (2002).

14) Gamage, M., Hayasaka, M. and Miki, T.: Virtual Path Hopping to overcome Network Failures due to Control Plane Failures in Connection Oriented Networks, *Proc. Joint Conference of the 10th APCC and the 5th MDMC*, China (Aug. 2004).

15) Weichenberg, G., Chan, V.W.S. and Médard, M.: A Reliable Architecture for Networks under Stress, *Fourth International Workshop on the Design of Reliable Communication Networks* (*DRCN*), Canada (Oct. 2003).

16) Nikolopoulos, S.D., Pitsillides, A. and Tipper, D.: Addressing Network Survivability Issues by Finding the K-best Paths through a Trellis Graph, *Proc. IEEE INFOCOM*, Japan (Apr. 1997).

17) Szviatovszki, B., Szentesi, A. and Juttner, A.: On the Effectiveness of Restoration Path Computation Methods. http//www.cs.elte.hu/~alpar/publications/proc/RestoPath.pdf

18) Guo, Y., Kuipers, F. and Mieghem, P.V.: Link-Disjoint Paths for Reliable QoS Routing. http://www.nas.its.tudelft.nl/people/Piet/papers/dimcra.pdf

19) Bejerano, Y., Breitbart, Y., Orda, A., Rastogi, R. and Sprintson, A.: Algorithms for Computing QoS paths with Restoration. http://www.paradise.caltech.edu/~spalex/pub/conferences/infocom03.pdf

20) Liu, G., Yang, Y. and Lin, X.: Performance Evaluation of K Shortest Path Algorithms in MPLS Traffic Engineering, *IEICE Trans. Commun.*, Vol.E87-B, No.4, pp.1007–1010 (Apr. 2004).

21) Leelanivas, M., Rekhter, Y. and Aggarwal, R.: Graceful Restart Mechanism for LDP, RFC 3478 (Feb. 2003).

22) Farrel, A.: Fault Tolerance for the Label Distribution Protocol (LDP), RFC 3479 (Feb. 2003).

23) Pan, P., Swallow, G. and Atlas, A.: Fast Reroute Extensions to RSVP-TE for LSP Tunnels. http://www.ietf.org/internet-drafts/draft-ietf-mpls-rsvp-lsp-fastreroute-07.txt

24) Cisco product features: MPLS Traffic Engineering-RSVP Graceful Restart. http://www.cisco.com/en/US/products/sw/iosswrel/ps1829/products_feature_guide09186a008027db17.html

25) Clouqueur, M. and Grover, W.D.: Availability Analysis of Span-Restorable Mesh Networks, *IEEE Journal on Selected Areas in Communications*, Vol.20, No.4 (May 2002).

26) Ogino, N. and Tanaka, H.: Routing and Re-Routing of Reliable Label Switched Paths with

Variable Bandwidths in MPLS over Optical Networks, *IEICE Trans. Commun.*, Vol.E87-B, No.7, pp.1834–1843 (2004).

27) Jerram, N. and Farrel, A.: MPLS in Optical Networks, White Paper of Data Connection Ltd, http://www.dataconnection.com.

28) Aboul-Magd, O.: The Documentation of IANA Assignments for Constraint-Based LSP Setup Using LDP (CR-LDP) Extensions for Automatic Switched Optical Network (ASON), *IETF* RFC 3475 (Mar. 2003).

29) Boutremans, C., Iannaccone, G. and Diot, C.: Impact of Link Failures on VOIP Performance, *Proc. NOSSDAV '02*, USA (May 2002).

30) Andersson, L. and Swallow, G.: The Multi Protocol Label Switching (MPLS) Working Group Decision on MPLS Signaling Protocols, *IETF* RFC 3468 (Feb. 2003).

31) Lang, J.: Link Management Protocol (Oct. 2003). http://www.ietf.org/internet-drafts/draft-ietf-ccamp-lmp-10.txt

32) Mannie, E.: Generalized Multi-Protocol Label Switching (GMPLS) Architecture, *IETF*, RFC 3945 (Oct. 2004).

33) Gamage, M., Hayasaka, M. and Miki, T.: Implementation of Virtual Path Hopping (VPH) as a Solution for Control Plane Failures in Connection Oriented Networks and an Analysis of Traffic Distribution of VPH, *Proc. 3rd International Workshop on QoS in Multi-service IP Networks* (*QoS-IP 2005*), Italy (Feb. 2005).

(Received September 30, 2005)
(Accepted May 9, 2006)
(Online version of this article can be found in the IPSJ Digital Courier, Vol.2, pp.452–464.)

Manodha Neilendra Gamage received his B.Sc. (Eng) in Electronics and Telecommunications from the University of Moratuwa, Sri Lanka in 1998. He received his M.E. from the University of Electro-Communications, Tokyo, Japan in 2003. He is currently a Ph.D. candidate at the University of Electro-Communications, Tokyo, Japan and expects to graduate in 2006. He joined Sri Lanka Telecom in 1998 as a telecommunications engineer, where he was involved in implementing ISDN-network and real-time video conferencing applications in Sri Lanka. His research interests include real-time interactive applications over the Internet such as VOIP, IP networks, and connection-oriented networking such as MPLS. He is a member of IEEE.

Mitsuo Hayasaka received his B.E. and M.E. from the University of Electro-Communications, Tokyo, Japan in 2000 and 2002. He is currently a Ph.D. student at the same university. His research interests involve QoS controls of real-time multimedia communications, peer-to-peer multimedia streaming, and multicasting techniques. He is a member of IEEE, IEICE, and IPSJ.

Tetsuya Miki received his B.E. from the University of Electro-Communications,Tokyo, Japan in 1965, and his M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1967 and 1970. He joined the Electrical Communication Laboratories of NTT in 1970, where he was engaged in research and development on high-speed digital transmission systems using coaxial/optical cable, fiber-to-the-home systems, ATM transport networks, and network operation systems. He was the Executive Manager of NTT's Optical Network Systems Laboratories from 1992 to 1995. He is currently a Professor at the Department of Information and Communication Engineering at the University of Electro-Communications. His current research interests include broadband wireless/fixed network architectures, photonic networks, and their applications. He received the IEICE Achievement Award in 1978 for his contribution to the early development of optical transmission systems. He is a Fellow of IEEE, and was a Vice-President of the IEEE Communications Society from 1998–1999. He is also a Fellow of IEICE, and was a Vice-President of IEICE from June 2003 to May 2005.