Network Working Group
Internet Draft
Category: Informational
Expires: May 2004

Adrian Farrel
Old Dog Consulting

November 2003

Interim Report on MPLS Pre-emption

draft-farrel-mpls-preemption-interim-00.txt

Status of this Memo

Abstract

This document is an interim report into pre-emption in MPLS systems.

At the 58th IETF, the MPLS Working Group determined that there are several interpretations of how pre-emption should be achieved within MPLS systems. This document is the result of the initial enquiries to vendors and other implementors into the question of how and why they offer pre-emption funtion in their MPLS inplementations.

This document is intended to be a short-lived document and only exists to document the current findings. It will be superceded or retired.

1. Introduction

The most recent charter of the MPLS working group includes the work item to develop a solution for "soft pre-emption" within MPLS systems. This work item was contested by some people who regard MPLS signaling (and in particular RSVP-TE as documented by [RFC 3209]) to already include mechanisms for soft pre-emption.

A common understanding of the correct behavior during pre-emption will be beneficial to vendors trying to get their hardware to inter-work, and to network operators trying to offer services in networks built from equipment from more than one vendor.

Without commenting on the correct interpretation of pre-existing documents, this document aims to establish the following.

- Behavioral characteristics of Admission Control in MPLS and GMPLS systems.
- Desired characteristics of pre-emption.
- A terminology for soft and hard pre-emption.
- A record of implemented pre-emption behavior.

If necessary, a subsequent document will describe the conclusions of the MPLS working group with regard to the correct interpretation of the procedures for pre-emption.

2. Background

2.1 Default PathErr Processing

[RFC2205] defines RSVP procedures including the procedures for handling PathErr messages that are used to report events or errors from a downstream node to an upstream node. The predominant use of PathErr in [RFC2205] is to report a problem that occurs while an RSVP path is being established. There is no specific text that refers to the use of a PathErr once a path has been established, but [RFC2205] does include the following text.

> PathErr messages are very simple; they are simply sent upstream to the sender that created the error, and they do not change path state in the nodes though which they pass.

2.2 The Origins of Priority in RSVP

[RFC2751] introduces the concept of priority to RSVP and suggests how it may be signaled using the Policy Data object.

Priority can be used in this context to modify the traditional first-come first-served Capacity-based Admissions Control (CAC) into a Priority-based Admisssions Control (PAC).

[RFC2751] concentrates mainly on PAC and says very little about the operation of pre-emption. The following text is relevant.

> When a previously admitted flow is preempted, a copy of the preempting flow's PREEMPTION_PRI element is sent back toward the PDP that originated the preempted PREEMPTION_PRI object. This PDP,

> having information on both the preempting and the preempted priorities may construct a higher priority PREEMPTION_PRI element in an effort to re-instate the preempted flow.

However, [RFC2750] describes the processing of policy errors as follows.

> Policy errors are reported by either ResvErr or PathErr messages with a policy failure error code in the ERROR_SPEC object.

2.3 Priority and Pre-emption in MPLS

[RFC3209] defines RSVP-TE and introduces the concept of setup and holding priorities for LSPs. This enables the implementation of PAC in MPLS systems, and facilitates pre-emption of a lower priority LSP by a higher priority LSP.

Pre-emption in this case means that some or all of the system resources previously reserved for the lower priority LSP are assigned to the higher priority LSP.

[RFC3209] contains the following text.

> When a new Path message is considered for admission, the bandwidth requested is compared with the bandwidth available at the priority specified in the Setup Priority.
>
> If the requested bandwidth is not available a PathErr message is returned with an Error Code of 01, Admission Control Failure, and an Error Value of 0x0002. The first 0 in the Error Value indicates a globally defined subcode and is not informational.

The 002 indicates "requested bandwidth unavailable".

If the requested bandwidth is less than the unused bandwidth then processing is complete. If the requested bandwidth is available, but is in use by lower priority sessions, then lower priority sessions (beginning with the lowest priority) MAY be preempted to free the necessary bandwidth.

When preemption is supported, each preempted reservation triggers a TC_Preempt() upcall to local clients, passing a subcode that indicates the reason. A ResvErr and/or PathErr with the code "Policy Control failure" SHOULD be sent toward the downstream receivers and upstream senders.

3. Ambiguity

Confusion and diverse implementations have arrisen owing to the lack of definitive instructions for error handling within [RFC3209]. The diversity of implementations has been driven both by assumed interpretations of [RFC3209] and by the needs of specific Admission Control implementations.

4. Admission Control

The implementation of Admission Control turns out to be a major factor in the decision about how to implement pre-emption within an MPLS system.

4.1 Capacity-based Admission Control (CAC)

CAC operates by allocating system resources to flows (for example, LSPs) as a function of the bandwidth requested during signaling and the capacity of the links or router that supports the flow. Each router is responsible for managing its own resources.

When a request for more bandwidth than can be supported by the links of router is receieved, the CAC request fails and a signaling error is returned. The CAC request may fail because the requested resources exceed the total available in the system. The request may also fail if some resources have already been reserved for other flows meaning that the requested resources exceed the currently avilable resources on the links or router.

4.2 Policy-based Admission Control (PAC)

PAC is a modification of CAC that allows the available resources to be subdivided by priority. This means that some of the resources can be reserved for use only by higher priority flows. High priority flows can use high and low priority resources, but low priority flows can use only low priority resources.

4.3 Pre-emption

PAC facilitates pre-emption. A resource request for a high priority flow may pre-empt a low priority flow, and commandeer its resources. The low priority flow is usually notified through signaling.

4.4 Best-Effort Traffic

Best-effort traffic does not have any resources specifically reserved for it. It is sent on the understanding that if there is no capacity at some point in the network it will be discarded. If, however, there are resources that are available either because they have not been reserved or because the flow for which they were reserved is not using them, then the best-effort traffic may get through.

4.5 Statistical Admissions Control

Some implementations of Admissions Control are statistical. This means that CAC requests are managed against a count of available resources. When a CAC request is successful the count of available resources is decremented.

PAC may also be managed in a statistical model.

In a statistical Admissions Control implementation no resources are specifically assigned to each flow, and there is no attempt to police the flows at transit routers. It is a requirement of successful operation that the edge nodes adhere to the implicit contract of

their resource requests. Policing is sometimes applied at the network edges or through management applications.

A consequence of this is that if one flow exceeds its contract, other flows will suffer.

It is hard to support best-effort traffic in a statistical CAC system.

4.6 Per-flow Admission Control

Per-flow Admission Control operates as CAC or PAC, and explicitly assigns resources for the use by the flow. This means that a flow that stays within its contracted limits is guaranteed resources to meet its contract regardless of the behavior of the other flows in the system.

A flow that exceeds its contracted limits may see its excess traffic discarded. However, if there are unused or unallocated resources in the system, the excess traffic may use these.

Similarly, best-effort traffic can be supported in this Admission Control model since it is clear that that traffic should only be allowed when there are spare resources.

5. Pre-emption Models

There are two pre-emption models applicable to MPLS systems.

5.1 Soft Pre-emption

In soft pre-emption, a higher priority LSP commandeers the resources previously assigned to a lower priority LSP. The lower priority LSP is not torn down and can continue to forward traffic on a best-effort basis.

Note that resource reservations on other LSRs are not changed, and the degradation to best-effort happens only at the point of pre-emption.

A notification is normally sent to upstream and downstream LSRs to warn them that the expected levels of service have been disrupted at one LSR along the LSP. This allows end-to-end or local repair to be performed to re-instate the desired level of service.

5.2 Hard Pre-emption

In hard pre-emption, a higher priority LSP commandeers the resources previously assigned to a lower priority LSP, and the lower priority LSP is torn down at the point of pre-emption. That is, the labels are released and the entry is removed from the LFIB. Any traffic that arrives with the old label is black-holed.

A notification message is sent to upstream and downstream LSRs to inform them of this event and, if the notified LSR is unable or unwilling to perform local repair, it may also tear the LSP by releasing resources and labels and forwarding the notification.

5.3 Head-end Teardown

Note that an ingress LSR that is informed of soft pre-emption may respond by explicitly tearing down the pre-empted LSP. Although this results in the release of labels and disruption of data traffic, it is not counted as hard pre-emption. The distinction between soft and hard pre-emption is made only at the LSR where the pre-emption occurs, and only at the time of pre-emption.

5.4 Applicability of Pre-emption Models

It will be observed that since the soft pre-emption model degrades the pre-empted LSP to best-effort, it is not well-suited to statistical Admission Control. In this mode hard pre-emption is more applicable.

Soft pre-emption can be used in per-flow admission control.

5.5 Methods of Notification

In the current confusion about the correct method of performing pre-emption both soft and hard pre-emption implemtations use a PathErr and ResvErr message containing the "Policy Control failure" error code to notify upstream and downstream LSRs of the pre-emption event.

There is no way to distinguish from the received message whether the pre-emption point performed soft or hard pre-emption.

5.6 Interworking of Pre-emption Models

The two pre-emption models do not interwork satisfactorily with the current usage of the same messages and error codes.

5.6.1 Hard Pre-emption in a Soft Pre-emption Network

Consider a pre-emption point that applies hard pre-emption in a network of LSRs that assume soft pre-emption. The other LSRs will assume that the LSP is up and will continue to send traffic which will be black-holed.

The Path refresh from the LSR immediately upstream of the pre-empting LSR will be rejected with an error of "Admission Control failure" because it will be seen as a new LSP setup request. This error might trigger the tear-dwon of the LSP or might simply be propagated to the ingress.

The pre-empting LSR will cease to send Resv refreshes so the LSP will eventually timeout from the upstream LSRs.

Similarly, the downstream LSRs will not receive Path refreshes and may receive a ResvErr or ResvTear in response to its Resv refresh.

In this case there is heavy reliance on the ingress LSP deciding that best-effort is not good enough. It must either re-route or tear the LSP.

5.6.2 Soft Pre-emption in a Hard Pre-emption Network

If the pre-emption point performs soft pre-emption and signals this event upstream and downstream to the remainder of the network, and if the remainder of the network assumes that hard pre-emption has been performed then the worst that happens is that labels are wasted at the pre-emption point.

Upstream of the pre-emption point, the notification of pre-emption causes the release of state, rsources and labels. This means that no more traffic will be passed to the pre-emption point on the pre-empted LSP. Also, no Path refresh will be sent, and the Resv refresh received from the pre-emption point may be rejected with a ResvErr or ResvTear. In time, the pre-emption point will time-out the LSP and release the labels.

Downstream of the pre-emption point, the Path refresh from the pre-emption point may cause the LSP to be re-established. This partial LSP will never carry traffic and will be removed when the pre-emption point times-out the LSP as described above.

6. Tying Resources to Labels

An LSP is defined by the existence of entries in the Label Forwarding Information Base (LFIB) at each LSR along the LSP. These entries map {incoming interface, incoming label} to {outgoing interface, outgoing label}.

6.1 Packet-Based Networks

An LSP may exist with or without reserved resources at each or any LSR along its path. An LSP with no resources reserved at any point is described as 'best effort'. An LSP with no resources reserved at a particular LSR or on a particular link is best effort through that LSR or on that link.

In an MPLS network, a distinction should be drawn between the pre-emption of resources (such as buffers) and the pre-emption of labels. It is not a requirement that the release of resources forces a release of label. The LSP may survive pre-emption with its resources removed (best effort) or reduced (degraded).

Packet LSRs that are incapable or unwilling to offer best effort

LSPs, MAY choose to offer only hard pre-emption. This might arise where admission control is purely an arithmetic accounting function, and edge nodes are required to police flows.

Where soft pre-emption is appplied, it is usual to only release resources at the LSRs where pre-emption occurs. That is, the resources are left assigned to the LSP at all other LSRs.

6.2 Non-Packet Networks

For non-packet switching types (such as lambda switching) there is a hard binding between resources and labels. In these cases, it is not possible to pre-empt resources without releasing the label.

Hard pre-emption is usually applied in non-packet networks.

6.3 Separating Signaling State from LSP State

Some implementations may choose to retain signaling state for an LSP even when the labels have been released during hard pre-emption.

This may be particualrly useful in optical networks where resources are manually deprovisioned and reprovisioned.

7. Methods of Signaling Pre-emption

The poll of vendors and implementors has shown that several methods for signaling pre-emption events have been implemented or considered for implementation. These are presented below without comment upon their efficacy or properness.

7.1 PathErr and ResvErr with Policy Control Failure Error Code

As described in [RFC3209] the pre-emption point sends a PathErr upstream and a ResvErr downstream. The messages carry the "Policy Control failure" error code.

This mechanism is used in both hard and soft pre-emption implementations. In some implementations the ingress automatically responds to the PathErr with a PathTear.

7.2 PathErr with State Removal

Using the mechanism described in [RFC3473] a PathErr is sent upstream carrying the "Policy Control failure" error code and with the Path_State_Removed flag set. At the same time, a PathTear is sent downstream.

This mechanism is used in hard pre-emption implementations.

7.3 ResvTear and ResvErr

A ResvTear is sent upstream and a ResvErr carrying the "Policy Control failure" error code is sent downstream. The ingress automatically responds to the ResvTear with a PathTear.

This mechanism has been suggested for use in hard pre-emption implementations.

7.4 Notify Message

The Notify message introduced in [RFC3473] is sent upstream and downstream carrying the "Policy Control failure" error code. The Admin_Status object is included with the D-bit set to indicate that LSP teardown is required.

This mechanism has been suggested for use in hard pre-emption implementations.

7.5 Resv Message with Record Route Object

A Resv message is sent upstream with a flag in the RRO subobject conveying the information about pre-emption for a specific hop.

This mechanism is introduced in [SOFT-PREEMPT] to provide a way to perform soft pre-emption in systems that use the technique described in section 7.1 to perform hard pre-emption.

8. Error Codes and State Removal

The Addmission Control error code in [RFC2205] is defined as carrying an Error Value of the form ssur cccc cccc cccc where it is mandated that u = 1 to denote

u = 1: RSVP may use message to update local state and forward
the message. This means that the message is informational.

Note that a ResvErr may carry the InPlace flag in the ERROR_SPEC object to indicate that the resources are still in place. This is of no particular help during pre-emption.

GMPLS [RFC3473] introduced the Path_State_Removed flag for inclusion in PathErr messages to indicate to an upstrem LSR that the downstream LSR has completely removed the Path state for the LSP. The implication being that the resources have been freed, the labels released, and the LSP torn down.

9. A Side Note on PathErr Processing

One other factor should be brought into consideration: How is a PathErr message handled when it is received for an established LSP and the PathErr carries an error code other than "Policy Control failure"?

10. Other Requirements

Two other notes on the requirements of pre-emption can be made.

10.1 Reducing the Impact of Pre-emption

Where the establishment of one LSP requires pre-emption of resources at multiple LSRs, it is desireable that the same LSP be pre-empted at each intermediate LSR when possible and subject to the local policy.

Where the establishment of two or more further LSPs requires pre-emption at different points in the network, it is desireable that the same LSP be pre-empted in each case where possible and subject to the local policy.

The coordination of LSPs for pre-emption is a matter for individual implementations. The protocol messages used to indicate pre-emption must make it possible for each LSR to gather sufficient information to perform this function.

10.2 When To Apply Pre-emption

It is normal to apply pre-emption on the reverse leg of LSP setup when processing Resv. However, when processing a Path message:

- The likely need for and possibility of pre-emption may be taken into account during path computation and routing.

- Pre-emption may be applied on the forward leg of LSP setup so that, if it is known that pre-emption will be required (if the LSP sets up successfully), pre-emption can be performed ahead of time.

- In bidirectional LSPs (see [RFC3473]), and in particular when the resource is bound to the label, it may be necessary to allocate resources for the upstream data flow while processing the Path message during LSP setup. This may require that pre-emption is performed when processing the Path message.

11. Summary of Deployed Solutions

Appendix A lists some of the deployed solutions according to an informal survey of implementors and vendors. This information was garnered from a private email survey conducted by the co-chairs of the CCAMP working group.

11.1 Soft Pre-emption

There is only one deployed solution for soft pre-emption.

- A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

There is a further solution for soft pre-emption under development.

- A Resv message MAY be sent upstream with a flag in RRO subobject conveying the information about pre-emption for a specific hop.

11.2 Hard Pre-emption

There are two deployed solutions for hard pre-emption.

- A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

- A PathErr with the code "Policy Control failure" and the Path_State_Removed flag (see [RFC3473]) set is sent upstream. At the same time, a PathTear is sent downstream.

12. Security Considerations

This is an informational document that introduces no new protocols nor protocol extensions. There are no security implications of this document.

The attention of implementors is drawn to the security sections of the documents describing the pre-emption signaling features that they are implementing.

13. Acknowledgements

Thanks to Dimitri Papadimitriou, JP Vasseur, Arthi Ayyangar and Kireeti Kompella for a detailed discussion and exposee of the problems.

14. Intellectual Property Consideration

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

15. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2205] Braden, R. (Ed.), Zhang, L., Berson, S., Herzog, S. and S. Jamin, "Resource ReserVation Protocol -- Version 1 Functional Specification", RFC 2205, September 1997.

[RFC2750] Herzog, S., "RSVP Extensions for Policy Control", RFC 2750, January 2000.

[RFC2751] Herzog, S., "Signaled Preemption Priority Policy Element", RFC 2751, January 2000.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V. and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3471] Berger, L. (Editor), "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L. (Editor), "Generalized MPLS Signaling - RSVP-TE Extensions", RFC 3473 January 2003.

16. Informative References

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", RFC 2026, October 1996.

[RFC3031] Rosen, E., Viswanathan, A., and Callon, R., "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[SOFTPREEMPT] Meyer, Maddux, Vasseur, Villamizar and Birjandi, "MPLS Traffic Engineering Soft preemption", draft-ietf-mpls-soft-preemption-01.txt, October 2003, work in progress.

17. Authors' Addresses

Adrian Farrel
Old Dog Consulting
Phone: +44 (0) 1978 860944
EMail: adrian@olddog.co.uk

18. Full Copyright Statement

Apendix A Implementation Reports

A.1 Company 1

Hardware vendor.

A.1.1 Soft Pre-emption

A Resv message is sent upstream with a flag in the RRO subobject conveying the information about pre-emption for a specific hop.

Not currently deployed. Implementation under development.

A.1.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

Significantly large deployment.

A.1.3 Processing of Other PathErrs for Established LSPs

Unknown.

A.2 Company 2

Hardware vendor.

A.2.1 Soft Pre-emption

Not currently supported.

A.2.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

Significantly deployment.

A.2.3 Processing of Other PathErrs for Established LSPs

Unknown.

A.3 Company 3

Hardware vendor.

A.3.1 Soft Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream. The ingress always responds with PathTear.

Some deployment.

A.3.2 Hard Pre-emption

Not supported.

A.3.3 Processing of Other PathErrs for Established LSPs

PathErr is informational only. It does not disrupt the state of existing LSPs.

A.4 Company 4

Software vendor.

A.4.1 Soft Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream,

and a ResvErr with the code "Policy Control failure" is sent downstream.

Considerable deployment.

A.4.2 Hard Pre-emption

Not supported.

A.4.3 Processing of Other PathErrs for Established LSPs

PathErr is informational only. It does not disrupt the state of existing LSPs.

A.5 Company 5

Hardware vendor.

A.5.1 Soft Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

Some deployment.

A.5.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" and the Path_State_Removed flag (see [RFC3473]) set is sent upstream. At the same time, a PathTear is sent downstream.

Some deployment.

A.5.3 Processing of Other PathErrs for Established LSPs

PathErr is informational only. it does not disrupt the state of existing LSPs.

A.6 Company 6

Hardware vendor.

A.6.1 Soft Pre-emption

Not supported.

A.6.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream. The ingress always responds with PathTear.

Under development.

A.6.3 Processing of Other PathErrs for Established LSPs

State is retained, but ingress always responds with PathTear.

A.7 Company 7

Software vendor.

A.7.1 Soft Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

Significant deployment.

A.7.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" and the Path_State_Removed flag (see [RFC3473]) set is sent upstream. At the same time, a PathTear is sent downstream.

A.7.3 Processing of Other PathErrs for Established LSPs

Depends on Path_State_Removed flag.

A.8 Company 8

Hardware vendor.

A.8.1 Soft Pre-emption

Not supported.

A.8.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

Some deployment.

A.8.3 Processing of Other PathErrs for Established LSPs

The LSP is torn down.

A.9 Company 9

Hardware vendor.

A.9.1 Soft Pre-emption

Not supported.

A.9.2 Hard Pre-emption

A PathErr with the code "Policy Control failure" is sent upstream, and a ResvErr with the code "Policy Control failure" is sent downstream.

Some deployment.

A.9.3 Processing of Other PathErrs for Established LSPs

Unknown.

A.10 Company 10

Hardware vendor.

A.10.1 Soft Pre-emption

Not supported.

A.10.2 Hard Pre-emption

Not supported.

A.10.3 Processing of Other PathErrs for Established LSPs

The LSP is torn down.

Some deployment.