IP Switching

October 16th, 1998

Paavo Kauranne

paavo.kauranne@ntc.nokia.com

Abstract

IP Switching was originally a set of informational RFCs defined by Ipsilon Networks Inc. IP Switching is often used as general term for many related, layer 3 switching protocols, e.g. Cisco's Tag Switching and IBM's ARIS. IP Switching, or layer 3 switching is a way to achieve fast networking, that is compatible with current IP based applications. This paper handles only Ipsilon's IP Switching.

IP Switching

1. Introduction

IP Switching is Ipsilon's (Ipsilon is Nokia IPRG nowadays) approach to achieve fast IP networking by detecting IP flows and creating separate ATM virtual connection for those flows instead of routing each packet separately. An IP flow is a sequence of IP packets, which originate from a certain source and are transported to a certain destination. A strong point in Ipsilon's approach is that it supports the legacy TCP/IP application and protocols.

Similar ideas to forward IP were invented about the same time by three different groups studying a way to do IP forwarding on a hardware switch, like ATM switch. Cisco introduced similar approach called Tag Switching later. [IPS_IEEE_97]

The idea behind switching the IP traffic, instead of forwarding it traditionally, is to reduce the total work of analysing the IP packets at every node. A substantial part of IP traffic can be classified as IP flows. It makes sense to spend a little more time analysing the IP packets at one node and then use hardware switches to forward the packet on separate, dedicated virtual connections through the network.

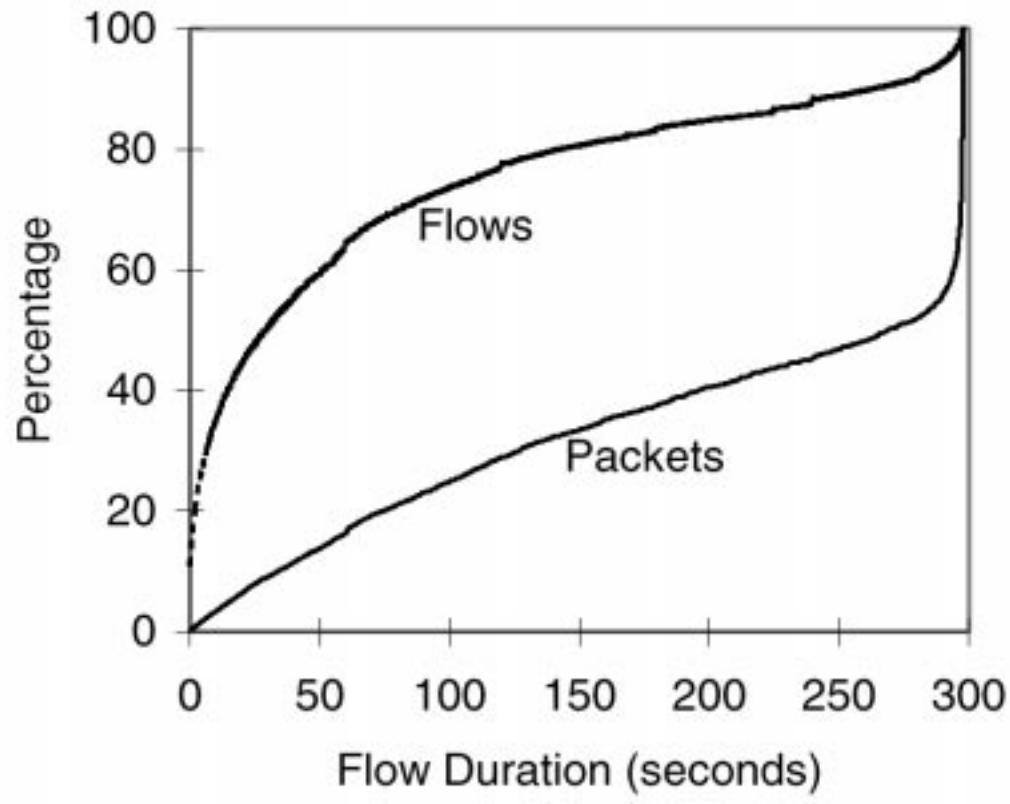

Figure 1. Cumulative Distribution of Flows and Packets vs. Flow Duration [IPS_ INTEROP96]

Ipsilon used statistics of real Internet backbone traffic. A five minute period of traffic was recorded starting at 5.15pm on Sep 25, 1995. The sample was taken at a FDDI ring, which connects the San Francisco Bay Area to the Internet backbone. The rationale was that this is the worst case traffic for IP Switching, since it can contain a wide variety of IP addresses. If IP Switching could be used with this kind of traffic, it would suite a campus or an enterprise wide network even better.

The five-minute sample was analyzed carefully. Two IP packets were considered to belong to an IP flow, if they had the same source and destination address. The figure 1 shows the result of the analysis. The result of the analysis is that 64% of flows has the duration less than 60 seconds, but they carry only 16% of the packets. This means that it is wise to forward the short duration flows, but switch the long duration flows.

In further analysis of the sample it was noticed that IP packets containing certain protocols were more suitable for IP Switching than others. The results are presented in the following table without going into the details of the analysis.

Protocol	port		%flows	%pkts	%bytes	flows/s	pkts/s	duration	pkts/flow	bytes/pkt
IP in IP		✓	0.04	2.73	2.57	0.09	456	173.1	2307	253
TCP ftp-data	20	✓	0.76	12.09	15.18	2.17	2018	118.2	525	338
TCP ftp-cntrl	21		1.55	0.74	0.23	6.50	124	38.6	16	83
TCP telnet	23	✓	1.39	4.81	1.61	4.24	803	114.3	114	90
TCP smtp	25		10.26	4.80	2.82	49.49	802	18.2	15	158
UDP dns	53		45.30	5.57	3.04	216.56	929	15.4	4	147
TCP gopher	70	✓	0.45	0.54	0.55	1.87	91	43.3	40	275
TCP http	80	✓	17.94	40.21	41.53	72.98	6717	56.5	74	278
TCP pop-v3	110		0.08	0.05	0.03	0.41	9	27.0	21	148
TCP authent	113		2.12	0.19	0.05	10.54	32	9.0	3	64
TCP nntp	119	✓	0.35	6.56	6.59	0.68	1096	176.7	627	270
UDP ntp	123		5.01	0.20	0.06	25.02	33	1.37	1.3	83
TCP netbios	139	✓	0.03	0.08	0.15	0.11	14	69.8	82	501
UDP snmp	161		1.35	0.26	0.11	6.14	43	17.9	6	115
TCP login	513	✓	0.09	0.24	0.14	0.31	41	88.1	92	156
TCP cmd	514	✓	0.01	0.13	0.07	0.06	21	49.1	316	149
TCP audio	1397	✓	0.00	2.20	2.62	0.01	367	167.9	15653	321
TCP AOL	5190	✓	0.18	0.46	0.38	0.51	77	129.8	84	223
TCP X-11		✓	0.08	0.66	0.53	0.18	111	160.6	276	217

Table 1. Flow Statistics per Protocol. [IPS_ INTEROP96]

The ✔symbol in the table indicates that the protocol appears to be suitable for switching.

2. Ipsilon's IP Switching

The IP Switch consists of two parts:

- IP Switch Controller
- ATM Switch

The IP Switch Controller is a high-end computer running a modified free UNIX variant. The IP Switch Controller connects directly to one port of the ATM Switch via an ATM interface on the IP Switch Controller. The IP Switch Controller uses General Switch Management Protocol (GSMP) to control the ATM Switch and Ipsilon Flow Management Protocol to negotiate IP flow switching with neighbouring node.

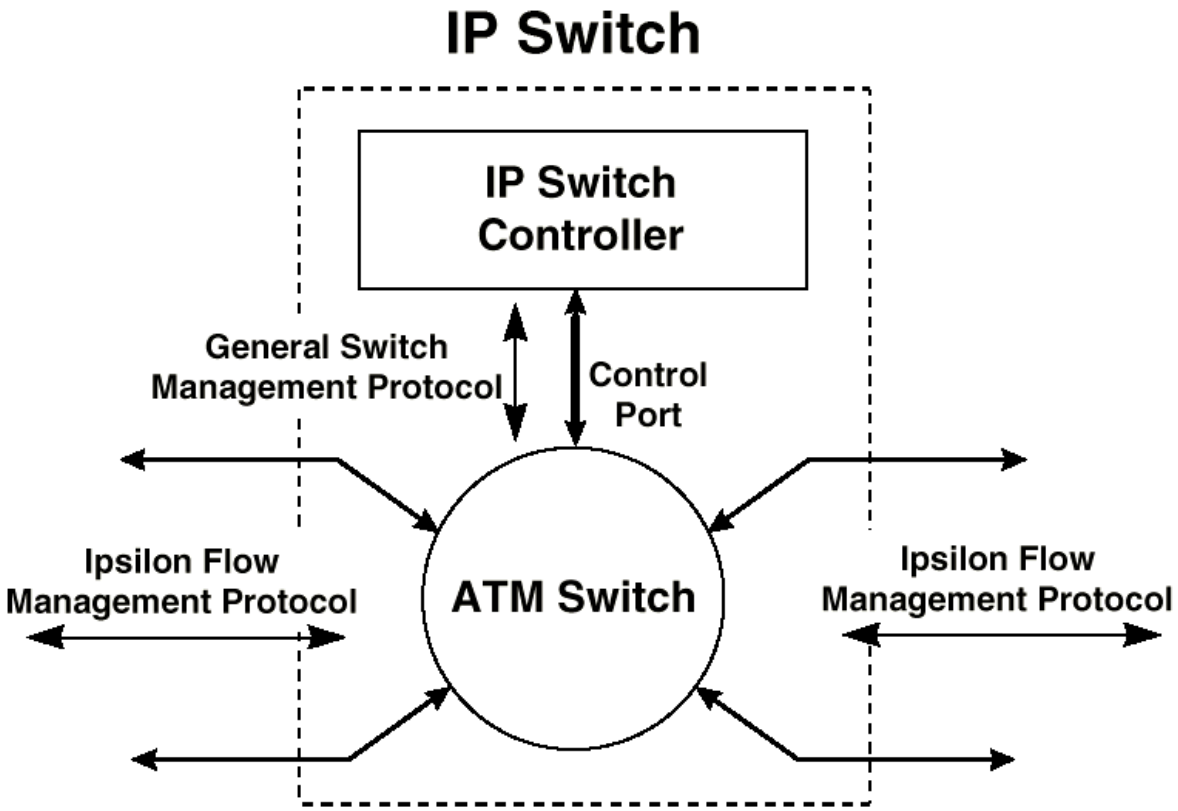

Figure 2. Structure of an IP Switch [IPS_IEEE97]

The IP Switch Controller takes care of the packet forwarding, flow detection, flow switching decision and controlling the ATM Switch. The ATM switch provides the hardware switching.

In the following figure the IP layer represents the IP Switch Controller and the data link layer represents ATM Switch. In normal packet forwarding (the dashed line) IP packets are taken to the IP layer for analysis. After the forwarding decision they are put back to the data link layer. In IP Switching, after the flow is detected and both the upstream node and the downstream node have a dedicated VC for the flow, the IP packets of the flow are switched at the data link layer (black solid line).

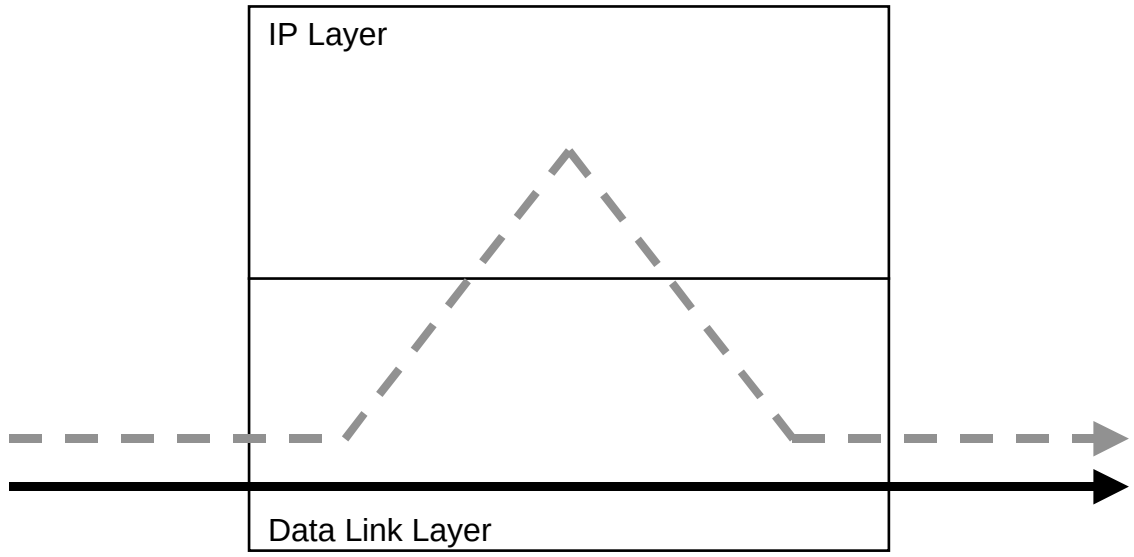

Figure 3. IP forwarding and IP Switching

The hardware switching of IP flows is fast, no IP layer functionality is needed for that. Before the IP flows are switched the IP switch has to do more instructions per IP packet than a traditional router. The IP switch analyses more fields than the traditional router:

- If a packed does not belong to a flow, the IP Switch classifies it anyway before forwarding it.
- If a packet belongs to a flow, the IP Switch has to label the packet and then forward it to the correct VC.

This means that in a network of IP Switches at least one node has to do more work per packet than in traditional router network. This is not as bad as it sounds, the labelling function is "transferred" upstream, towards the edge of the network, away from the more critical, often congested core routers of the network. Since the switched flows do not burden the IP Switch controller, there are more resources for the IP packets that are traditionally forwarded.

Ipsilon Networks Inc. has defined a set of informational RFCs, that specify two key parts of IP Switching, IFMP and GSMP, also the RFC 1954 is important, it tells how the IP flows are to be mapped on ATM data links.

- Ipsilon Flow Management Protocol v1.0, IFMP, publishes as RFC 1953 in May 1996.
- Transmission of Flow Labelled IPv4 on ATM Data Links v1.0, published as informational RFC 1954 in May 1996
- General Switch Management Protocol v1.1, GSMP, published as RFC 1987 in August 1996. An updated RFC of GSMP v2.0 was published in March 1998 as RFC 2297.

2.1 IFMP

This RFC defines a way for a node to ask the adjacent node to attach a layer 2 label to a specified IP flow. The specification of IFMP does not require the layer 2 (data link layer) technology to be ATM, but often refers to RFC 1954, which defines Transmission of Flow Labelled IPv4 on ATM Data Links. It is mentioned that the label definition and the flow encapsulation are to be specified separately for each data link layer technology as it has been done for ATM in the RFC 1954.

Since the topic of this session is IP over ATM, from now on in this paper, the layer 2 technology is supposed to be ATM.

A node classifies all IP packets it forwards, this means the node analyses all those fields of the IP packet, which are used to identify an IP flow. This classification determines if the future packet belonging to the same IP flow are to be switched via an own VC or forwarded to the default VC.

As a node recognises an IP flow, that is to be switched, it sends a Redirect request message to the upstream neighbouring node. The node receiving the Redirect message from downstream can either ignore or accept the Redirect request. No acknowledgement messages are sent in either case.

If the node accepts the request, it starts to label the packets that belong to the IP flow specified in the Redirect request and redirects them to the data link specified by the label.

Redirect Protocol

All IFMP redirect protocol messages use the same format, the Op Code field specifies, which type of message it is. Current version 1.0 of the IFMP specifies five Op Codes: Redirect, Reclaim, Reclaim Ack, Label Range and Error.

Although the messages use same format, the message body varies; it can contain one or more IFMP Redirection protocol message elements.

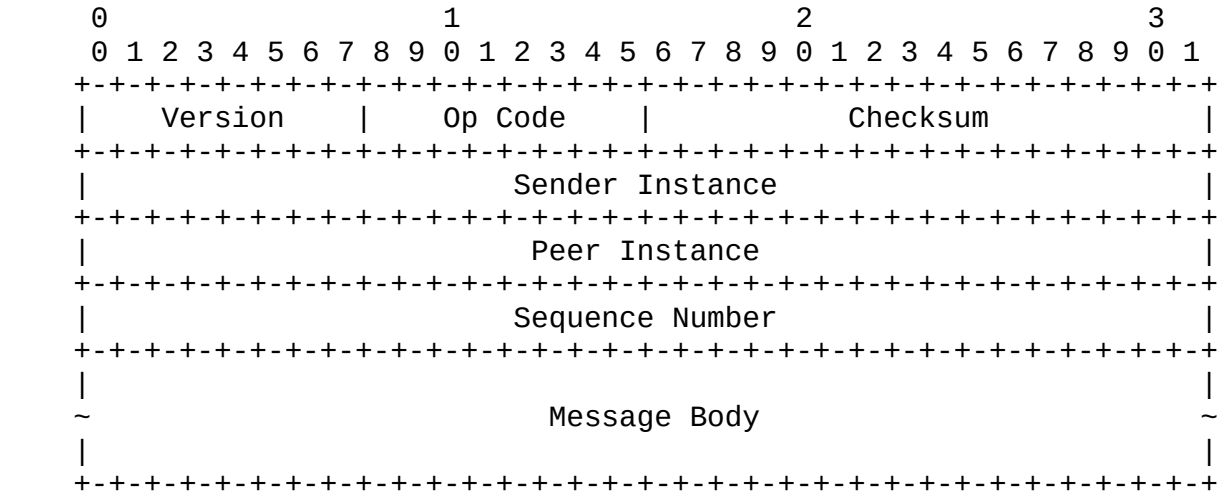

Figure 4. Redirect Protocol Message format [RFC1953]

Redirect

Redirect message is used for asking the upstream node to use certain label (VPI, VCI) for a IP flow specified as Flow Identifier in the message or to refresh timer for a IP flow that is already redirected. The message element for redirect is as follows:

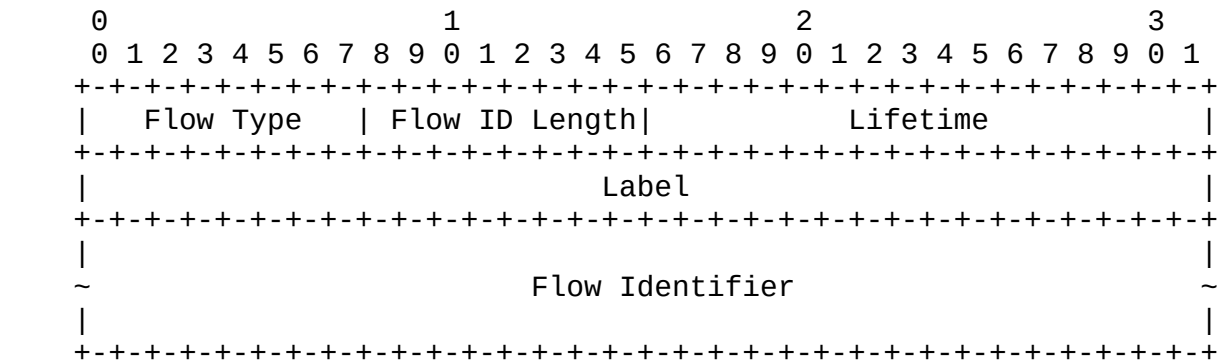

Figure 5. Message element for redirect message. [RFC1953]

Lifetime field specifies the time the redirect request is valid, if a new, identical redirect request has not been received by the upstream node, the flow is forwarded to the default VC and the VC specified for the flow is released. The label field is used to identify the Data Link Layer addressing, in ATM case, the label is:

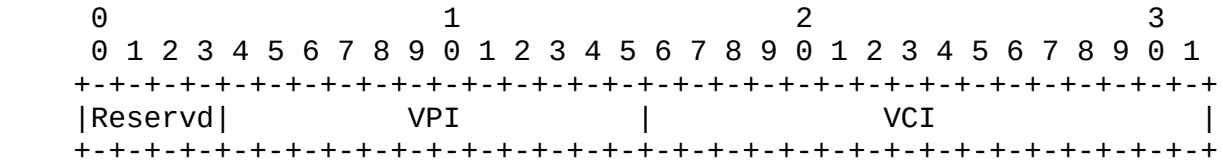

Figure 6. Label for ATM at Data Link Layer [RFC1954]

There are two types of flow identifiers used, type 1 is used to identify a flow that originates from a certain port of a certain host, carries certain protocol and destines at certain port of a certain host. The identification of the flow contains all these fields and a couple of more:

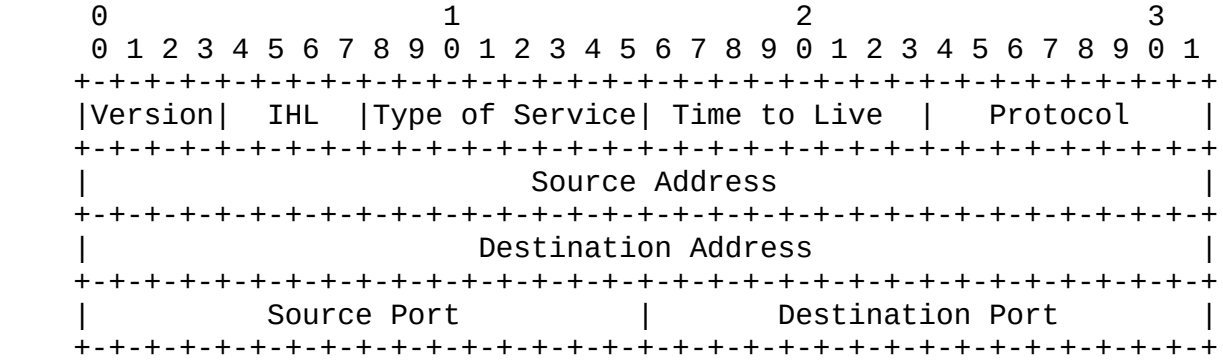

Figure 7. Flow type 1 identifier. [RFC1953]

The flow identifier type 2 does not know about the protocol or the source and destination port fields of the packets belonging to aflow, it is simply used to identify a flow from one host to another.

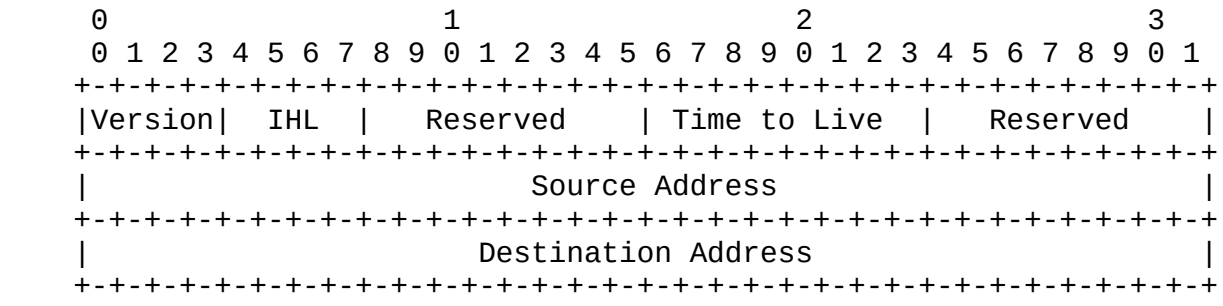

Figure 8. Flow type 2 identifier. [RFC1953]

Reclaim and Reclaim Ack

Reclaim is used for asking the upstream node to release the VCs for specified flows and to forward the packets via the default VC. The upstream node uses Reclaim Ack to acknowledge the release of the reclaimed VCs. The message element for Reclaim and Reclaim Ack messages are similar to the one in Figure 4, except the Lifetime field is reserved for future use in these message elements.

Label Range

Label range message can be sent as a response message to the redirect request if any of the redirect message elements specified a label (VCI) that the upstream node could not handle. Label range message is used to specify the usable range of labels for this link. The label fields format for ATM was presented in figure 5. Only one label range message element can be included in one Label range message.

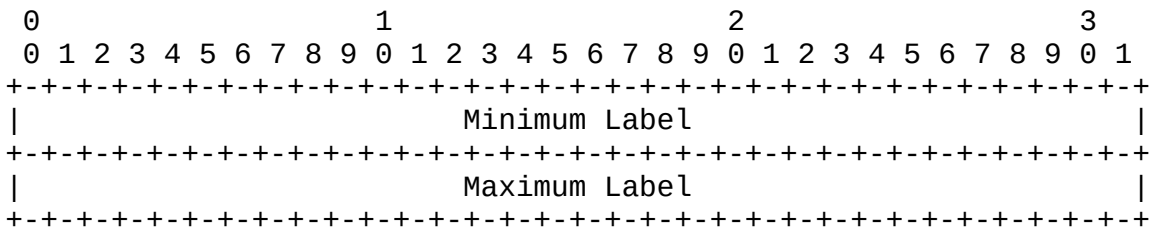

Figure 9. Label range message element. [RFC1953]

Error

A node can send an Error message in response to any IFMP Redirection Protocol message. Only one message element can be included in the Error message, the format of the Error message element is following:

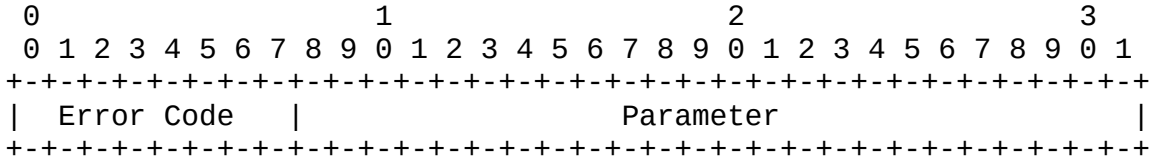

Figure 10. Error message element. [RFC1953]

Adjacency Protocol

IFMP also specifies an adjacency protocol for nodes to be able discover the identity of the neighbouring node and to synchronise the state across the data link. In order for the node to send any Redirection Protocol messages the link state has to be synchronised. [RFC1953]

2.2 GSMP

General Switch management Protocol is a simple protocol to control an ATM Switch. The protocol is a master slave protocol, where the GSMP Controller (IP Switch controller) is the master and the ATM Switch is the slave. There is, however, a mechanism for the switch to report asynchronous events to the GSMP Controller. Events, such as the physical line status of a link has changed; new port has been added to the ATM switch; a port has been taken off the ATM switch and the ATM Switch has received a cell with VPI/VCI combination, which is not currently allocated to any connection.

A switch is assumed to contain multiple "ports". Each port is a combination of one "input port" and one "output port". Some GSMP requests refer to the port as a whole whereas other requests are specific to the input port or the output port. ATM cells arrive at the switch from an external communication link on incoming virtual channels at an input port. ATM cells depart from the switch to an external communication link on outgoing virtual channels from an output port. [RFC1987]

There is a kind of flow control for the Event messages that the ATM Switch sends. As a certain type of Event is sent the ATM Switch is not allowed to send further Events of the same type before the GSMP Controller has given the Switch the right to send that kind of Event again. This is to hinder the ATM Switch flooding the GSMP controller with similar Event messages. As the ATM Switch is not allowed to send every Event it notices, it increments an event counter for every Event. The event counter value is passed to the GSMP Controller in Port Management acknowledge messages and in Event messages. As the GSMP Controller notices that the count of Events has changed it becomes aware that it has missed some Events due to the flow control and can act accordingly.

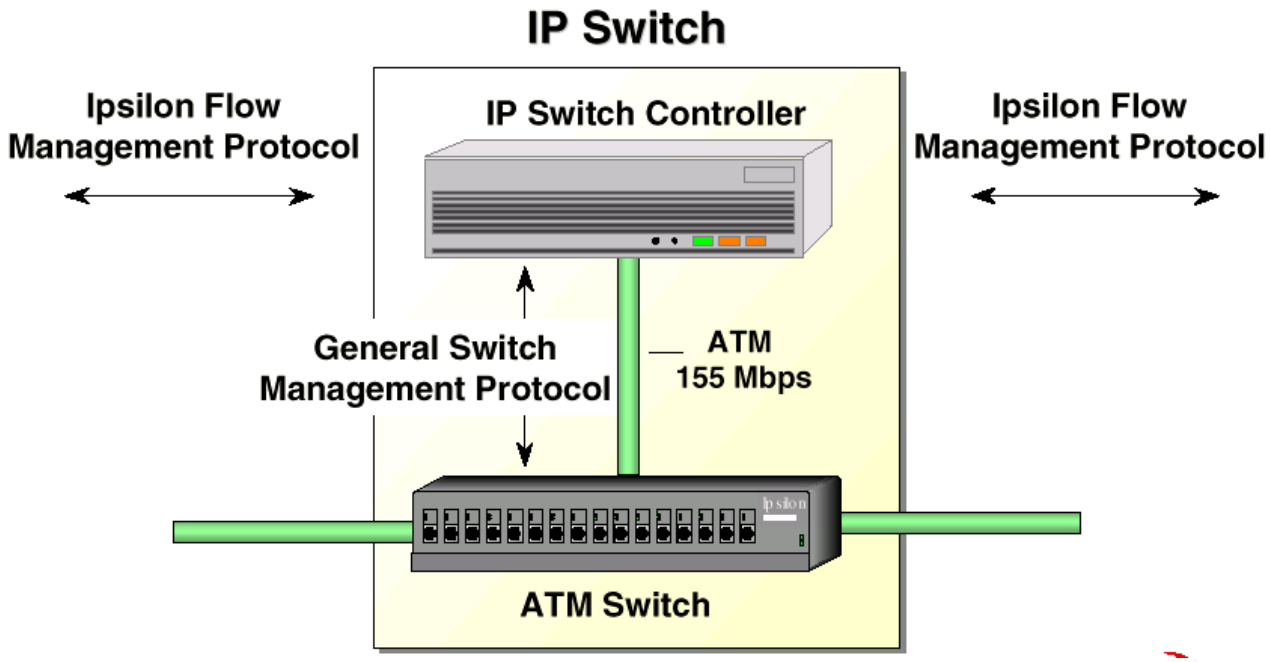

Figure 11. Environment for GSMP [IPS_OPENSIG96].

As with IFMP, the GSMP also defines an adjacency protocol to synchronise the control channel between the GSMP Controller and the ATM Switch and to discover the identity of the entity in the other end of the link. No other GSMP messages than adjacency messages may be transmitted before the adjacency protocol has achieved synchronised state across the control channel.

The GSMP Switch can create and release point to point connections; add and release leaves on point to multipoint connections; query for ATM Switch configuration and statistics and manage ATM Switch ports. GSMP requests are of four types: Connection management requests; Port Management requests; Switch Configuration queries and Switch Statistics requests. All of the GSMP messages, except the adjacency protocol, have same basic layout of the message, which is presented in the figure 11.

The GSMP Controller has control over an ATM Switch Port administrative status; it can either Bring Up or Take Down the ATM Switch Port. If the operational status of the Port (line status) is disabled then the Port cannot transmit or receive any cells, no matter what the administrative Port Status is. Every Port is

The GSMP specification does not tell how the Ports are configured or what is the interface to configure the switch. As the ATM Switch notices that there is a new Port configured, it can send an event message to notify the GSMP Switch controller about a "new port in the hood". Other possibility would be to periodically poll the ATM Switch configuration with the appropriate GSMP request.

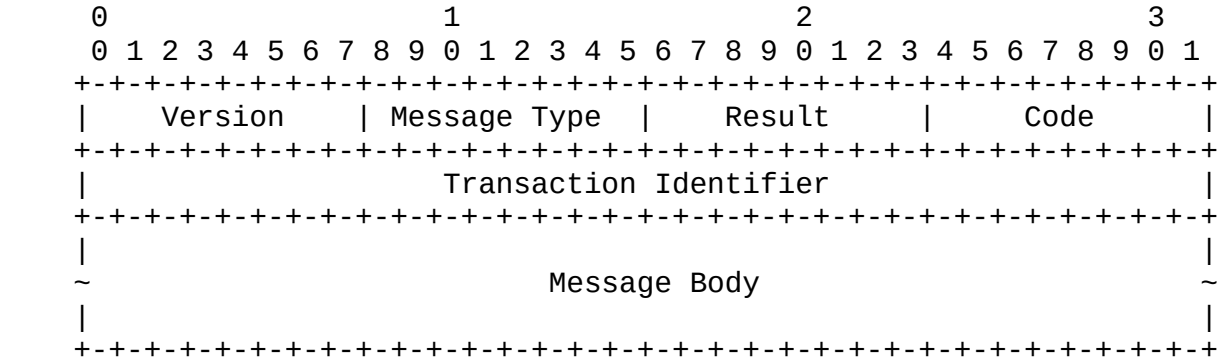

Figure 12. GSMP Message format. [RFC1987]

The Version field specifies the version of GSMP, in the RFC1987 this is specified to be 1. Message Type field tells which GSMP message the message is. The Result field is used a bit differently in request and acknowledge messages:

In request messages the Result field tells if the GSMP Controller wants to get an acknowledge for successful case also or only for unsuccessful case.

In acknowledge message the Result field is used to indicate whether the request was carried out successfully or not. If the ATM Switch does not manage to carry out the GSMP Controller's request, it can use the Code field together with the Result field to give further information about the nature of the failure.

All the messages, except Configuration messages, contain Port Session Number field, which is used to specify the "life" of the Port. When a Port is created first time a random number is set to Port Session Number for the new Port.

Every time a Port is brought up with Port Management message a new random number is set to be the Port Session Number. In order for the GSMP Controller to be able to control the Ports of the ATM Switch, it must know the current Port Session Number of the Port it wants to manipulate. This is to make sure the GSMP Controller does not act on the basis of obsolete knowledge about the Ports.

The message format differs according to the request. The Connection management messages use the same format, except the Move Branch message, which adds some extra fields to the format.

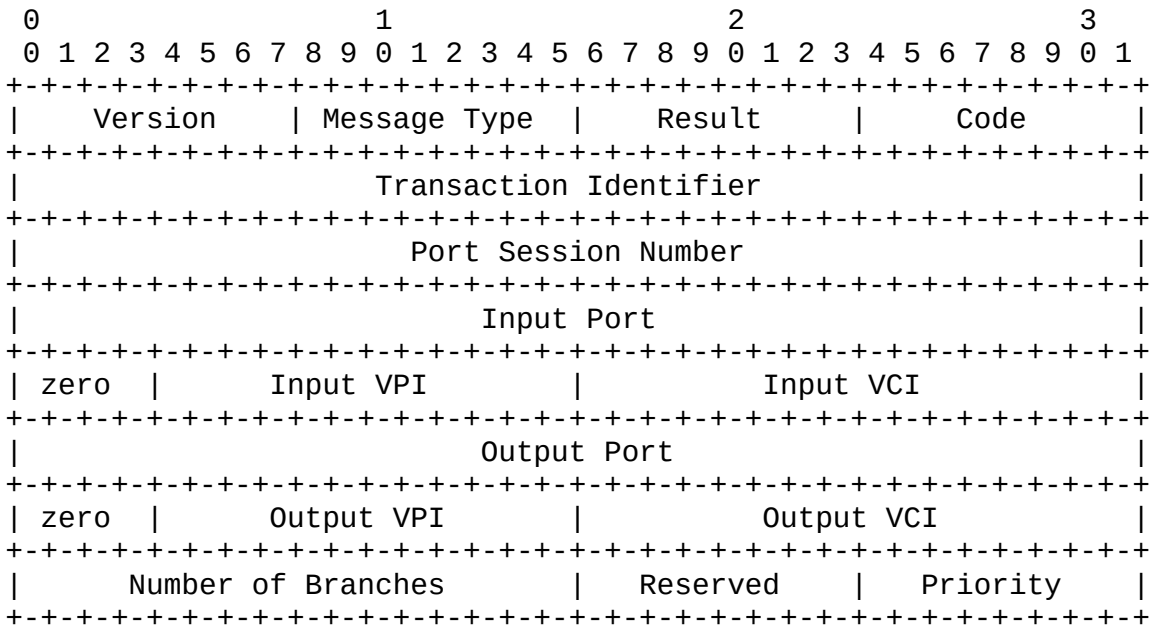

Figure 13. Connection Management message body. [RFC1987]

GSMP specifies the connection with a pair of Port, VPI and VCI fields. Since the version 1.1 of GSMP does not have take Quality of Service into account, it has a Priority field to handle this. It is said that if two connections with different priorities share the same output port, it is then likely that the cell belonging to the higher priority connection leave the port before the cell with lower priority if they arrive to the switch at the same time.

Connection Management Messages

The connection management messages are used to control the connection an the ATM Switch, the GSMP connections are unidirectional SVCs The ATM Switch uses some kind of default settings for the traffic parameters, since they are not given in the GSMP connection management messages.

Add Branch

Add Branch message is used to either create a new point to point (ptp) connection, if the Input Port, VPI, VCI combination is not in use already, or add a leave to existing connection, if the Input Port, VPI, VCI combination is in use already. If the GSMP controller specifies No_Success_Ack in the Result field then the ATM Switch would not acknowledge if the request were carried out successfully.

Delete Branch

Delete Branch is used to remove a ptp connection or to remove a leave from a point to multipoint (ptm) connection. The leave to be deleted is specified by Input and Output Port, VPI and VCI combinations.

Delete Tree

Delete Tree message is used to remove a whole connection. In this message the Input Port, VPI and VCI specify the connection.

Verify Tree

Verify tree message is used to verify that both the GSMP Controller and The ATM Switch share the same opinion about the number of leaves that belong to the connection specified by the Input Port, VPI and VCI. This, along with the corresponding acknowledge message, is the only GSMP Connection Management message, which uses the Number of Branches field.

Delete All

Delete All message is used to remove all the connections from a certain port. The Input Port field specifies the port.

Move Branch

Move branch is the only GSMP Connection management message with different message layout that the others. Instead of the Output Port, Output VPI, Output VCI fields there is the following structure:

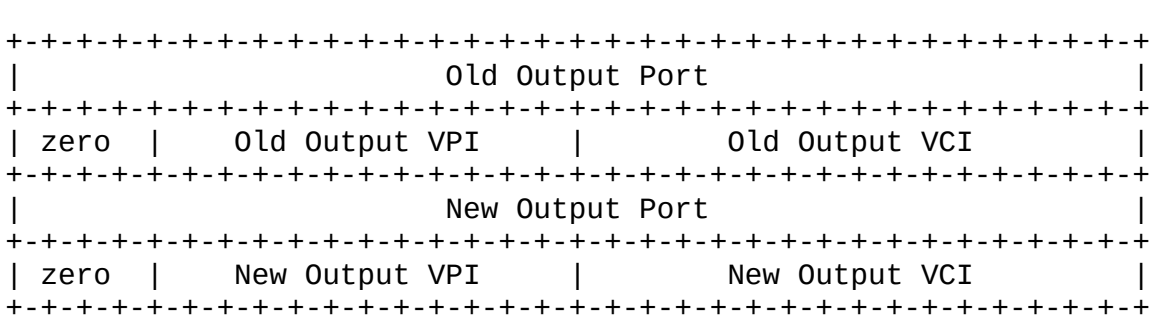

Figure 14. Different fields of Move Branch message. [RFC1987]

This means that only one leave of a connection can be moved at one request and the root of the ptm cannot be moved.

Port Management messages

Port Management messages have a common format as well, the Function field differentiates the Port Management messages.

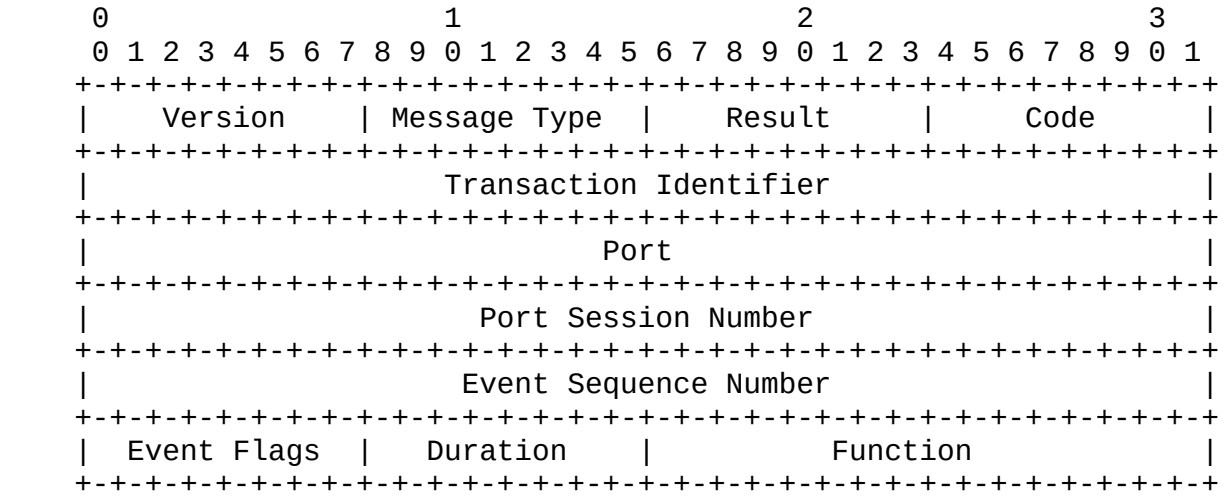

Figure 15. Port Management message format. [RFC1987]

Bring Up

Bring Up message is used to set a certain Port Status to Available state. The Port is identified with Port field. All the connections that arrive at the port must be deleted, a new Port Session Number generated and all the dynamically assigned VPI/VCI values for the Port are to be released. The new Port Session Number will be carried in the acknowledge message.

Take Down

Take Down message sets the Port Status to Unavailable state. No cells are to be transmitted or received at the Port.

Loopbacks

Internal, External and Bothway Loopbacks messages are used to test the hardware. Internal loopback directs the cells leaving the Output Port back to the Input Port, External loopback loops cells arrive at the Input Port back to the link without entering the port. In Bothway loopback both Internal and External loopbacs are performed. The time to spend in loopback is specified by the Duration field. After the loopback the Port is returned to service.

Reset Input Port

Reset Input Port message removes all the connections that arrive at the Input Port. The Port Status is set to Unavailable.

Reset Event Flags

Reset Event Flags message is used to clear the Flags that specify which Event messages are not allowed to be sent for the particular Port. Any combination of the Event flags may be cleared with one message.

Statistics Messages

Statistics message is used to query for ATM Switch statistics. The statistics are hardware counters e.g. cell and cell discard counts. There are two classes of Statistics messages defined: VC Activity messages and Port and VC Statistics messages. The VC Activity message is user to see if a one or more VCs have been carrying traffic recently. The Port and VP Statistics message is used to get the statistics counters of a certain Port or VPI/VCI.

Configuration Messages

Configuration messages are used to query for Switch or Port configuration information. The Switch configuration message asks for Switch type, Switch Name and Switch Firmware version number. The Port configuration messages query for Port information, either for a certain Port or for all the Ports. The information available is in the following:

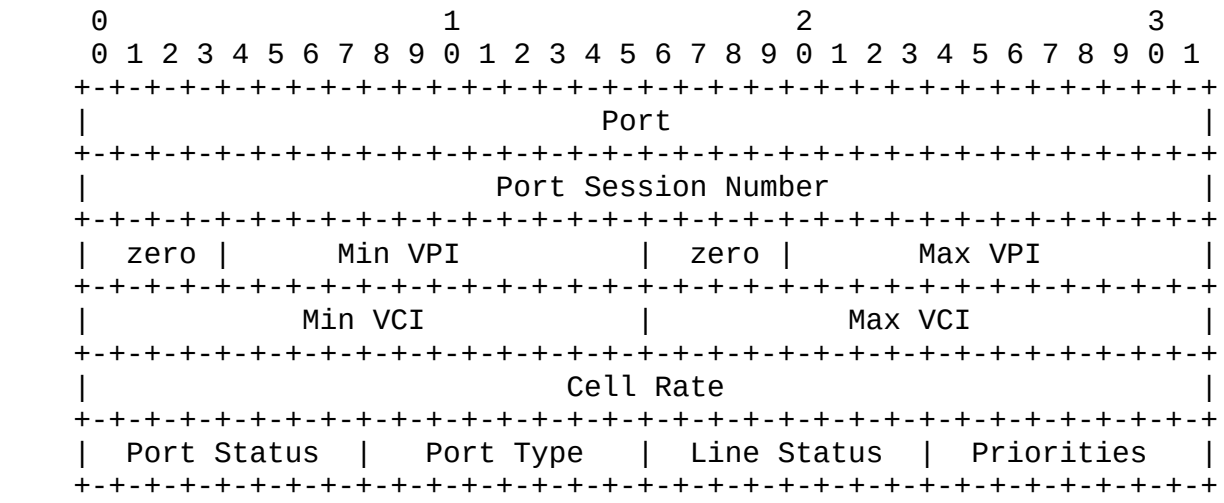

Figure 16. A record for Port Information. [RFC1987]

The information returned with Port Configuration and All Ports Configuration messages is presented above. These messages are the way of GSMP Controller to get to know the current status of the ATM Switch.

Event Messages

Event messages are used to inform the GSMP Controller about asynchronous events in the ATM Switch. As said before there is a kind of flow control for the Event to hinder the ATM Switch flooding the GSMP Controller with Event messages. The GSMP Controller does not acknowledge the Event messages. Event messages are of the following format:

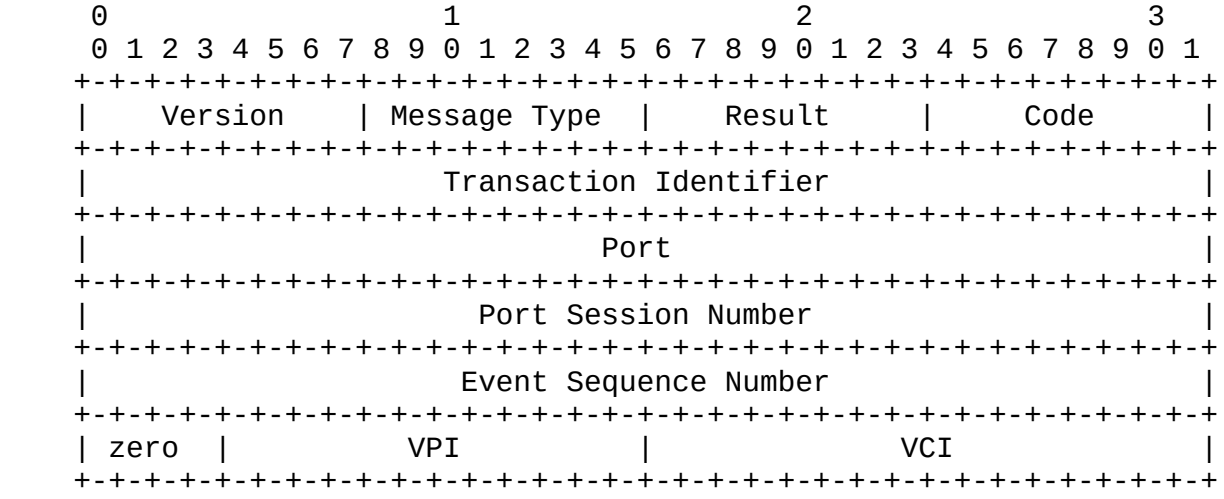

Figure 17. Event message format. [RFC1987]

Message Type field specifies, of which type of the Event message is.

Port Up

Port Up message is used to inform the GSMP Controller about the state of physical line going back to normal operational state. A new Port Session Number should be generated for the Port.

Port Down

Port Down message is used to sent the information about physical link state going to unavailable, i.e. to report a link failure.

Invalid VPI/VCI

Invalid VPI/VCI message is used to inform the GSMP Controller about Port receiving a cell that has got VPI//VCI combination that is not currently used by any of the connections defined for the Port. This is the only Event message, where the VPI and the VCI fields are used.

New Port

New Port message inform the GSMP Controller about a new port added to the ATM Switch. The port number and the Port Session number are delivered to the GSMP Controller. The state of the New Port is not defined in the specification.

Dead Port

Dead Port message is used to inform the GSMP controller about a certain Port removed from the ATM Switch.

3. References

[RFC1953] Newman, P., et. al. "The Ipsilon Flow Management Protocol Specification for IPv4 Version 1.0", Informational RFC, May 1996

[RFC1954] Newman, P., et. al. "Transmission of Flow Labelled IPv4 on ATM Data Links Ipsilon Version 1.0", Informational RFC, May 1996

[IPS_IEEE_97] Newman, P., et. al., "IP Switching and Gigabit Routers."
< http://www.iprg.nokia.com/~pn/papers/ieee_comm96.pdf>
IEEE Communications Magazine, Jan 1997

[IPS_ INTEROP96] Newman, P, et. al. "Flow labeled IP: Connectionless ATM Under IP."
< http://www.iprg.nokia.com/~pn/papers/interop96.pdf >
Engineer's Conference, Interop, April 1996

[RFC1987] Newman, P., et. al. "General Switch Management Protocol Specification Version 1.1". Informational RFC, August 1996

[IPS_OPENSIG96] Newman, P., "GSMP: Quality of Service Strikes Back"
<http://www.iprg.nokia.com/~pn/presentations/opensig97/opensig97.pdf>.
Opensig Fall Workshop, Columbia University, Oct. 1997