Smart IP Switching: A Hybrid System for Fast IP-based Network Backbones

David Lloyd, Telenor RDI, Clonshaugh Industrial Estate, Dublin 17, Ireland
david.lloyd@dublin.fou.telenor.no
&
Donal O'Mahony, Networks & Telecommunications Research Group,
Computer Science Dept, Trinity College, Dublin 2, Ireland
Donal.OMahony@cs.tcd.ie

Abstract

The process of routing, that is, the handling of Internet Protocol (IP) packets at layer 3 (the network layer), is a potential bottleneck in large intranets and the public Internet. Asynchronous Transfer Mode (ATM) provides a hardware capability of switching cells at very high speeds at layer 2 (the link layer). A number of schemes for supporting IP over ATM hardware have emerged in recent years. These techniques may be broadly categorised by the approach they employ to set up paths through the network, namely, traffic-driven and control-driven. We propose a hybrid scheme, which while being essentially traffic-driven, exhibits advantages of both the traffic-driven and control-driven approaches.

1. Introduction

Routers are essential to form the backbone of the Internet, and are also used in the local area to control the scope of broadcast traffic. The process of routing, that is, the handling of Internet Packets (IP) packets at layer 3 (the network layer), requires substantial processing on the part of the router. Although many new innovations are being introduced to allow routing at very high speeds (layer 3 switches), this will always be slower than switching (layer 2). This is particularly the case when routing is compared against the switching of fixed size cells, as is the case with Asynchronous Transfer Mode (ATM).

2. IP over ATM

Many schemes for supporting IP over standard ATM Forum compliant ATM, have emerged over the years. These schemes which are generally referred to as *IP over ATM*, include Classical IP over ATM [1], LAN Emulation (LANE) [2], and Multi-Protocol over ATM (MPOA) [3]. The schemes emulate an IP network in various ways over a network of links set up using standard ATM signalling. These schemes are complex, since there are two networks to manage and control, the ATM network and the emulated IP network.

3. IP on ATM hardware

The complexity of *IP over ATM* schemes inspired a number of schemes, which make use of ATM hardware to speed up IP, while eliminating the complexities of full ATM signalling. These *IP on ATM hardware* schemes, are the subject matter of this paper. The IETF Multiprotocol Label Switching (MPLS) working group, have produced a framework [4], which classifies the various schemes and attempts to lay the way forward for standardisation.

3.1 Approaches to IP on ATM hardware

In an IP network, packets contain layer 3 addresses and consequently cannot be switched at layer 2, without the prior establishment of state information at each node along the path from source to destination. Usually, this takes the form of the attachment of a label to the packet and the establishment of a layer 2 path through each node along the route for frames bearing this label.

Techniques for solving this problem differ primarily in the setting up of state information in intermediate nodes. Two principal approaches have emerged: control-driven and traffic-driven.

3.2 Control-driven Approaches

In the control-driven approach, an attempt is made to establish all the required paths through the network in advance of data traffic transmission. This requires identification of ingress and egress nodes (network entry and exit nodes respectively) to and from the backbone network and associating labels with the paths from each ingress node to all possible egress nodes.

Once paths have been established, packets entering the network are labelled at the ingress node, permitting intermediate nodes to switch them at layer 2 all the way to their egress node. The control-driven approaches allow flow aggregation, that is, the grouping of one or more individual flows, to take place at any node along the path to the egress node. This is normally achieved by using one of Virtual Channel (VC) merge capable ATM hardware, or Virtual Path (VP) merge. Clearly, there is substantial overhead in setting up these paths to begin with, and the failure of any node in the network will mean that substantial re-configuration must take place. Furthermore, VC merge capable ATM hardware is not readily available, and VP merge techniques are non-trivial. Examples of this approach are Cisco's Tag Switching [5], and IBM's Aggregate Route-based IP Switching (ARIS) [6].

3.3 Traffic-driven Approaches

In the traffic-driven approach, the network starts out as a conventional routed network. The routers examine the traffic as it travels past and attempt to identify 'flows'. Flows are groups of IP packets that are associated in some way, usually a sequence of packets sent from a particular source system and heading to the same destination system.

When a node detects an IP packet flow that has potential of being long-lived, it allocates a locally significant layer 2 label and requests that its upstream neighbour applies it to all further packets of the specified flow. If the upstream node accepts the request, then a dedicated VC based on the label is established and the ATM switch fabric (switching hardware) is updated. If a node has an incoming and outgoing dedicated VC for a specified flow, it initiates a cut-through path. This allows subsequent packets of the flow to be switched rather than routed.

The primary example of the traffic-driven approach is Ipsilon's IP Switching [7], while Toshiba's Cell Switch Router (CSR) [8] works on a similar principal. Since flow detection and switching takes place in a localised manner at each individual backbone node, the system is highly resilient in the face of network failure, as the absence of traffic causes the network to revert to conventional routing with the associated re-configuration capability.

The traffic-driven approach suffers from the fact that initial packets in a flow must be routed before a switched path can be set up. Short-lived flows may not exist for a sufficiently long period to have a switched path set up. There is an overhead in terms of signalling traffic that is incurred in order to set up a switched path. Furthermore, flow aggregation is not an inherent property of the traffic-driven approaches. Control-driven approaches, on the other hand inherently perform flow aggregation and allow all packets that can be switched to be switched without any per-flow signalling traffic being generated.

4. Smart IP Switching

We propose a new scheme, which we call *Smart IP Switching*. This is a hybrid system that attempts to combine the best features of both the control-driven and traffic-driven approaches. It is essentially traffic-driven, however, it allows aggregation of flows destined for different end-systems which exit the backbone network through the same egress node. This leads to the establishment of a switched path from ingress to egress node similar to a control-driven mode of operation. Flow aggregation in Smart IP Switching is based on longest match Classless Inter-Domain Routing (CIDR) prefixes [9]. It increases the probability of packets being switched and also lowers the per-flow signalling traffic overhead. In addition, the concept of long-term dedicated VCs is introduced which allow switched paths to be kept open in anticipation of later use.

4.1 Conventional IP Switching

In a conventional IP Switching Network, IP packets are carried in ATM cells from one node to another. ATM uses locally significant channel identifiers that are broken into a Virtual Path Identifier (VPI) and a Virtual Channel Identifier (VCI) within the path. By default, IP packets are fragmented as required by ATM Adaptation Layer (AAL) -5 and sent in the default virtual circuit (VPI=0, VCI=15). The IP datagrams are re-assembled at each node, and a conventional routing operation is performed on them which results in their being once again encapsulated in an AAL-5 Protocol Data Unit (PDU), labelled with the default VPI/VCI and dispatched on the output port.

IP switches continuously monitor the flow of IP packets, and when a sequence of packets that qualifies as a potentially long-lived flow is detected, an Ipsilon Flow Management Protocol (IFMP) [10] *REDIRECT* message is sent to the upstream node in the network. The REDIRECT message is a request to the upstream node that it label all further packets belonging to the specified flow with a particular flow label (VPI/VCI pair). This will cause subsequent packets to be sent by the upstream node within the dedicated VPI/VCI established for this purpose.

If a node notices that a dedicated channel has been set up for a flow at both input and output ports of an IP switch, it responds by writing an entry into the switch table allowing what is known as 'cut-through' switching to take place. In a similar fashion, each node along a path makes this decision independently, resulting in a switched path being set up all the way from ingress to egress node. The cut-through paths are set up with a defined lifetime. The lifetime is refreshed by the sending of further REDIRECT messages upstream by each downstream node. Each downstream node along the path will continue to refresh the lifetime of the cut-through path as long as packets continue to use it. If the flow ceases, after a short period, the cut-through paths will expire returning the network to normal routing based service.

4.2 Smart IP Switching Architecture

Smart IP Switches take a similar approach except that as well as managing potentially long-lived flows, the scheme will manage potential aggregate flows. In the following sections we introduce and discuss the concepts involved in the architecture.

4.2.1 Unique Flow Identifier (UFI)

Detection of a potential aggregate flow is based on matching the destination address against the routing table in that node to generate an identifier, which we call the *Unique Flow Identifier (UFI).* A particular flow or aggregate flow may be represented at any point in a network by a locally generated UFI. The format of the UFI is illustrated in Figure 1.

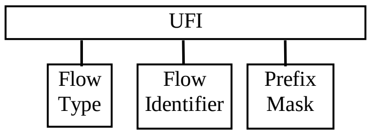

Figure 1 : Unique Flow Identifier (UFI)

REDIRECT messages are sent, and cut-through paths established based on the UFI.

4.2.2 An Aggregate Flow Type

The UFI can represent the standard Ipsilon flow types 0, 1 and 2, as defined in [10], in addition to a new flow type we call *flow type 3*. Flow type 3 is specifically defined to represent aggregate flows. When the UFI represents a flow type 3, any packets that arrive at the ingress node and are destined for hosts with the same CIDR prefix can avail of the switched path. This is particularly useful in handling situations where a heavily accessed Internet site (e.g. Netscape, Microsoft or Altavista) is implemented using a pool of servers within the same IP sub-network.

4.2.3 Merging of Flows

Flow aggregation is the merging of two or more individual flows. The methods for this described in the literature are VC merge and VP merge. VC merge is the merging of flows which may come from a number of separate incoming VCs (VPI/VCI pairs) into the same outgoing VC (VPI/VCI pair). Under normal circumstances, VC merge cannot be achieved without the use of VC merge capable hardware. IP datagrams are carried in AAL-5 Protocol Data Units (PDUs) in all of the IP on ATM hardware schemes. Since AAL-5 relies on cells being delivered in-order, care must be taken to ensure that cells from different PDUs do not become interleaved, as it would be impossible to recover the individual PDUs from the merged cell stream.

VP merge is the merging of one or more incoming Virtual Paths (VPs) onto the same outgoing VP. When it is employed in the control-driven approach, it is necessary to ensure that none of the incoming VPs, use the same VCI to carry different flows that are to be merged. This is because, when VPs are merged, there is a *one-to-one* mapping of one or more incoming VCIs, to the same outgoing VCI. Therefore, in order to avoid the possibility of cell interleave in the VP merge scheme, it is necessary to allocate VCIs in blocks to all potential sources. Smart IP Switching must also ensure that cells from different AAL-5 PDUs are not interleaved, and the techniques it employs are discussed in the following sections.

4.2.4 Ingress-Piping

The only point in a backbone where it is possible to merge flows onto the same VC is at an ingress node. This was taken into consideration in the Smart IP Switching implementation. Furthermore, potentially packets can be injected into the backbone at any node. To allow for this, each node must maintain a dedicated VPI/VCI for each UFI that it is currently managing, which we will call the *ingress pipe*. This is allocated initially when the first REDIRECT message is received for the flow, and is used for subsequent packets that enter the backbone network at this point.

4.2.5 Dedicated Virtual Channel (VC) Set-up

When a potential aggregate flow is detected, an IFMP REDIRECT message is sent to the upstream node, requesting that the node attach a particular label to all further packets belonging to the specified flow. Smart IP Switching specifies a flow type 3 in a manner similar to the way a conventional IP Switch will specify a flow type 0, 1 or 2, except a CIDR prefix is encoded into the REDIRECT message element[1]. Upstream nodes will set up outgoing dedicated VC in the same manner as a conventional IP Switch.

4.2.6 Cut-through Operation : Down-piping

When a cut-through path is to be established for a flow type 3 whose ingress point is at some upstream node, the flow is shifted from the ingress pipe to a separate dedicated VC allocated from a VC pool. We call this process *down-piping* and it effectively ensures that AAL-5 flows can never be inadvertently interleaved. The down-piping operation is illustrated in Figure 2, in which Smart IP Switch (SIPS) 2, creates a cut-through path to the *next available* outgoing dedicated VC.

4.2.7 Virtual Merge

At intermediate nodes in a backbone where flows that enter a switch from different interfaces must be dispatched on the same interface, cell interleave of AAL-5 PDUs must be avoided. When one or more new tributary flows are joined with an existing flow that is in cut-through mode, they are switched to the *next available* outgoing VC on the outgoing interface. This action is shown in Figure 3, in which the flows coming from SIPS4, on the incoming dedicated VC are switched to the next available free VC on port 2 of SIPS2. There is no time delay in setting up a new VC, since the adjacent switches maintain a pool of pre-established VCs between them. The association between all the flows that leave SIPS2 on port 2, and share the same UFI is what we refer to as a *virtual merge*, and is shown in figure 3. Unlike VP merge, virtual merge has no intrinsic value in itself, it is merely a descriptive term.

[1] In our implementation, the CIDR prefix was encoded into a flow type 3 identifier.

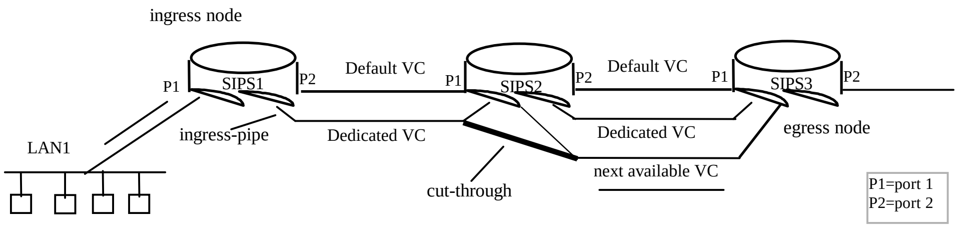

Figure 2: Cut-through (Down-piping)

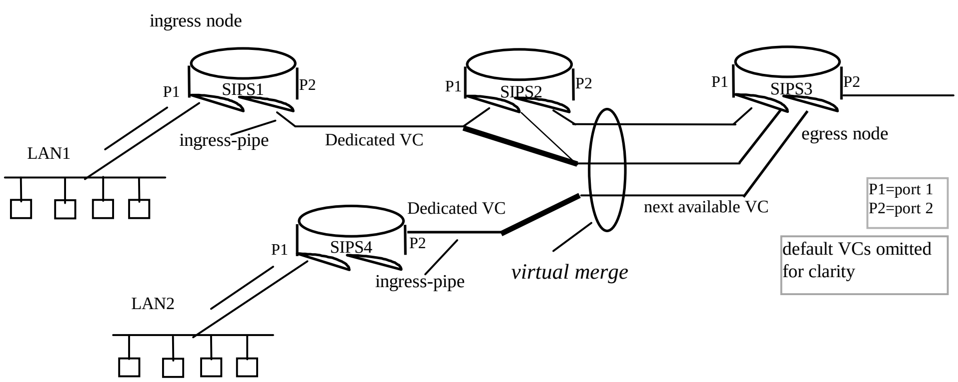

Figure 3: Virtual Merge

At a downstream node all flows which constitute the virtual merge may be associated together by the fact that they are represented by the same UFI. As more flows join the mainstream cut-through flows, they are treated in the same manner as above, and switched to the next available VC from the pool. Each of these individual VCs may contain one or more flows, since each one was an ingress-pipe at some upstream node.

4.2.8 Flow Management

In the Smart IP Switching scheme, a Flow Information Base (FIB), is used in the management of flows. This is similar to the Ipsilon approach except that in a SIPS, the FIB is keyed on the UFI. FIB entries that represent downstream nodes and are of flow type 3, have associated with them a table which is used to keep track of upstream sources. This table, known as the *source VC table*, will contain an entry for every upstream ingress node. The purpose of keeping track of all upstream sources is firstly, to ensure that all dedicated VCs along the path from egress to the current node that remain active, continue to receive REDIRECT messages to refresh their lifetime. Secondly, when a flow type 3 expires, all its cut-through paths must be deallocated. Each ingress node is represented in a source VC table by a representative IP address. The representative IP address is the source IP address of the first flow that initiated the setting up of the dedicated VC for a particular ingress pipe.

Although REDIRECT messages must be sent all the way back from the egress node to the ingress node, the Smart IP Switching scheme achieves this on a localised basis. In conventional IP Switching, REDIRECT messages are sent at regular intervals to the relevant upstream node as long as the associated dedicated VC is active. In the case of Smart IP Switching, when refreshing an aggregate flow each of these REDIRECT messages will have a number of message elements attached. Each message element will be representative of one upstream ingress pipe that remains active. A separate REDIRECT message will have to be sent to each upstream node that is attached on a different physical interface. When an upstream node receives a REDIRECT message for a flow type 3, as well as refreshing the dedicated VC between it and the downstream node that originated the message, it will refresh the lifetime of each source VC table entry for which there is a message element. In a similar fashion, when this node constructs a REDIRECT message to send upstream, it will attach an individual message element to represent each entry in its source VC table.

4.2.9 Longevity of Virtual Channels (VCs)

An additional property of Smart IP switched paths which differs from conventional IP Switching is that each dedicated VC is classified as being either long-term or short-term. When initially established, a dedicated VC is considered to be short-term. If aggregation occurs within its lifetime, then its classification is changed to long-term and a longer expiry timer is set. Furthermore, Smart IP Switching nodes also keep track of paths that have recently expired. If these are re-activated within a short period, then they are set up as long-term dedicated VCs. This scheme alleviates the problem caused in conventional IP Switching in which flows that consistently recur cause the repeated establishment and removal of dedicated VCs. Short-term dedicated VCs are synonymous with the dedicated VCs set up in conventional IP Switching. Long-term dedicated VCs endeavour to achieve the longevity of the pre-established VCs set up in the control-driven approaches, in areas where there is enough traffic to warrant them.

4.2.10 Classification Precedence

In cases where a Smart IP Switch detects a flow that qualifies as a potentially long-lived flow, but also as a potential aggregate flow, it shall be set up as an aggregate flow. The reason for this is more potential benefit may be attained by switching an aggregate flow than by switching an individual flow.

4.2.11 VC Pool Management

A Smart IP Switch on initialisation will open up a specified number of dedicated VCs between each of its adjacent nodes. The use of a VC pool means Smart IP Switches can manage VCI allocation on a localised basis. The problem of cell interleave avoidance does not require the administration and distribution of VCI blocks to all potential ingress nodes as in the case of the VP merge solution employed in control-driven approaches. When a particular node is performing a cut-through operation for a flow type 3, it selects the next available VC from the VC pool. When the downstream node receives a packet on the *unassigned* incoming VC, it will update the source VC table of the associated FIB entry, and remove the VC from the pool. When the dedicated VC entry times out, it is put back in the pool. This localised VC pool management introduces a problem due to the fact that it is normally the downstream node that makes the decisions on which VC (label) to use for the next dedicated VC. The problem arises when an upstream and a downstream node select the same VCI for unrelated flows. The problem may be resolved by a simple strategy whereby, when an upstream node selects the next available VCI, it will do so on an odd VPI, and when a downstream node selects the next available VCI, it does so on an even VPI. If the number of available VCs on the pool drops below a specified threshold, more VCs should be added to the pool. The exact number VCs to keep available on the pool is a matter of policy.

4.2.12 CIDR Fall-back

It is possible that at any point in time, CIDR prefixes in different routers across a network may be inconsistent. It is likely that a particular CIDR prefix will become more specific (longer) as the destination network is approached. We use the term *CIDR fall-back* to describe the mechanism used in Smart IP Switching to handle this problem. It is based on the fact that the UFI is always generated on a localised basis, and the prefix mask it contains will be the longest match possible between the routing table and the FIB. It is possible that after passage of the first packet from ingress to egress there exists a number of dedicated VCs, each based on a UFI with a different CIDR prefix. In this case, the CIDR prefix based on the longest match between the routing table and FIB of the egress node has propagated to the next upstream node, the node that is one hop from the egress.

When the second packet traverses the same path, the same prefix will have propagated to the next upstream node, the node that is two hops away from the egress. This process will continue with one extra packet being expelled for every SIPS with an inconsistent CIDR prefix. Obsolete dedicated VCs that have been set up in the process will quickly expire since they will not be used, and therefore will not be refreshed.

5. Simulated Smart IP Switched Network

We have constructed an ATM network emulator using the Vendor Independent Network Control Entity (VINCE) [11] software package from the U.S. Naval Research Laboratories to validate the Smart IP Switching idea. This allows a network of simulated Smart IP Switched nodes to be set up and fed with test traffic. The package allows simulated networks of arbitrary topology to be constructed for test purposes. The simulated network topology we used is based on the Irish IP Research Network (HEANET) and is illustrated in Figure 5. The test system is capable of operating in three modes: a conventional routed network, an IP Switched network, and a Smart IP Switched network. Traffic traces from two sources were used. The first set (TEST SET 1) was reverse engineered from data captured at FIX-WEST, a major Internet interchange point, and available through [12]. In the absence of real routing table data, the routing tables of all SIPSs were populated with routes to the sites illustrated in Figure 5. The results based on monitoring traffic at SIPS *heuston.hea.ie* are shown in Table 1. This showed that IP Switching was able to divert 54.0% of IP packets from being routed, to being switched, with a control-packet overhead of 7.1%. Smart IP Switching showed a significant performance improvement over this with 99.2% of all IP packets being switched, with a control-packet overhead of 1.4%.

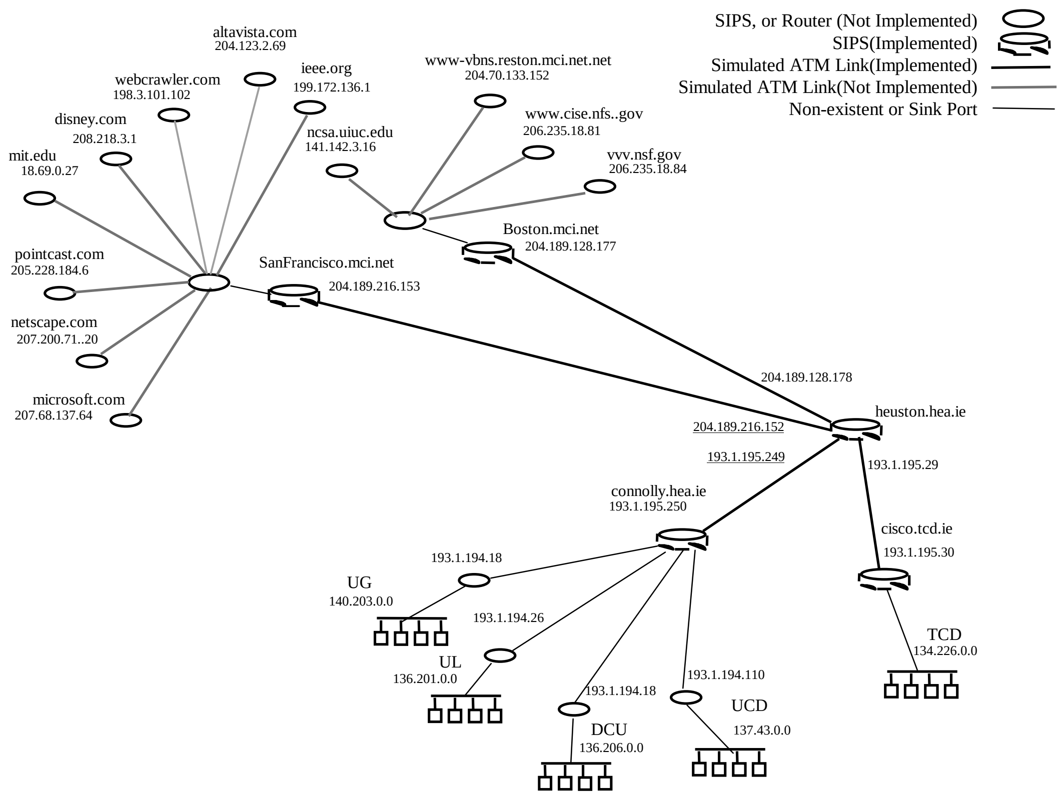

Figure 5: Simulated Network Topology

	IP Switching	Smart IP Switching
No. Packets Received (L3)	4573	123
No. Packets Forwarded (L3)	4756	198
No. Packets Switched (L2)	4790	8797
No. Control Packets Received	494	51
No. Control Packets Forwarded	677	126
Total Transit Packets	8869	8869

Table 1: TEST SET 1 Results (heuston.hea.ie)

In order to further validate the idea, a short traffic dump (TEST SET 2) was taken at an ingress point to the Irish IP Research network (HEANET). The results of TEST SET 2 are shown in Table 2.

	IP Switching	Smart IP Switching
No. Packets Received (L3)	4743	1256
No. Packets Forwarded (L3)	4836	1435
No. Packets Switched (L2)	4718	8862
No. Control Packets Received	235	802
No. Control Packets Forwarded	238	981
Total Transit Packets	9316	9316

Table 2: TEST SET 2 Results (heuston.hea.ie)

These results were not as favourable as TEST SET 1, however still showed significant benefit from the scheme. In IP Switching mode, 50.6%[2] of packets were switched with an overhead of 2.5% while in Smart IP Switching mode 95.1% of traffic was switched with an overhead of 9.5%.

These results show that by employing the Smart IP Switching scheme, the proportion of IP packets switched rather that routed is increased by more than 40%. The control traffic required in the scheme has increased on average by only 1.3%. Even taking into account that our first traffic-trace was reverse engineered, and both traces were short, the results show a significant improvement over conventional IP Switching, and approach the performance of control-driven approach.

6. Conclusions

We have devised a new scheme referred to as Smart IP Switching which takes a traffic-driven approach to the establishment of switched network paths in a manner modelled on Ipsilon's IP Switching. By introducing aggregation and long-lived flows, we have improved the

[2] This percentage may be smaller than might normally be expected, due to the small size of our traffic trace.

performance to a level that is comparable with control-driven approaches such as Cisco's Tag Switching while still retaining the underlying simplicity, scalability and reliability of the traffic-driven approach. Analysis of the system using test data derived from real traffic at two different points in the Internet confirms the systems ability to allow the building of scalable switched Internet backbones with improved performance.

References

[1] M. Laubach, "Classical IP and ARP over ATM", IETF RFC-1577, January 1994.

[2] The ATM Forum Technical Committee, "LAN Emulaton Over ATM, Version 1.0", January 1995.

[3] The ATM Forum Technical Committee, "Multi-Protocol Over ATM, Version 1.0", July 1997.

[4] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, and A. Viswanathan, "A Framework for Multiprotocol Label Switching", IETF Internet Draft, draft-ietf-mpls-framework-02.txt, November 1997.

[5] Y. Rekhter, B. Davie, D. Katz, E. Rosen, and G. Swallow, "Cisco Systems' Tag Switching Architecture Overview", IETF RFC-2105, February 1997.

[6] A. Viswanathan, N. Feldman, R. Boivie, and R. Woundy, "ARIS: Aggregate Route-Based IP Switching", IETF Internet Draft, draft-viswanathan-aris-overview-00.txt, March 1997.

[7] P. Newman, T. Lyon, and G. Minshall, "Flow Labelled IP: A Connectionless Approach to ATM", Proc. of IEEE Infocom, San Francisco, March 1996, 1251-1260.

[8] Y. Katsube, K. Nagami, and H. Esaki, "Toshiba's Router Architecture Extensions for ATM: Overview", IETF RFC-2098, February 1997.

[9] V. Fuller, T. Li, J. Yu, and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", IETF RFC-1519, September 1993.

[10] P. Newman, W. L. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon, and G. Minshall, "Ipsilon Flow Management Protocol Specification for IPv4, Version 1.0", IETF RFC-1953, May 1996.

[11] E. Hoffman, A. Mankin, and M. Perez, "VINCE, Vendor Independent Network Control Entity", U.S. Naval Research Laboratory (NRL), December 1993.

[12] National Laboratory for Applied Network Research (NLANR), Internet, http://www.nlanr.net/INFO.