PROJECT: ATM Forum Technical Committee

****

SOURCE: Ross Callon
Bay Networks
3 Federal Street
Billerica, MA 01821
phone: +1-508-436-3936
email: rcallon@baynetworks.com

Jason Jeffords
Cabletron
Technology Drive
Durham, NH 03824
phone: +1-603-337-7019
email: jeffords@ctron.com

Hal Sandick
IBM Corporation
800 Park Office Drive
Research Triangle Park, NC
phone: +1-919-254-4614
fax: +1-919-254-5483
email: sandick@vnet.ibm.com

Joel M. Halpern
Newbridge Networks Corp.
593 Herndon Parkway
Herndon, VA 22070-5241
phone: +1-703-708-5954
email: jhalpern@Newbridge.COM

****

Title: Issues and Approaches for Integrated PNNI

****

DATE: April 15-19, 1996 -- Anchorage

****

DISTRIBUTION: PNNI Subworking Group

****

ABSTRACT: This contribution gives a brief overview of technical issues related to use of Integrated PNNI (I-PNNI) as a routing protocol for support of the Internetwork Protocol (IP). This implies that PNNI would support routing for both ATM and IP simultaneously in an IP over ATM environment.

This contribution is intended to generate discussion and an initial consideration of issues. This contribution does not attempt to present a complete discussion of all issues. It is understood that further refinement and completion of details of the protocol will occur during the process of standardization of Integrated PNNI. It is expected that some of the details presented in this contribution will change during the process of standardization of integrated PNNI.

Support for other Internetwork Layer Protocols (such as

Appletalk, APPN/HPR, CLNP, DECnet, IPv6, and IPX) is for further study. It is felt that progress on integrated PNNI will be more straightforward if we first define support for one protocol (such as IP) and then extend I-PNNI to support other Internet Level protocols.

****

NOTICE: This contribution has been prepared to assist the ATM Forum. This document is offered to the Forum as a basis for discussion, and is not a binding proposal on the contributors, nor on any other company. The statements are subject to change in form and content after further study. Specifically, we reserve the right to add to, amend, or modify the statements contained herein.

****

1. INTRODUCTION

This contribution provides an overview of the major issues and detailed technical design of Integrated PNNI (I-PNNI) for use as a routing protocol for support of IP [5]. Section 1 provides a brief overview of I-PNNI. Section 2 defines terms used in this specification. Sections 3 and 4 give a general description of the protocol. Section 3 describes how I-PNNI works in a flat (non-hierarchical) network. Section 4 shows how I-PNNI can be extended for use in a multilevel hierarchy.

This contribution discusses support for IP only. Other internet level protocols (such as Appletalk, APPN/HPR, CLNP, DECnet, IPv6, and IPX) are for further study.

This contribution assumes that the reader is familiar with PNNI Phase 1 [1]. This contribution will make use of concepts, terms, and methods described in PNNI Phase 1, without any attempt to explain the PNNI Phase 1 mechanisms in this document.

This contribution provides an initial draft overview of a possible design of I-PNNI. It is expected that some of the methods presented here will change during the process of standardization of I-PNNI. Also, it is likely that many of the concepts and methods presented here may require more detailed and clearer description before the standard is finalized. The intention of this document is to give a detailed initial overview of mechanisms, in order to initiate discussion and help in the process of generating more complete and detailed future contributions.

It is intended that the I-PNNI specification would be developed as an "incremental delta specification", which would define the additional features which need to be used in addition to PNNI in order to support IP routing. We feel that it would be editorially impractical to attempt to re-write the existing PNNI specification as part of the I-PNNI specification, and that a reference is more appropriate. In addition, we expect that PNNI Phase 1 will be updated and enhanced in the future, and it would be highly desirable if the I-PNNI spec could be updated to refer to any new PNNI spec by updating the reference, rather than by trying to make parallel edits to both specifications. Note that this is the approach which was successfully used in the Integrated IS-IS specification [11].

1.1 What is I-PNNI?

Introduction and use of ATM requires operation of internet level protocols (such as IP) over ATM. Critical to the operation of IP over ATM is the issue of how routing is to be accomplished in this environment. Routing issues include how SVCs will be routed in the ATM subnetwork, selection of when and where SVCs will be set up, and routing of IP packets over the combination of legacy media and the ATM subnetwork.

This specification describes a routing protocol for use in the IP over ATM environment, based on an extension of the ATM PNNI routing protocol. This solution is known as "Integrated PNNI" (or I-PNNI) since it uses PNNI to integrate routing for ATM and for multiple internet level protocols running over ATM. Note that I-PNNI offers one option for routing in the IP over ATM environment. A description of the various alternate options is contained in a companion contribution [3].

PNNI Phase 1 [1] is a switch to switch protocol for use in private ATM networks. PNNI provides routing and signaling for ATM SVCs. PNNI provides powerful QOS routing capabilities, ease of configuration, topological flexibility, and scalability.

The approach outlined here is compatible with ATM switches which run PNNI Phase 1 [1], but which have no knowledge of internetwork layer protocols such as IP. It is not acceptable to require that normal ATM switches have any knowledge whatsoever of IP or other similar internetwork protocols. Fortunately PNNI has been designed to be very flexible and extensible, so that compatibility of I-PNNI with switches running PNNI Phase 1 is straightforward.

I-PNNI is intended specifically for operation in the IP over ATM / Multi-Protocol Over ATM environment. Thus integrated PNNI offers support for the capabilities and features that are likely to be important in this environment. I-PNNI is therefore compatible with the use of virtual networks, such as are provided by LAN Emulation [9] and RFC 1577 Logical IP Subnets [10]. Similarly, I-PNNI is compatible with both "conventional routers" (i.e., routers in which the route computation function and the IP forwarding function are co-located in one physical device), as well as "virtual routers" (i.e., routers in which the route computation is performed in one physical device known as a route server, and the IP forwarding function is located in different physical devices known as edge device forwarders).

1.2 Interaction with ATM-Attached Hosts

In general, I-PNNI does not change nor dictate the methods that hosts use to interact with the IP or ATM networks. The vast majority of hosts should not be directly involved in the operation of routing protocols (neither internet level nor ATM level routing protocols). Hosts will behave in accordance with other standards. Legacy attached hosts will make use of existing IP standards. ATM-attached hosts will make use of standards such as UNI, LAN Emulation, NHRP, RFC 1577, and/or the MPOA Host behavior. The host behavior will therefore be independent of which approach is used for routing in the IP over ATM environment.

In many cases ATM-attached hosts will make use of NHRP in order to obtain IP address to ATM address mappings for other hosts. In general, such hosts will forward NHRP queries to a router or route server. Routers and route servers which are running I-PNNI will make use of the routing information obtained from I-PNNI, along with other information such as caches obtained from previous NHRP queries, to either answer the NHRP query directly or to forward the NHRP query to an appropriate next hop router / route server.

For short term transactions it may be undesirable to incur the overhead of a query plus a call setup, and the host may instead forward IP packets directly to a router without any prior associated NHRP query. Similarly, hosts which do not use NHRP may forward packets directly to routers.

In some unusual cases hosts may choose to participate in I-PNNI. This is expected to be a relatively rare occurrence. For example, in most cases it would be undesirable to have each host appear in the topology state database. However, it is possible that a small number of major servers may choose to run I-PNNI in order to optimize routing to and from the servers, and to eliminate the need to query for the location of the server. It is straightforward for I-PNNI to support this behavior. If a host chooses to run I-PNNI, then it announces itself as "transit restricted" for SVCs (using the standard PNNI

encoding), and also as "transit restricted" for IP packets. It also announces reachability to its own IP address(es) and ATM address(es).

1.3 Example Network

An example network is illustrated in this section, with the intention that latter sections can refer to this example network in explaining specific details of I-PNNI. In figure 1 there are ten routers, numbered R1 through R10, as well as three ATM switches A, B, and C. Integrated PNNI allows a single routing protocol to be used in this environment. Thus I-PNNI is used for routing between ATM switches (e.g., A, B, and C) and also for routing between routers (e.g., R1 to R11).

Route servers and edge device forwarders are not explicitly illustrated in figure 1. This is in order to simplify the example, and is not meant to minimize the importance of route servers and edge device forwarders. In fact, the routers illustrated in figure 1 may be either conventional routers (in which the routing protocol function and IP forwarding function are done in the same physical device) or virtual routers (in which the routing protocols are done in route servers, and the IP forwarding is done in edge device forwarders). Similarly, the protocols must support routing to virtual networks that may be distributed amongst multiple routers and amongst hosts attached to a combination of ATM and legacy LAN segments.

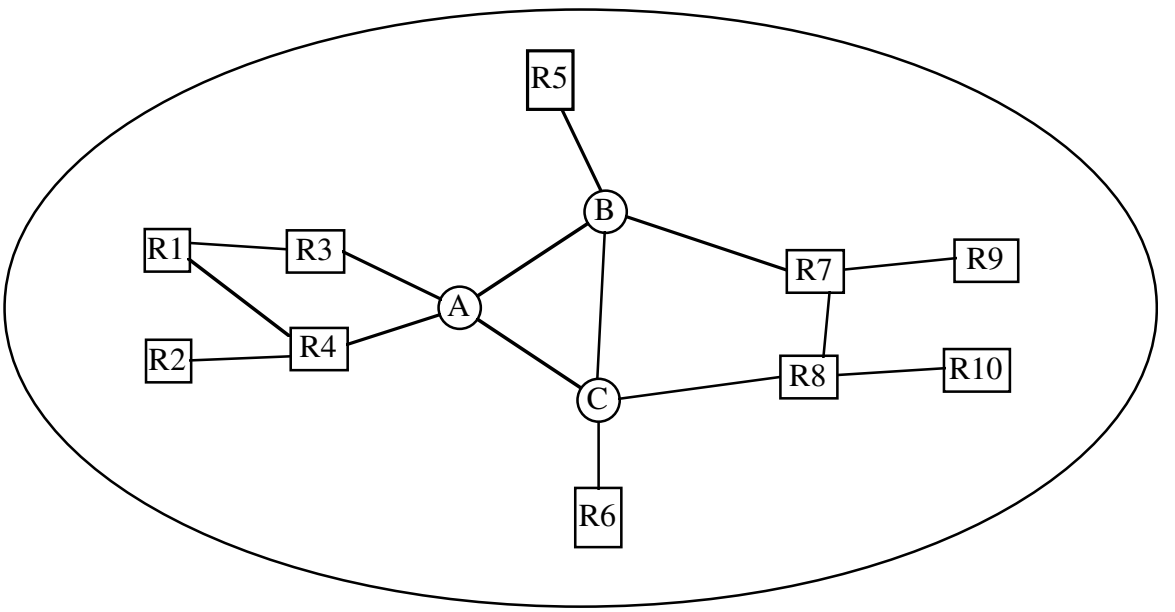

Figure 1: A Group of Switches and Routers

2. TERMINOLOGY

ATM-Attached Host
A host which has a direct physical attachment to a ATM network.

ATM Switch
A physical device which forwards ATM cells between virtual circuits on one or more physical ATM links, according to virtual circuit identifier (VPI and VCI) in the header of each cell. The device may
or may not participate in running UNI and/or NNI in-band signaling
and routing protocols.

Conventional Router
A physical device which is capable of forwarding internet layer packets (such as IP, IPX or IPv6 packets) based on Internet layer information, and which

also participates in running one or more internet layer routing protocols (such as OSPF, I-PNNI, etc.). A conventional router is equivalent to a single physical device which contains the functions of both a route server and a non-routing edge device.

Edge Device Forwarder
A physical device which is capable of forwarding Internet Layer packets (such as APPN, IP, IPX, or IPv6 packets) based on Internet layer header information, but which does not participate in any internet layer routing protocols. A forwarder must get its forwarding table information from a Route Server.

Host
A system which is acting as the end user of data communications services (i.e., either ATM services and/or Internet Layer Protocol services).

Integrated Routing
A routing model which uses a single routing protocol, I-PNNI, for routing at both the ATM and Internet Layers.

Layered Routing
A routing model in which routing for ATM is kept as independent as possible from internet layer routing. For example, PNNI Phase 1 may be used for routing of ATM SVCs, while OSPF may be used for routing of IP packets and NLSP may be used for routing of IPX packets. This approach requires that the internet layer routing protocols model the operation of the ATM network, for example as a LAN, as a point to point link, or as a combination of multiple LANs and/or point to point links.

Legacy-Attached Host
A host which has a direct physical attachment to a "legacy" network, i.e., any network other than an ATM network.

Legacy Network
Any network which is not ATM.

Peer Addressing Model
An addressing model in which internet layer addresses are algorithmically mapped to ATM layer addresses.

Route Server
A physical device which runs one or more internet layer routing protocols (such as APPN, OSPF, I-PNNI, etc..), and which provides forwarding information directly to non-routing edge devices.

Separated Addressing Model
An addressing model in which internet layer addresses are independent of ATM layer addresses. This implies that a mapping function will be needed in some cases to map from an internet layer address to the corresponding ATM address.

Virtual Router
A set of physical devices which together offer the function of one or more routers. A virtual router will consist of a route server and one or more non-routing edge devices.

3. I-PNNI MECHANISMS IN A FLAT NETWORK

In order to simplify the description of I-PNNI functions, those functions which are needed for hierarchical operation are listed separately in section 4. This section specifies the functions which are necessary for operation in a flat (non-hierarchical) network, or within one peer group in a larger network.

3.1 General Overview of PNNI Operation

I-PNNI allows a single instance of PNNI routing to be run by both routers and switches. PNNI therefore is run between a switch and neighboring switches,

between a switch and neighboring routers, and between routers. Thus all ten routers in figure 1 (including routers which have no ATM interface) and all three switches run the normal PNNI routing protocol, and all appear as "nodes" in the topology database. The interface between a router and a switch (such as the link between switch A and router R3 in figure 1) run the PNNI protocol and operates as a normal PNNI interface. PNNI Hellos and PTSEs/PTSPs are exchanged between all nodes illustrated in figure 1. Similarly, all nodes and links are advertised using normal PNNI metrics. Thus the router to router links are advertised using the same metrics as are used for inter-switch links.

This implies that router to switch link is considered to be a normal PNNI link. A switch thinks that neighboring router is another switch. An SVC from a router via switches to another router has its DTL set by the router which originates the SVC.

3.2 Avoiding ATM Routing via Routers

Given that I-PNNI is run between all nodes in the combined router/switch topology, it implies that all nodes see the complete topology. However, it clearly is not acceptable for switches to route SVCs *through* routers (i.e., an SVC may end at a router, but it cannot traverse a router to reach a node on the other side of the router). For example, suppose that the link from switch B to switch C in figure 1 is congested, and cannot forward any additional SVCs. Suppose furthermore that switch B is processing a call request to a called address reachable via switch C. In this case switch B may choose a path via node A. However, the switch may not choose a path via nodes R8 and R9.

Routers participating in I-PNNI advertise themselves as "transit restricted" nodes (this uses the standard PNNI encoding in order to advertise the transit restricted status, as specified in [1]). The definition of "transit restricted" is that the node should be treated as non-transit, except that other information included in the PTSE may override this indication, provided that this other information indicates the conditions under which the node can be used for transit. Thus, for example, the IP-specific information included in an I-PNNI PTSE could specify that the node can forward IP packets, even if it is nontransit for the purposes of SVCs.

3.3 Exchange of Hellos

The exchange of Hellos is essentially unchanged from PNNI. Each node (whether a router or an ATM switch) exchanges PNNI Hello packets with all neighbors. This allows each node to determine the identity of the neighbor, and the status of the link. In addition, this allows each node to determine whether the neighbor is in the same peer group, thereby controlling the scope of transmission of PTSEs.

Database synchronization between neighboring nodes is similarly unchanged from PNNI.

3.4 Flooding of PTSEs

All nodes (including switches and routers) announce links to neighbor nodes in the normal manner in PTSEs. The PTSEs are flooded throughout the peer group in the normal manner (i.e., each node floods each new PTSE to all neighbor nodes which are in the same peer group).

Routers are eligible to be elected Peer Group Leader just like any other node operating PNNI routing. However, if I-PNNI is to be used in a multi-level hierarchy, the PGL needs to be I-PNNI knowledgeable (i.e., the PGL may be either a router or a switch; however, the PGL needs to be running software which is knowledgeable about IP and I-PNNI). In general this may be accomplished either by giving the ATM-attached routers a higher leadership priority (relative to switches), or by putting I-PNNI knowledgeable routing software on switches.

3.5 ATM Addresses for Routers to Use in I-PNNI Packets

I-PNNI implies that IP routers will transmit I-PNNI packets. This implies that there are a number of places in which routers will need to use ATM addresses. For routers which do not have ATM interfaces, this implies that they will somehow need to obtain ATM addresses.

There are a number of ways that this may be accomplished. Note that the ATM private address space is relatively large, and there is no shortage of ATM private addresses. Thus any administration which is running ATM at all, and which therefore needs to obtain ATM private addresses for some systems, should be able to obtain a large enough block of addresses to make assignment of addresses to all systems in the network practical. Also, we might assume that an IP routing domain needs to have IP addresses for use by routers which is unique at least in the context of the set of routers in one domain. Thus we could define a simple mapping of IP addresses into ATM private addresses solely for the purpose of providing one way to obtain unique ATM private addresses for the routers to use in the operation of I-PNNI. Note that an approach similar to this was employed by Integrated IS-IS (which provides a manner to map IP addresses into OSI NSAP addresses, solely for the purpose of providing one possible feasible way to obtain a unique NSAP address given a unique IP address). In fact, given that NSAP addresses and ATM private addresses use the same address space, we could consider using precisely the mapping as used by Integrated IS-IS.

3.6 Advertisement of Reachability

All nodes participating in PNNI announce reachability to those addresses that they can reach. ATM switches will announce reachability to the ATM addresses that they can reach, based on prefixes of 20-byte ATM private addresses. Similarly, routers announce reachability to those IP addresses which they can reach, based on prefixes on 4 byte IP addresses (note: extension to 16 byte IPv6 packets should be straightforward, but has been left for future work).

ATM-attached routers running I-PNNI will also announce that they can reach their own ATM private address. For example, consider router R3 in figure 1. It has an ATM interface, corresponding to its link to switch A. Note that switch A thinks that this is a normal PNNI interface. Switch A therefore does not announce reachability to R3's ATM address. Rather, R3 announces reachability to this address. In the case of multi-homed routers, this allows the multi-homed router to optionally have only a single ATM address, and results in better routing of SVCs destined to that router in some cases.

We propose that IP addresses be carried independently of ATM addresses. Maintaining independence of addressing allows ATM and IP addresses to be assigned independently, allows ATM switches to ignore IP reachability information (other than the need to store PTSEs from all nodes in the peer group), and avoids any potential confusion between the two address spaces. The "normal mode" operation of I-PNNI supports the separated addressing model. However, support for peer model addressing is possible, as discussed below.

Reachability to IP addresses is announced using TLV-encoded extensions, which will be ignored by switches running normal PNNI. These extensions make use of the TLV extension capability defined in PNNI Phase 1. Also, the attributes defined in PNNI are used to indicate that a switch which does not understand the extended field (i.e., which does not understand I-PNNI) can safely ignore the information (i.e., pass the information on unchanged if included in PTSES to be forwarded, store the information for purposes of database synchronization, but don't use the information).

3.6.1 Direct Reachability

Clearly it is necessary to allow routers to announce that they can directly reach particular IP addresses (corresponding to host routes, subnets, or summary routes). This is done in the same manner as other IP routing protocols (specifically OSPF and I.IS-IS), by announcing a prefix of an IP address. This information is included in PTSEs initiated by a router.

Traditionally, IP address prefixes have been advertised by specifying a 32 bit IP address plus a 32 bit mask. In theory this would allow for non-contiguous masks, or for address masks which are not "high-order-justified". However, for a variety of very good reasons the IETF/IESG/IAB a number of years ago required that addresses must be assigned such that masks are always contiguous and high-order-justified. Therefore, we propose that IP prefixes can be advertised using an IP address plus a count of the number of bits in the address which are considered significant. This is consistent with current IP address usage.

When a router announces direct reachability to a particular address prefix, this implies that the router is directly attached to the associated IP hosts or subnet. In order to deliver the packet to the associated destination(s), the IP packets may be transmitted to the router directly.

3.6.2 Support for Query Reachability

"Virtual Networks" is a networking model which allows a system's IP address and logical IP connectivity to be independent of the system's location in an ATM network (although any given virtual IP network may optionally be physically restricted to some range of an ATM network). Examples of virtual networks are provided by LAN Emulation, and by RFC 1577 Logical IP Subnets.

The use of virtual networks implies that a host may be located on a particular logical IP subnet, while a logically neighboring host (with an adjacent IP address) may be located in a different part of the actual physical ATM network. This implies that a router which advertises reachability to a virtual network might be physically close to part of that virtual network, but might be physically remote from other parts of the virtual network. In addition, even if a router is physically close to one particular part of a virtual network, for ATM-attached hosts direct reachability via an ATM connection to the host may in some cases be preferable. For example, this may provide higher bandwidth, lower delay, or better QOS support than would be provided using indirect connectivity via a router.

For this reason, an address mapping query-response function (using NHRP) may be used to map directly from the destination host's IP address to the host's ATM address. In this case, optimal routing to a specified IP address may require that an NHRP address resolution query should be directed to the appropriate router (i.e., the router which is announcing reachability to the destination subnet), which will in turn reply providing the ATM address to use for the best path to that host.

Note that query reachability is also useful in hierarchical networks. Here a single real physical system (the peer group leader) will be transmitting PTSEs describing the capability of a logical group node (LGN). Thus, the PTSEs associated with the LGN may advertise summary reachability to multiple IP systems using one or more IP address prefixes. However, the optimal ATM address to use to reach any possible IP address might not be advertised in the summary reachability advertisement, and instead the optimal ATM address to reach any one particular IP address may be determined by using an NHRP query.

In addition to the use of virtual networks, there are likely to be some routers (whether ATM-attached or not) which are running I-PNNI and that have direct reachability to "real" physical IP subnets. In these cases a query would not be useful.

This implies that it is highly desirable for the routing protocol to distinguish between routers which have direct physical reachability to an IP subnetwork (for which the IP packet should be transmitted to the router), versus routers or router servers which have indirect reachability (for which an NHRP address resolution query may be directed to the router if an optimal path is required). Integrated PNNI therefore distinguished between these two types of reachability.

If a router announces "query reachability" to a particular IP address, this implies that the router is capable of reaching the associated IP hosts or subnets, but is not necessarily on the optimal path to the associated host or subnet. Thus a more optimal route may be achieved if an NHRP query is first used to determine the optimal ATM address to use to reach any particular IP address which matches the advertisement.

For very short term transactions, it may not be worth the delay and processing implicit in sending a query, waiting for a response, and then setting up an SVC. For this reason, even when query reachability is advertised, packets may be sent to the router for delivery to the associated IP address. However, better long term service will be obtained if first an NHRP address resolution query is sent to the associated router, the router responds with the correct ATM address to use for the IP destination, and then an SVC is set up to the specified ATM address in order to forward the IP packet.

Note: We might want to consider the case of virtual networks which are running over other media other than ATM. For example, a virtual Ethernet may be distributed amongst multiple bridge-like devices, which are interconnected via an FDDI ring. Whether this is worth considering for I-PNNI is an issue for the PNNI working group to consider.

3.6.3 In-Care-Of Addresses

There may be some cases where a router or route server wants to advertise reachability on behalf of another system. For example, a router may know of a major server which is either attached to a virtual network served by the router. Similarly, a logical group node may want to advertise the address of a major server which is included in a lower level peer group. In this way, for those "popular" addresses which are frequently contacted by other systems, it is possible to avoid the need for a query/response.

In order to announce "in care of" reachability to a particular IP address or address range, the router advertises the ATM address that should be used to deliver IP packets to the specified destination or group of destinations. In this case the router advertises an IP address prefix plus a single ATM address. In order to deliver packets to any IP destination which matches the specified address prefix, an SVC may be set up to the specified ATM address.

Note: Support for "mapped addresses", in which the ATM address is an algorithmic transformation of the IP address, is for further study.

3.7 Maintenance of SVCs between Routers

In order to forward IP packets between routers via the ATM subnetwork, it is of course necessary to make use of SVCs that traverse the ATM subnetwork. There is an issue to be resolved regarding when and how to set up such SVCs. There are a wide range of options here. The two extreme ends of available options may be: (i) Such SVCs may be set up only when required based on IP packets waiting to be forwarded; (ii) Such SVCs may be set up on entirely an a priori basis.

Option 1: Set up on demand

> I-PNNI allows routers to view the entire network topology, including routers and ATM switches. This implies that even if no SVCs are set up, an optimal end to end route may be computed across the network topology, and an IP packet may be forwarded along the optimal path. When an ATM-attached router needs to forward the packet over the ATM subnetwork, the route computation will allow computation of the next router across the ATM subnet. If the router already has an SVC set up to the next hop router, then the packet is forwarded. Otherwise the SVC is set up on demand.
>
> With this method, over time a set of SVCs will be set up which allow the majority of packets to use existing SVCs. Under used SVCs may be closed on

a timer driven basis (what it means to be "under used" will be a function of the number of SVCs that the network can support).

This approach has a potential problem in that the "first packet" to any particular next hop router may experience considerable latency while waiting for the SVC to be set up. Also, in general there may be limits on the number of SVCs which can be set up in a given period of time, and on the number of SVCs which may simultaneously be in use. This method could therefore suffer from robustness problems if the required SVC cannot be set up for a particular packet.

Option 2: Set up a priori

With this method, the routers open up a partial set of "a priori overlay SVCs" amongst themselves immediately upon network initialization, before the arrival of any specific traffic. The establishment of SVCs is based on the knowledge gained from running I-PNNI, and allows creation of a set of SVCs that span the ATM network. For any two routers X and Y which are participating in I-PNNI in the same peer group there might not necessarily be a direct SVC from X to Y. However there will be some path from X to Y using only other routers in the same peer group as intermediate hops.

There are a number of ways that this default set of SVCs may be set up. One method that has been proposed is to base this upon the router ID of each router. For example, consider the router ID as a circular number space. Each router may set up a direct SVC to the "k" next higher numbered routers. In the example, if k equals 1, then R3 would set up an SVC to R4. R4 would set up an SVC to R5, etc. Finally R8 (the highest numbered router on the ATM network) would set up an SVC to R3.

Note that these two options are actually two extremes, with a continuum of possibilities in the middle. For example, one option would be to use the default set of SVCs for best effort IP traffic, but use an optimal route based on the complete (legacy and ATM) topology for QoS traffic (using a "custom" SVC if necessary to traverse the ATM subnetwork).

We believe that it will be highly desirable to use the second method for best effort IP traffic. This will ensure that some path is available to any particular IP destination. This is desirable for short duration IP flows (for which it is not worthwhile to set up an SVC), and to assume the reliable existence of a path. For longer duration flows and for QoS traffic it is likely that the optimal path should be set up on demand.

Point to multipoint SVCs may need to be set up for efficient support of IP multicast. The best method of supporting IP multicast using I-PNNI is an important area for further study.

3.8 Route Computation for Best Effort IP Traffic

As described above, all routers and switches within a peer group obtain complete topology of the peer group, including knowledge of which reachable address prefixes are reachable via each node. Also, switches have information which leads them to believe that the routers are non-transit for forwarding SVCs. This information therefore gives both routers and switches the information that is necessary to route calls and packets correctly.

Within a peer group, so long as directly reachable addresses are advertised, this also provides the information necessary to perform IP to ATM address resolution. For example, suppose that in figure 1 there is an IP packet at router R8 which is destined to a host which is attached to router R1. R9 can determine (based on a standard routing computation) that the correct path is via switch C, to switch A, to R4, to R1. This implies that R9 can either use an existing SVC (if an SVC is already set up between R9 and R4), or can open an SVC to R4. If it is necessary to open a new SVC, then given that the routing computation computes the entire path from R9 to R1, the routing

computation also automatically determines the address to which the SVC should be setup.

In hierarchical situations there may be a need to use NHRP queries. This is discussed in section 4. Also, where virtual networks are used, a more optimal path may be provided if NHRP queries are used.

The route computation to be used may differ for: (i) Best effort IP packets (using hop by hop routing); (ii) QOS-specific IP packets using hop by hop routing; and (iii) QoS-specific IP flows using some sort of route setup. However, the methods used for QoS route computation is related to other issues of QoS support, which are discussed in the following section (3.8). This section will only consider the route computation for best effort IP packets.

For best effort IP traffic, I-PNNI uses hop by hop routing. This implies that the routers within a peer group must perform a consistent route computation, in order to protect against routing loops (this requirement is the same as occurs with any link state routing protocol, including OSPF). For this reason I-PNNI specifies the standard route computation to be used for best effort hop by hop IP packets.

It is proposed that the standard route computation for best effort IP packets consist of a simple Dijkstra computation on the administrative weight.

There is an issue regarding whether existing SVCs should be advertised into integrated PNNI (and by implication, whether they will be considered in the best effort route computation). In general the latency associated with setting up a new SVC is multiple orders of magnitude greater than the time required to forward an IP packet. It is therefore highly desirable for IP packets to use pre-existing SVCs. By advertising existing SVCs into I-PNNI, and treating such SVCs as a normal link for the purpose of the Dijkstra computation, we can maximize the probability that a packet can be forwarded without requiring a new SVC be set up. Depending upon which SVCs are set up on an a priori basis, and the metrics associated with the links implied by the set up SVCs, it is in fact possible to ensure that IP packets will always be forwarded over existing SVCs.

If the route computation considers the set of existing SVCs only for the best effort computation, then the amount of information required by advertise the SVC can be minimized by using an I-PNNI-specific format, with only the administrative weight metric advertised.

3.9 QoS Support

One of the potential benefits of I-PNNI is that it provides the first fully QoS-aware routing protocol for IP. However, there is nothing in the current IP stack to make use of this capability. The RSVP mechanisms being defined to date for IP assume that the packets will follow default IP routing. Of course, one of the reasons for this assumption was that there were no protocols which could do any better.

Therefore, having the capability to perform QoS sensitive path selection for IP is a necessary enabler for actually delivering such a capability to the users. There are a number of issues that will need to be addressed when that is being done. This brief sketch is provided only to give a perspective on what is possible. It is quite likely that much of the work required to fully utilize QoS sensitive routes (for example, extensions to RSVP) will most appropriately be standardized within the IETF.

3.9.1 Packet Classification with Hop by Hop Packet Forwarding

Given the ability to calculate QoS-sensitive routes for IP packets, it is necessary to determine which specific IP packets are to be mapped to which QoS routes.

One possible approach, within a single organization, is a-priori packet

classification, used with hop-by-hop forwarding. If all of the routers within the organization's network classify packets the same way, and if all of them have the same routing classification-sensitive routing algorithm, QoS routing can be deployed without any changes to the host software. There are nonetheless a number of issues to deal with. One small problem is that in order to ensure consist routing amongst all routers in the network, the packets would probably need to be classified into a relatively small number of service categories. Given the limitations of service categories, it is likely that such support would end at the administrative boundary of an organization.

Additionally, with strict packet classification, there is a need to be careful about what metrics are used. Experience has shown that hop by hop forwarding of connectionless datagram traffic with dynamic traffic sensitive metrics (such as ACR) can produce unstable systems. Delay or Maximum Cell Rate sensitive routing would not have this problem. Care would have to be taken before performing hop by hop routing using the available cell rate (given that ACR must by definition vary according to traffic load).

Finally, there is some difficulty in ensuring that all routers have a consistent view of the classification and routing algorithms. This is the aspect of this solution that may be amenable to standards work, if it is determined to be useful upon more detailed analysis.

3.9.2 Flow setups

Following the model of RSVP, an alternative approach to the QoS routing difficulties is to use a flow-setup mechanism. This is a close relative of both the ATM routing work, and the ongoing IETF integrated services approach. Use of some sort of flow setup and state maintenance would allow use of source routing and of traffic sensitive metrics, thereby greatly increasing the flexibility of QoS sensitive routing. While the IETF has not yet decided that it is interested in this topic, one can see how it could be approached.

The RSVP-like flow setup could use some sort of maintenance mechanism to maintain and verify the QoS routing selected path. This mechanism would range in dynamics from the RSVP soft-state through various intermediate mechanisms up to a full VC maintenance as we use with ATM. Additionally, packets could be marked by the sender as requiring QoS routing. This would ensure that transient routing changes did not produce any replicating QoS sensitive multi-cast packets.

There are other alternatives being investigated by researchers in this area. The above just suggest that there are ways that IP could be enhanced to make use of a QoS routing capability once one is provided.

3.10 Pseudonodes and Designated Routers

Integrated PNNI may be used over broadcast media such as Ethernets and other local area networks. This implies that integrated PNNI requires support for designated routers and pseudonodes (in the same manner that these are supported in OSPF and IS-IS).

However, the operation of pseudonodes and designated routers is local to a LAN. From other nodes in the peer group, a LAN plus "n" routers with interfaces to the LAN would appear as n+1 normal nodes (one of which represents the LAN, the other n of which represent the n routers). Thus, in order to run integrated PNNI over broadcast LANs requires that support for pseudonodes be specified in the Integrated specification, and that routers support these features. However, this has no impact on the basic ATM PNNI specification, and no additional "hooks" are needed in PNNI in order to support this capability.

Figure 2 illustrates an example in which Routers R1, R2, and R3 are interconnected via a broadcast LAN.

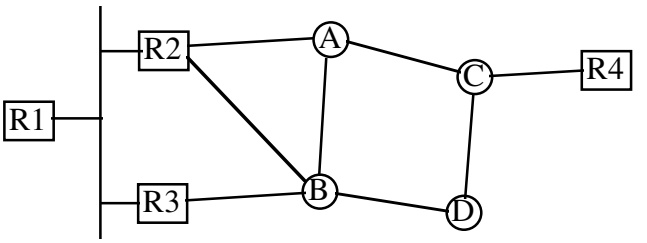

Figure 2: Network with Broadcast LAN

In this case, routers R1, R2, and R3 need to implement specific functions (designated router election and announcement of a pseudonode). The resulting topology (as announced in link state
Updates) is illustrated in figure 3.

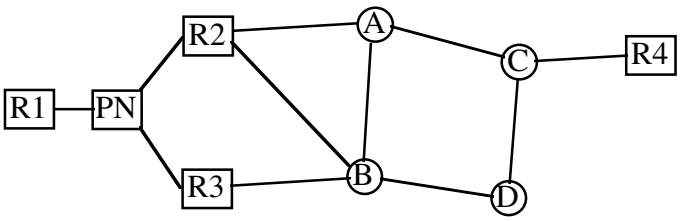

Figure 3: Representation of Broadcast LANs

From the perspective of A, B, C, D, and R4, the topology looks exactly as if a normal multiprotocol router ("PN") had point to point links to R1, R2, and R3.

Routers run an election algorithm on the LAN, using information contained in Hellos transmitted over the LAN. This results in one node being elected designated router for that LAN. This one node then broadcasts the I-PNNI PTSEs corresponding to the pseudonode (the node which represents the LAN). The details of this are for further study. However, we expect that the details can be taken almost verbatim from the OSPF specification.

3.11 External Routes

This will follow the normal IP method: routers are able to announce "external routes" for IP address prefixes outside of the I-PNNI routing domain. On the most part this will make use of the same capabilities (and probably the same formats) as are available with OSPF.

There is one special case which appears with IP over ATM: One could imagine a case where the actual external connectivity is via an ATM switch rather than a router. In some cases external traffic will need to go via the router anyway in order to go through appropriate packet filters. However, in some cases the traffic may be able to go via SVCs which terminate outside of the routing domain. In these cases determination of the address to be used for setting up the SVC will in many cases need to make use of an NHRP query with is transmitted to the domain boundary router, which in turn may forward the NHRP query to other routes outside of the domain. Thus, for external routes it is desirable to be able to distinguish "direct reachability" from "query reachability" in an manner which is very similar to the same distinction attached to internal routes.

3.12 Encapsulations

3.12.1 Encapsulation for IP Packets transmitted over ATM

For IP packets transmitted over point to point links corresponding to SVCs set up by I-PNNI, this uses RFC 1483 encapsulation.

3.12.2 Encapsulation for IP Packets transmitted over Legacy Media

This is outside of the scope of this contribution, and similarly is outside of the scope of the future I-PNNI Specification.

3.12.3 Encapsulation for PNNI Packets Transmitted over ATM

When transmitting PNNI protocol packets over ATM, the standard ATM Forum PNNI encapsulations will be used. This is necessary in order for I-PNNI to inter-operate with ATM switches which operate according to the ATM Forum Phase 1 PNNI Standard [1].

3.12.4 Encapsulation for PNNI packets over legacy media

I-PNNI can operate over legacy media. This implies that we need to define the encapsulation of PNNI packets over legacy media. There are two options that can be used here:

(1) PNNI packets may be embedded in IP packets.

This approach has the advantage that IP can operate over essentially any media, and thus by defining the PNNI encapsulation in IP we have defined the encapsulation over a wide range of media types.

(2) PNNI packets may be encapsulated directly over link layers.

This would require that an encapsulation be defined for each relevant link layer. A significant number of link types can be handled by defining a SNAP/SAP.

We feel that either of these approaches would be feasible, and that one approach should be chosen during the process of standardization of I-PNNI.

3.13 Private Virtual Networks

We could imagine a case where a backbone or carrier network is carrying IP traffic on behalf of multiple campus networks or customer networks, where the IP addresses used in the campus networks are not globally unique. In this case, it is likely that there would need to be some way to identify the different IP address spaces, and that the reachability information carried on behalf of the campus networks would need to be kept independent of other uses of addresses in I-PNNI. For example, there may be a special reachability information which would say "There is this special addressing region 'XXXXX' which uses private addressing, we can reach IP address 'yyyyy' in that addressing region".

Note that other IP routing protocols do not have special mechanisms for support of PVNs. However, such support would be valuable in any new IP routing protocol (particularly given the current issue with the limited IP address space). Support for private virtual networks should be discussed and considered during the process of standardization of I-PNNI.

4 HIERARCHICAL OPERATION

I-PNNI uses the same multi-level hierarchy mechanisms as are defined for PNNI. Some additional details need to be discussed, but the hierarchical operation is kept as much like the mechanisms of PNNI as possible. ATM switches again operate precisely as is defined for PNNI.

An example hierarchical I-PNNI network is shown in figure 4. Here there are three peer groups X, Y, and Z. As in PNNI, each peer group is summarized into a single higher level logical group node (corresponding to nodes X, Y, and Z, respectively).

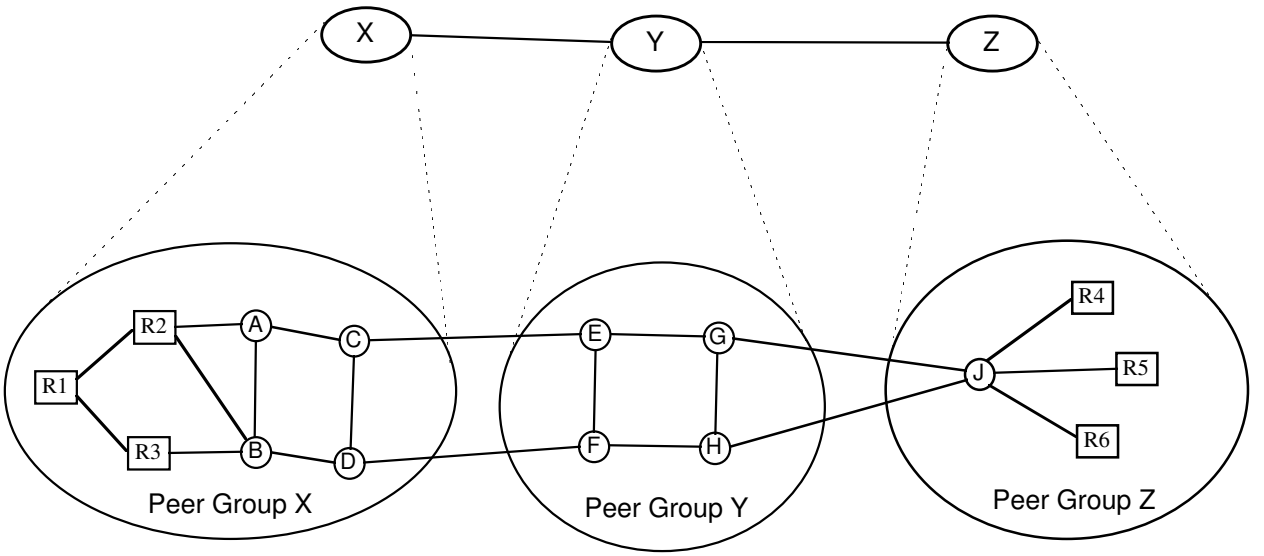

Figure 4: Example Hierarchical Network

In I-PNNI, each peer group may include a combination of ATM switches and routers. This implies that the corresponding logical group node needs to represent the capabilities of a combination of ATM switches and routers. For example, the reachability advertised by a logical group node may include reachability to both ATM addresses and IP addresses.

Hierarchical operation of I-PNNI is basically straightforward, except for a few "nits" which need to be resolved. Some of the issues which require a bit of are include:

- How to ensure that the path for an SVC never includes a router as an intermediate hop, even in the case where the path crosses a logical group node which is representing a peer group including a mixture of routers and ATM switches.

- How to ensure that the called address in SVCs between routers represents the correct address at which a corresponding IP packet should leave the ATM cloud.

- How to ensure that the PTSEs transmitted by logical group nodes accurately describes the characteristics of the corresponding peer group.

- The details of protocol operation need to be specified.

These issues are discussed below. However, first we point out some useful background information and consider a few "ugly points" and discuss how they can be resolved.

4.1 Types of Peer Groups

For proper operation of I-PNNI in a hierarchical network, it may be useful to classify peer groups into three categories. Depending upon how details of the protocol are defined, this may be useful for the purposes of hierarchical summarization of peer groups, and for aiding in the opening of SVCs between peer group leaders (both of which are discussed below). Graceful migration between categories is possible.

4.1.1 Classification of Peer Group Types

Issue: In the following categories we have not considered devices which are both routers and switches. Assuming that such devices will exist in the future, it is probably feasible for them to take the part of either routers

or switches in any of the following peer group types. However, we need to think more about this.

The three types of peer groups are:

(1) Pure ATM Peer Groups

A Pure ATM Peer Group operates using the standard PNNI Phase 1 specification. All nodes in the peer group are ATM switches. Similarly, the peer group leader can be a normal PNNI ATM switch. No knowledge of I-PNNI is needed by any node in the pure ATM peer group.

Note, in fact it is acceptable for some I-PNNI routers to be placed in a pure ATM Peer group. For example, this might be done at the beginning of a transition period. This is acceptable provided that the set of ATM switches in the peer group are contiguous (in order for ATM forwarding to work), provided that all border nodes are switches, and provided that the routers don't need to communicate via I-PNNI with any system outside of the peer group. If later the routers do need to communicate via I-PNNI with routers outside of the peer group, then the PGL will need to become I-PNNI knowledgeable, implying that the peer group becomes a "mixed" PG.

If a pure ATM peer group contains any logical group nodes, these must describe lower level pure ATM peer groups.

(2) Mixed Peer Groups

A mixed peer group may contain a mixture of ATM switches (which implement standard Phase 1 PNNI) plus routers (which operate I-PNNI). The set of ATM switches must be contiguous (in order for SVCs to operate between any two switches). The Peer Group Leader must be I-PNNI knowledgeable, and also must be attached to ATM (for example, the PGL may be a Router with an ATM interface, or a Switch running I-PNNI, or a workstation with ATM interface). Any combination of routers and switches may be border nodes (subject to the issue described in section 4.3.2 below).

If a mixed peer group contains any logical group nodes, these may represent any combination of lower level pure ATM peer groups, mixed peer groups, and router centric peer groups.

(3) Router-centric Peer Groups

Consider a peer group that does not meet the requirements for the first two types of peer group. Thus either the ATM switches within the peer group are not connected, or the PGL is not ATM attached. In this case, suppose that there are multiple ATM switches which are border nodes in the peer group, with outside links to ATM switches in some other peer groups. In this case, there are two problems which can occur: (i) If the ATM switches within the peer group are not connected, then the peer group cannot in some cases be used for ATM calls originating and/or terminating in other peer groups; (ii) If the PGL is not ATM attached, then the ATM switches which are border nodes cannot successfully exchange the correct information to allow the PGL in this peer group and the PGL in the neighboring peer group to contact each other.

These problems can be corrected by having all border nodes in the peer group be routers. This results in a third possible type of peer group, known as a router centric peer group.

In a router-centric peer group, all border nodes must be routers. The PGL must be I-PNNI knowledgeable. However, any I-PNNI knowledgeable node may be PGL (there is no requirement that the PGL be ATM attached). There does not need to be any ATM switches whatsoever in a router-centric peer group. If there are any switches, then the switches do not need to be contiguous. The switches may be thought of as existing primarily for the purpose of interconnecting routers. However, the switches may also be used for ATM hosts which are attached to the ATM cloud. Direct ATM connectivity will be possible only

within the peer group, and only between nodes which are interconnected via a contiguous set of ATM switches.

If a router centric peer group contains internal logical group nodes (i.e., logical group nodes which do not have any outside links), then these may represent any combination of lower level pure ATM peer groups, mixed peer groups, or router centric peer groups. If a router centric peer group contains any border logical group nodes (i.e., logical group nodes which have outside links) then these must represent lower level router centric peer groups.

4.1.2 Discussion of Peer Group Types

These rules may be summarized and justified as follows:

- If there is IP operating inside a peer group, and if the availability of IP needs to be visible from outside of the peer group, then the PGL must understand I-PNNI (i.e., the peer group cannot be a pure ATM peer group).

- If there is a pure ATM switch on the boundary of a peer group, then it is possible that the PGL in a neighboring peer group will be attempting to contact the PGL of this peer group using normal PNNI mechanisms. In this case the PGL must be ATM-attached so that the normal PNNI mechanisms will work in this case for establishing connectivity between PGLs.

- If you want ATM connectivity to work within a peer group, including the case where the peer group is a transit peer group for a call with originates and terminates outside of the peer group, then the set of ATM switches within the peer group must be contiguous.

4.2 Summarizing A Peer Group Into the Higher Level

I-PNNI uses the same hierarchy mechanisms as PNNI. Thus for example each peer group elects a single real system to operate as peer group leader. The peer group leader specifies the identity of the higher level peer group. Also, the peer group is summarized at the next higher level as a single node known as the logical group node (LGN). The LGN represents the capabilities of the overall peer group. For example, if the LGN represents a mixed peer group, then the PTSEs transmitted by the LGN will advertise reachability to both IP and ATM addresses.

Section 3.2 discussed the requirement to avoid routing SVCs via routers in flat (non-hierarchical) networks. The same requirement occurs in hierarchical I-PNNI networks. In this case, however, note that the peer group represents multiple real physical systems, possibly including both ATM switches and routers.

For pure ATM peer groups, note that all border nodes are ATM switches, and the ATM switches within the peer group need to be contiguous. This implies that a pure ATM peer group is summarized in the normal PNNI manner, and no restrictions are required.

For a mixed peer group, things are a bit more complex. In this case there may be some border nodes which are routers, and some which are ATM switches. The logical group node representing the peer group therefore cannot be marked as transit-restricted, since this would prevent transit traffic from entering via one border ATM switch and exiting via a different border ATM switch. Instead it is necessary that the transit-restricted indication is placed on higher level horizontal links corresponding to outside links from routers. Fortunately, the mechanisms for tagging outside links, corresponding uplinks, and corresponding higher level horizontal links allow the attribute tags to be carried up to the higher level links.

It is also desirable to avoid aggregating transit-restricted links from routers with normal (non-restricted) links from switches. We propose that this

may be accomplished by having links from routers use a different default aggregation token.

For router centric peer groups, we have a choice: The obvious option would be to have the corresponding logical group node be marked as transit restricted. An alternate approach would be to have the links from the logical group node all be marked as transit-restricted. This alternate approach would make the treatment of a logical group node and its links be identical for mixed and for router-centric peer groups.

4.3 SVCs between Peer Groups

4.3.1 Use the "Real" ATM Called Address

Consider the situation illustrated in figure 4. Here there are two adjacent peer groups A and B. The nodes on the boundary between the peer groups are all ATM switches. However, peer group A includes multiple routers, including one or more routers which are not ATM attached.

Consider a hypothetical SVC from a node within peer group B (perhaps node G) destined to node R1 in peer group A. For example, suppose that node G is elected peer group leader of peer group B, and node R1 is elected peer group leader of peer group A (admittedly this would violate a restrictions described above -- it will soon be clear why this restriction exists). Also, suppose that node G happens to have a higher node ID than node R1, implying that G should initiate the SVC between peer group leaders G and R1. G initiates the associated call request, with a higher level DTL that specifies {B,A}, and a lower level DTL that specifies {G,E}. Thus node G forwards the call request via node E. Node E then forwards the call request to node C. At this point node C needs to forward the call request to node R1.

Here a problem occurs: Node C discovers that it does not have a route to node R1. The problem that both nodes R2 and R3 are announcing that they are "transit-restricted", meaning of course that they are nontransit for the purpose of forwarding normal ATM SVCs.

The solution is to note that the described situation includes an error: Call requests (used to set up SVCs) need to specify a called ATM address which correctly identifies the ATM-attached point where the SVC ends. In the above example, the PGL of peer group B was trying to set up an SVC to node R1, where in fact R1 is not a correct end point for any SVC.

In normal cases, this situation will not occur. The correct operation of I-PNNI and of NHRP will result in a situation where SVCs will be destined to an ATM address which is in fact ATM reachable. There are however two special situations which have to be dealt with individually: SVCs between peer group leaders, and SVCs to NHRP servers. There is also an issue to be clarified regarding outside links between routers and ATM switches. These are discussed in the following sections.

4.3.2 Outside Links between Routers and Switches

Consider the network illustrated in figure 5. Suppose that router R5 needs to open an SVC to R3. In this case the ATM address of R3 would be advertised as reachable via R3. This would imply that at the higher level it would appear to be reachable via logical group node X (the logical node corresponding to peer group X). Suppose therefore that the SVC is routed via node C to node A. In this case A has no way to forward the SVC to R3. Although R3 has an ATM interface, it is not reachable from other ATM equipment within R3's peer group.

Earlier, we made a restriction that the ATM switches within a peer group needs to be contiguous. It is clear that this needs to be extended: If we define the "ATM Subnet" within a peer group to be the set of ATM switches plus ATM UNI and NNI interfaces, then the ATM Subnet within a peer group must be contiguous, and interconnected via ATM switches.

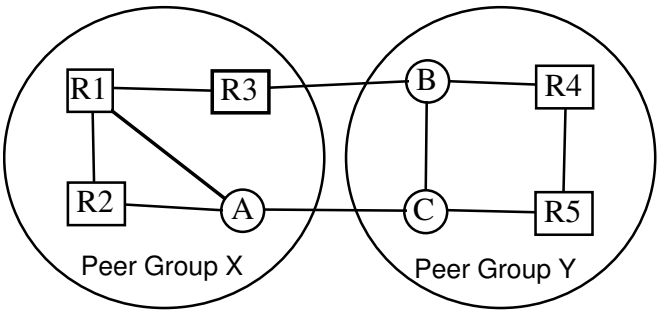

Figure 5: Outside link between a router and a switch

To the casual observer this may appear to be corrected in figure 6, by the addition of a link from R3 to A. In this case, suppose that R3 is using a single ATM address for both ATM interfaces. In this case when the SVC is delivered to A, it is possible for it to be forwarded to R3.

However, there is still a problem in this case: Specifically, the interface between R3 and B won't be used. Since the ATM address of R3 is advertised as reachable via logical group node X, any call requests to R3 will enter peer group X before attempting to reach R3. Note that this still breaks our new restriction, since one of the UNI interfaces of R3 is not interconnected with the rest of the ATM subnet in peer group X.

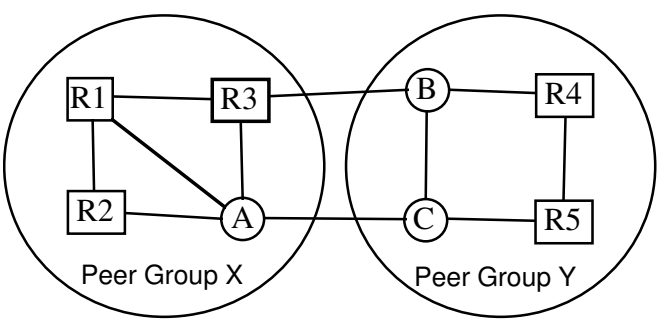

Figure 6: Slight Variation on Outside Links

A solution is to not allow outside links between a router and an ATM switch. If most cases peer groups may be configured so that there are never ATM switches with routers as outside neighbors. However, in some cases the normal manner in which to configure peer group boundaries will put a router in one peer group with a neighboring switch in another.

Where such a link is required, then one of the systems needs to operate as a "split system". A split system is defined to be a single real physical system that is represented by two (or more) logical nodes in the PNNI / I-PNNI protocol operation (as described in Annex C of [1]).

This is illustrated in figure 7. Here system R3 is operating as two logical nodes, one in peer group X and one in peer group Y. This allows the logical outside link between the two peer groups to be between two routers. The ATM address assigned to the interface from R3 to X is advertised as reachable via node r3' in peer group Y. The other characteristics of router R3 (such as IP reachability) are advertised in the PTSEs transmitted by node R3 in peer group X. The outside link between peer groups X and Y becomes a link between routers in this case.

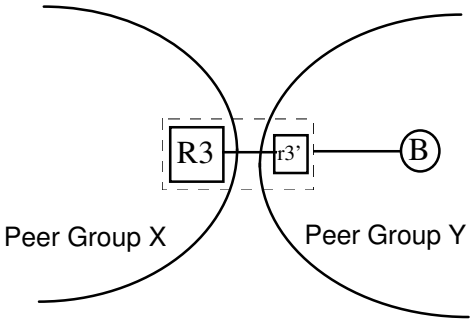

Figure 7: The correction for the outside link problem

At first glance it might appear that either the router or the switch could operate as a split system in this case. However, the router needs to be the node which operates as a split system for two reasons: (i) We do not want to require that switches understand anything beyond the basic PNNI specification; (ii) If the ATM switch operates as a split system (to appear in the peer group X in our example), then the ATM equipment within peer group X would not be connected.

Routers which are running Integrated PNNI will place an I-PNNI specific indication in their Hellos to allow this case to be automatically detected. This eliminates the need for any manual configuration to specify when a router needs to operate as a split system for this purpose. Also, routers which are running I-PNNI may send an initial Hello which leaves their own peer group blank, so that they can switch the one interface to being in the same peer group as their neighbor in the case that the neighbor is an ATM switch.

Having the router be split in this manner implies that peer group X ends up having a router in it, which implies that the PGL needs to understand I-PNNI. At first glance this might seem unfortunate, particularly if other peer group (peer group Y in our example) has no other routers in it. However, note that this requirement actually makes sense: In the example peer group Y may be the peer group in which a router to router SVC leaves the ATM cloud. Thus some system in peer group Y must be capable of answering NHRP queries (there is no system outside of Y which would be capable of knowing where the call is leaving the ATM network). Having peer group Y become a mixed peer group implies that the PGL will be I-PNNI knowledgeable, and will be capable of answering NHRP queries.

Again, we believe that there will be some systems which are both routers and switches. We believe that this can be handled in a straightforward manner, but some care will be needed to precisely and unambiguously specify precisely how a router/switch will operate if the outside neighbor is a switch versus if the outside neighbor is a router.

4.3.3 SVCs between PGLs

The previous section points out that the outside neighbors (i.e., ATM border nodes in adjacent peer groups with a link between them) will either both be an ATM switch or both be a router. We therefore have two situations to consider: PGL interaction via outside neighbor switches, and PGL interaction via outside neighbor routers.

4.3.3.1 PGL Interaction via Outside Neighbor Switches

Note that this case will occur only when both PGLs are ATM-attached. This implies that the normal mechanisms described in the PNNI specification can be used in this case.

4.3.3.2 PGL Interaction via Outside Neighbor Routers

Note that this case will occur only when the PGLs are I-PNNI knowledgeable in both peer groups. This implies that the PGLs are able to understand I-PNNI specific mechanisms in order to contact each other. We expect to define mechanisms which are very similar to those defined for PNNI. However, in the case that the interaction is via routers, then clearly the PNNI packets which are exchanged between PGLs in neighboring peer group cannot be done via a direct SVC between PGLs. Rather, the border router may need to forward PNNI packets between PGLs. Most likely this would be done by encapsulating the PNNI packets in an IP packet destined for the other PGL, and forwarding the packet via the border node.

There may of course be cases where an I-PNNI knowledgeable PGL in a mixed peer group determines that there is more than one way to contact the PGL in a neighboring peer group, with at least one way being via an SVC via a border ATM switch, and at least one way being via IP encapsulation via a border router. In this case the PGL may choose which border node to use in the same manner that it would choose between multiple border ATM switches in the normal PNNI case (i.e., a local decision may be made, but if communication via one border node does not work then communication via other border nodes should be attempted).

4.3.4 SVCs to NHRP servers

For mixed peer groups, the summarized information announced in the logical group node's PTSEs cannot in general contain enough information to identify the ATM exit point ATM address required to reach every IP address that is reachable in the peer group. Rather, nodes outside the peer group, when attempting to set up optimal SVCs to IP addresses in the peer group, will first send an NHRP query to an NHRP server in the peer group.

In fact, NHRP may be used in several cases in PNNI:

- To answer host queries from hosts who consider the router to be local.
- To answer queries to locate hosts that are located on a virtual network which is served by the router (this is explicitly possible when query reachability is advertised by an ATM-attached router).
- In a hierarchy, to answer queries from nodes outside of the peer group.

We propose that all I-PNNI knowledgeable systems must be capable of being NHRP servers. For example, this allows hosts to in all cases query their local router for NHRP information.

For mixed peer groups, the PGL will announce the ATM address used to reach the NHRP server in the peer group. There are a couple of possible options in this case:

(i) One option would be for this to be simply the PGL's address. This would require that the PGL operate as NHRP server for queries from outside the peer group about destinations inside the peer group.

(ii) The PGL could select a different address (which could correspond to the PGLs address, but with a different unique selector) which is used to reach the NHRP server. Within the peer group, every ATM-attached system which is I-PNNI knowledgeable announces reachability to this special address. This implies that an incoming SVC intenced to carry an NHRP request might be routed to any ATM-attached I-PNNI-knowlegeable system in the peer group. This might allow the NHRP load to be spread amongst multiple systems in the peer group.

The first option is simpler, and will make more efficient use of SVCs for NHRP requests. In particular, PGL to PGL SVCs will already exist, and can be used. This will make even more efficient use of SVCs if NHRP requests for destinations outside of the peer group are funneled via the PGL. This tradeoff places slightly more NHRP query load on the PGL, but makes significantly more

efficient use of SVCs for queries. This will also optimize the use of NHRP caches in PGLs. We therefore recommend the first approach.

4.5 Logical Group Node Advertisements

To summarize, the PTSEs advertised by a logical group node contains the following information (this is not necessarily a complete list):

- The ATM addresses reachable in the corresponding peer group. Generally this makes use of one or more summary addresses.

- The IP addresses reachable in the corresponding peer group. Again, this makes use of one or more summary addresses. IP addresses are completely independent of ATM addresses (except in one case, where

- The address to be used to reach NHRP servers in the logical group node.

- Horizontal links to peer nodes. Generally, router to router links are advertised as transit restricted. Router to router (transit restricted) links may be aggregated with other router to router links. Similarly, switch to switch links may be aggregated with other switch to switch links. However, router to router transit restricted links should not be aggregated with switch to switch links.

- Other standard PNNI information may also be advertised: For example, this includes uplinks to other higher level nodes.

4.6 Designated Routers and Pseudonodes in a Hierarchy

We will need to determine that manner that DRs and PNs can be applied in the case that the set of routers on a broadcast LAN belong to multiple peer groups, possibly at different levels of the hierarchy.

Our initial proposal is that the DR election algorithm would be performed independently for each peer group. Thus one DR would be elected for to each peer group which has one or more routers on the LAN. The DRs, acting on behalf of the Pseudonodes, would exchange Hellos enhanced with the information that PNNI requires on outside links. This would allow the pseudonodes to declare uplinks, again in the normal PNNI manner. This would eliminate the need for each router on the broadcast LAN to advertise uplinks.

4.7 Spanning Set of SVCs in a Hierarchy

In section 3.7 we explained how an a priori spanning set of SVCs may be set up between routers within a peer group. This allows short duration best effort IP packets to be forwarded without the requirement to set up SVCs on demand. A similar issue exists in a hierarchy. The best method of setting up an a priori spanning set of SVCs in a hierarchy is for further study.

5. REFERENCES

[1] PNNI Subworking group, "PNNI Draft Specification", edited by R. Cherukuri and D. Dykeman, ATM Forum 94-0471R15, January 1996.

[2] R.Callon et al, "The Relationship between MPOA and Integrated PNNI", ATM Forum 96-0352, April 1996.

[3] R.Callon et al, "Methods for Routing of Internetwork Level Protocols over ATM", ATM Forum 96-0353, April 1996.

[4] R.Callon et al, "An Overview of PNNI Augmented Routing", ATM Forum 96-0354, April 1996.

[5] J.Postel, "Internet Protocol", RFC 791, September 1981.

[6] D.Katz, D.Piscitello, B.Cole, J.Luciani, "NBMA Next Hop Resolution Protocol (NHRP)", Internet Draft, December 1995.

[7] MPOA Subworking group, "Requirements for the MPOA protocol", edited by C.Brown, ATM Forum 95-0004R2, April 19, 1995.

[8] MPOA Subworking group, "Baseline Text for MPOA", edited by C.Brown, ATM Forum 95-0824R5, January 1996.

[9] LAN emulation subworking group, "LAN Emulation over ATM, Version 1.0", edited by B. Ellington, AF-LANE-0021.000, January 1995.

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

[11] R.Callon, "Use of ISO IS-IS for Routing in TCP/IP and Dual Environments", RFC 1195, December 1990.

[12] R.Callon, "Integrated PNNI for Multi-Protocol Routing", ATM Forum 94-0789, September 1994.