INTERNET-DRAFT
<draft-schulter-ipv6atm-framework-01.txt>

P. Schulter
Digital Equipment Corp.
February 12, 1996

A Framework for IPv6 Over ATM

Status of this Memo

This document is an Internet-Draft. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its ˝areas and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use ˝Internet-Drafts as reference material or to site them other than as "work in progress."

To learn the current status of any Internet-Draft, please check the "lid-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), nic.nordu.net (Europe), munnari.os.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

This Internet Draft Expires July 15, 1996.

Abstract

Work in specifying the IPv6 [IPv6-SPEC] has now progressed to a point that the base set of IPv6 specs are on a standards track ˝and work on IPv6 implementations is beginning at many organizations. So far, the IPv6 specifications have dealt with broadcast media LANs [IPv6-ETHER]with very little attention to ATM. The IP over ATM WG is now starting to look at IPv6 over ATM [IP-IPV6ND]. ˝Many of the problems encountered in running IPv4 over ATM [IP-FRAME, IP-ATM, IP-ATMU, IP-ATMMC, NHRP] must also be dealt with when running IPv6 over ATM. Since ATM networks are point-to-point networks that offer no connectionless broadcast or multicast capabilities, many of the IPv6 protocols can't be directly ˝applied to the ATM datalink. Instead, some software framework built on top of the ATM datalink is needed to handle those protocols or functions which rely on connectionless multicast capabilities. Among these functions are Neighbor Discovery, Router Discovery, and address configuration. Another difficulty with ATM is that ATM networks have a flat address space, and are expected to ˝become very large (even global). It is desirable to logically partition a large ATM network into IPv6 subnets and to be able to maintain IPv6 subnet semantics while still maintaining the ability to interconnect any two nodes on the ATM network so that ATM QoS services can be provided within the framework of IPv6 ˝networking.

This document describes a framework for adapting IPv6 and its ˝base protocols [IPv6-SPEC, IPV6-ND, IPV6-ADDRCONF, IPV6-ADDRARCH, ˝IPV6-DHCP] to ATM networks while maintaining the complete ˝functionality and semantics of these protocols. This is done by defining an IPv6 "link" for ATM (and NBMA) networks in a way which preserves all the fundamental IPv6 concepts and functionality while still taking advantage of the unique capabilities of ATM.

Status of this Memo....1
Abstract....1
1. Introduction and Background....4
1.1. Requirements....6
2. Definitions of Terms....7
2.1. IPv6 Terminology....7
2.2. ATM Terminology....8
2.3. New Terminology....9
2.4. Addresses....10
3. IPv6 Over ATM Framework Description....10
3.1. NDS Trees....19
3.2. NDS Tree Configuration....21
3.2.1. Manual NDS Tree Configuration....23
3.2.2. Autoconfiguration of Trees....24
3.3. Forming Logical Links....26
3.3.1. Neighborhoods....28
3.4. Forming Sites....29
3.5. Beyond Sites....30
4. NDS Operations....31
4.1. Router Advertisements....33
4.1.1. Exporting Prefix Information....34
4.1.2. Routerless Logical Links....34
4.1.3. Movement of Prefix Information Outside Logical Links....35
4.1.4. Importing Prefix Information....36
4.2. Router Solicitations....38
4.3. Neighbor Solicitation....38
4.4. Neighbor Advertisements....40
4.5. Anycast Addresses....41
4.6. Neighbor Unreachability Detection....43
5. Address Configuration....44
5.1. Link-Local Addresses....44
5.2. Stateless Address Configuration....45
5.3. Stateful Address Configuration....45
5.4. Manual Address Configuration....46
5.5. Node Relocation....46
5.6. Duplicate Address Detection....48
6. Multicasting....49
7. VC Characteristics....49
7.1. NDS VCs....50
7.2. Data VCs....50
8. Security....51
Acknowledgments....51
Authors' Address....51
References....51

1. Introduction and Background

The basic problem in adapting IPv6 to ATM networks is that ATM networks do not provide a connectionless multicast link, and therefore, does not inherently provide a framework for full IPv6 Neighbor Discovery [IPV6ND]. Further, ATM networks can be very large (even global in scope) which can result in networks that ˝are composed of thousands or even hundreds of thousands of nodes. This means that ATM networks have much different link layer semantics than broadcast media. Generally, all nodes on an ATM network can establish connectivity to any other node on the same network without regard to geography. Additionally, ATM even provides subaddressing so nodes on different ATM networks which are interconnected (generally two private networks connected through a public network) can directly connect. While this reachability scheme can result in much greater flexibility in interconnecting nodes within a larger ATM network (for example, nodes in geographically diverse locations can be logically ˝grouped into logical networks), it also makes the partitioning of ˝nodes on the larger ATM network into a set of smaller logical networks or links more difficult.

Additionally, IPv6 differs from IPv4 in that address resolution and address configuration are part of the base IPv6 protocol and are located in the network layer rather than the datalink layer. That is, the Neighbor Discovery protocols which IPv6 uses to perform neighbor and router discovery are an integral part of IPv6, and any mechanism that is used to adapt ATM to IPv6 must deal with the Neighbor Discovery protocols. This is in contrast to IPv4 where the address resolution protocols are not part of ˝the base IP protocols and are part of each individual datalink layer (i.e., ARP for broadcast media, ATMARP or NHRP for ATM). In IPv4 new datalink layers could define their own address resolution protocols as necessary (as was done with ATMARP) since this function is left to the datalink. Thus, new datalinks could be added without affecting the IPv4 network layer. In IPv6 all datalinks must handle IPv6 Neighbor Discovery packets and use ˝them for address resolution, router discovery and address configuration. Not using Neighbor Discovery would require modifying the IPv6 network layer to accommodate a specific datalink.

Further, IPv6 provides for some extra features not in IPv4. One of these features is address autoconfiguration. Any mechanism used to adapt IPv6 to ATM must provide for address autconfiguration since this is expected to be the primary way in which nodes will configure their network layer addresses. Next, IPv6 has also been defined with network layer security features ˝as part of the base protocol. These security features are applied ˝to address configuration and resolution since they are defined at ˝the network layer. Any IPv6 over ATM solution must also take IPv6 security into consideration and preserve the security features built in to address resolution and configuration. Finally, IPv6 includes an address architecture [IPV6-ADDRARCH] which provides for network layer addresses (specifically link-local and site-local addresses) which are not globally visible. This addressing architecture must also be maintained for IPv6 over ATM.

The IPv6 protocols currently rely on connectionless broadcast and multicast capabilities of legacy LAN technologies to perform

functions such as IP to physical address resolution. Further, these protocols rely on the physical partitioning of networks to establish the boundaries for the creation of subnets and for routing and address configuration. To adapt the base IPv6 protocols to ATM the first thing that must be done is to define what an IPv6 subnet and "link" is on an ATM network. That is, how a set of ATM nodes is partitioned into an autonomous group which share common address prefixes, and between which the Neighbor Discovery protocols can be used. Such a "link" should be defined so that all the base IPv6 protocols (including ND) can be run ˝over it with no modifications to endsystems or routers, and so that ˝the administration of the network layer software is the same as that on other media.

Once an IPv6 "link" is defined for ATM networks, mechanisms and policies for running IPv6 protocols over the link must be ˝defined. These mechanisms should meet the following goals:

- the concept of the IPv6 subnetwork/link must be maintained
- the scope of link-local and site-local addressing must be maintained
- all functionality provided by the current IPv6 Neighbor Discovery protocols must be provided
- there must be no modifications required to the IPv6 network layer in order to run IPv6 over ATM
- the Neighbor Discovery protocol semantics must me maintained
- the resulting protocols and architecture must scale to arbitrarily large networks
- nodes must be permitted to join multiple subnets/links
- nodes must be permitted to join any subnet/link on the wider ATM network without regard to the node's geographic location
- nodes on different subnets/links must be permitted (but not required) to establish connectivity directly through the ATM network without going through a router. This is needed so that connections with specific QoS requirements can be established so that IPv6 can make full use of ATM QoS capabilities
- the link must provide for redundancy and failover of critical, non-ATM resources

Optionally, any IPv6 over ATM architecture should provide for the highest degree of autoconfiguration as the ATM and IPv6 protocols will allow. Ideally, all elements of an IPv6 over ATM network should be self configuring with as little human intervention and management as possible.

Finally, any protocols developed for IPv6 over ATM should be defined for both current UNI 3.1 and future UNI 4.0 networks. Since it is expected that UNI 4.0 will be widely deployed by the time IPv6 goes into wide use, the new capabilities and features ˝of UNI 4.0 should be taken into account in places where they can improve the functioning of the protocols. However, since initial implementations will be written to existing UNI 3.0/3.1 networks, the protocols must also work under UNI 3.0/3.1.

1.1. Requirements

Throughout this document, the words that are used to define the significance of the particular requirements are capitalized. These words are:

- **MUST** — This word or the adjective **"REQUIRED"** means that the item is an absolute requirement of this specification.
- **MUST NOT** — This phrase means the item is an absolute prohibition of this specification.
- **SHOULD** — This word or the adjective **"RECOMMENDED"** means that there may exist valid reasons in particular circumstances to ignore this item, but the full implications should be understood and the case carefully weighed before choosing a different course.
- **SHOULD NOT** — This phrase means that there may exist valid reasons in particular circumstances when the listed behavior is acceptable or even useful, but the full implications should be understood and the case carefully weighed before implementing any behavior described with this label.
- **MAY** — This word or the adjective **"OPTIONAL"** means that the item is truly optional. One vendor may choose to include the item because a particular marketplace requires it or because it enhances the product, for example, an other vendor may omit the same item.

2. Definitions of Terms

Relevant terminology from the IPv6 Protocol [IPV6-SPEC], IPv6 Addressing Architecture [IPV6-ADDR], IPv6 Stateless Address Autoconfiguration [IPV6-ADDCONF], IPv6 Neighbor Discovery [IPV6-ND], then Classical IP and ARP over ATM [IP-ATM], UNI 3.0 [ATM-UNI30], UNI 3.1 [ATM-UNI31], UNI 4.0 [ATM-UNI40], and PNNI [ATM-PNNI] will be provided, and finally new terminology introduced by this document will be given.

2.1. IPv6 Terminology

- **IPv6** — Internet Protocol Version 6.
- **IPv4** — Internet Protocol Version 4.
- **Node** — a device that implements IPv6. In this document all nodes implement IPv6 over ATM. Node are always connected to ˝an ATM network and may be connected to any number of other networks.
- **Router** — A node that forwards IP packets not explicitly addressed to itself.
- **Host** — Any node that is not a router
- **link** — A communications facility or medium over which nodes can communicate at the link layer, i.e., the layer immediately below IPv6. Examples are Ethernet (simple or bridged), ATM networks, PPP links, X.25, Frame Relay, and FDDI; and internet (or higher) layer "tunnels", such as tunnels over IPv6 or IPv6 itself.
- **Neighbors** — Nodes attached to the same link.
- **Interface** — A node's attachment to a link

- **address** —An IP layer identifier for an interface or a set of interfaces
- **packet** — An IP header plus a payload.
- **Communication** — Any packet exchange between nodes that requires that the address of each node used in the exchange remain the same for the duration of the exchange. Examples are a TCP connection or UDP request/response.
- **Unicast-address** — An identifier for a single interface. A packet sent to a unicast address is delivered to the interface identified by that address.
- **Multicast-address** — an identifier for a set of interfaces (typically belonging to different nodes). A packet sent to a multicast address is delivered to all interfaces identified by that address.
- **Link-layer address** — a link-layer identifier for an interface. Examples are an IEEE 802 address for Ethernet links, and ATM NSAP or E.164 addresses (with optional subaddresses) for ATM networks.
- **link-local address** — An address having link-only scope that can be used to reach neighboring nodes attached to the same link. All interfaces have a link-local address.
- **Site-local address** — An address having scope that is limited to the local site
- **global address** — an address with unlimited scope
- **interface token** — A link dependent identifier for an interface this is (at least) unique per link.
- **Anycast address** — an identifier for a set of interfaces (typically belonging to different nodes). A packet sent to ˝an anycast address is delivered to one of the interfaces identified by that address (the "nearest" one, according to the routing protocols measurement of distance).
- **On-link** — An address that is assigned to an interface on a specified link. A node considers an address to be on-link if:
  - it is covered by one of the link's prefixes
  - a neighboring router specifies the address as the target of a Redirect message
  - a Neighbor Advertisement message is received for the (target) address
  - a Neighbor Discovery message is received from the address
- **off-link** — the opposite of "on-link"; and address that is not assigned to any interfaces on the specified link.

2.2. ATM Terminology

- **Point-to-Point (PtP)** — An ATM connection (virtual circuit) connecting exactly two ATM nodes. The same set of nodes can have multiple PtP connections to each other.
- **Point-to-Multipoint (PtM)** — An ATM connection that connects one node to many nodes. PtM connections are unidirectional with data flowing from the "root" node to many "leaf" nodes. PtM connections are created by placing a special call to each node to be included in the PtM connection. The same set of nodes ˝can have multiple PtM connections to each other.
- **Call** — The process of establishing a connection (virtual circuit) between two nodes (parties). One node actively initiates connection establishment (calling party) and the other node passively receives the call (called party). ˝Circuit

traffic parameters are established and negotiated during the call.

- **Release** — The process of terminating a single connection between two more nodes. A node is always notified if a connection is released for any reason.
- **Group Address** — This type of address is specific to UNI 4.0 (UNI 3.0/3.1 does not support group addressing). An address which identifies a set of ATM nodes. Calls to a group address are routed by the ATM network to exactly one of the nodes identified by the address. Group addresses are assigned a specific scope when they are registered with the network. This scope defines which nodes within the ATM network routing hierarchy are able to reach the specific nodes identified by the group address. Group addresses are also referred to as anycast addresses in the UNI 4.0 specification, but the term group address is used in this document to avoid confusion with IPv6 anycast addresses.
- **Address registration** — The process by which an ATM node informs the ATM network of its link-layer addresses. ATM nodes can register multiple addresses and the addresses can take several forms. A node must register an address to be able to ″establish connections to other nodes.
- **UNI** — User-Network Interface. The ATM interface and protocols between the nodes and the network.
- **PNNI** — Private Network-Network Interface. The ATM interface and protocols between switching elements of an ATM network. The also defines how an ATM network topology is build and how calls are routed within that topology.

2.3. New Terminology

- **Logical Link (LL)** — a set of ATM nodes which can reach each other using only link-local addresses and which can broadcast to all other nodes in the LL. This is analogous to a LAN segment.
- **Neighbor Discovery Server (NDS)** — An entity which provides a service for performing IPv6 Neighbor Discovery in an ATM network.
- **NDS(n)** — This notation denotes a Neighbor Discovery server at a specific level in the NDS hierarchy.
- **GA(n)** - This notation denotes a specific ATM group address which is used to place calls to an NDS at level n.
- **"Broadcast"** — In this document this term refers to physically transmitting an IPv6 packet to all nodes or other entities within a specific scope. This term does not refer to any function of the ATM network, and is used in quotes so as not ″to be confused with a hardware broadcast function.
- **Site** — A collection of Logical Links in which all nodes share a common IPv6 site address prefix. All nodes in a Site can ″reach any other node in the Site using IPv6 site addresses. Since this term is used to specify a specific logical grouping of ″LLs and nodes, and does not directly relate to the ATM UNI 4.0 ″site group addressing scope, it is capitalized to distinguish it from the ATM term "site".

2.4. Addresses

- **all-nodes multicast address** — The link-local scope address to reach all nodes. FF02::1
- **All-routers multicast address** — the link-local scope address to reach all routers. FF02::2
- **Solicited-node multicast address** — a link-local scope address that is computed as a function of the solicited target's address. See [IPV6-ND]
- **link-local address** — a unicast address having link-only scope that can be used to reach neighbors. All interfaces **MUST** have a link-local address. Routers **MUST NOT** forward packets with link-local source addresses. See [IPV6-ADDR]
- **unspecified address** — A reserved address value that indicates the lack of an address (e.g., the address is unknown). It is never used as a destination address. See [IPV6-ND] and [IPV6-ADDR].
- **DHCP Server/Relay Agent address** — the link local scope address used to reach all IPv6 DHCP Servers and Relay Agents. FF02::1:0 [IPV6-DHCP].
- **tentative address** — an address that has been assigned to an interface, but which has not yet been verified using duplicate address discovery. A node can not use a tentative address as the source address in a packet, or receive packets which contain the tentative address as the destination, until duplicate address detection has been performed [see IPV6-ADDCONF].

3. IPv6 Over ATM Framework Description

The IPv6 over ATM framework described here defines what an IPv6 "link" is on an ATM network, and a mechanism by which IPv6 protocols can be applied to an ATM IPv6 "link". In IPv6 terms a link is a collection of nodes which share common prefixes, can communicate using link-local addresses, and which can use ND to resolve link-layer addresses, perform address configuration, and router discovery. Communication to off-link nodes is ˝accomplished via routers. The definition of an IPv6 ATM link described here uses the same model of a link with one important addition which ˝is necessary for the media. This is, all nodes on the same link are expected to create one or more VCs over which data is exchanged (that is, all nodes on a link must directly connect). On an ATM network, a link has the same semantics as it does on broadcast media. Also, just as with broadcast media, the link also defines the scope of ND messages.

To form an IPv6 link on an ATM network, nodes are connected to a set of servers which are used to aid the IPv6 Neighbor Discovery protocols. These servers, called Neighbor Discovery Servers ˝(NDS) are organized in a tree structure rather than a flat structure. Organizing the framework into a tree structure is done for scalability, redundancy, and failover. A tree structure also makes network autoconfiguration possible in a UNI 4.0 ˝environment where group addresses and reachability scopes can be used.

An NDS server is a service that can be located on any ATM node independent of other IPv6 services. They can function independently of IPv6 ATM interfaces and are not addressable IPv6 entities. The NDS servers are primarily ND packet relay services and there are no node to NDS protocols other than for ˝registration of nodes when they join a Logical Link (the registration process is primarily for security and configuration purposes). Once a

node joins a Logical Link no new protocols are required to ˝perform any ND function.

All ATM nodes on an ATM network are all technically "on-link" since they can all connect to any other node and exchange data. An ATM network can be logically partitioned (potentially across arbitrary geographic boundaries if needed) into many IPv6 subnetworks. This is necessary to logically group nodes sharing some organizational relationship together into different groups which have some degree of administrative isolation from other groups. However, even after logically partitioning ATM networks into logical IPv6 subnetworks, all ATM nodes are still physically "on-link" to all other ATM nodes. The ability for any two nodes to connect has not been affected at the link-layer, only network layer reachability has been affected. Note that this case is different from the normal case with legacy LANs. Using legacy LANs, IPv6 subnetworks are generally physically as well as logically disjoint. With ATM, subnetworks are logically ˝disjoint, but not physically disjoint.

Given this type of partitioning of an ATM network, it is possible to define two levels of "on-link". The first is the physical level where two nodes are considered "on-link" they can establish a VC to each other. This level of "on-link" is determined by the physical connections of the ATM network(s). The second level is the logical level where nodes within the same logical grouping ˝are considered "on-link" at the network layer. Nodes which are not logically "on-link" are still physically "on-link". Hosts which are physically "off-link" must be logically "off-link" also.

The framework described here manages these two levels of "on-link" in such a way as to extend the definition of "on-link" to include multiple IPv6 subnets while still maintaining the subnetwork ˝model (the logical "on-link" level) and the ability for any two nodes to directly connect (the physical "on-link" level). This allows existing IPv6 semantics to be maintained while still allowing "VC cut through" between nodes on different subnetworks. Extending the model of the "link" is possible because of the characteristics of the ATM datalink, and because the IPv6 "link" is simply an abstraction built on top of the underlying datalink. However, this extension does not change the way IPv6 treats "on-link" and "off-link" nodes.

To do this, the concept of a Logical Link (LL) is used. A ˝logical link has semantics almost exactly like those of a legacy LAN physical link. That is, a logical link is a set of nodes which can directly communicate with each other and over which ND is ˝used for address resolution, address configuration, and router discovery. Most importantly, a Logical Link provides the isolation and logical grouping of a set of nodes into a "community" which can be managed like a LAN. Thus, Logical Links control the logical determination of "on-link" at the network level, not at the physical level.

The Logical Link concept differs somewhat from the IP subnetwork concept. A Logical Link defines node groupings, but does not necessarily need to equate with IP subnetting. Since the operation of a Logical Link mainly provides the semantics of a physical LAN, it is possible that many IP subnets could be configured on the same Logical Link. This, a LL could support multiple Logical IP Subnets (LISs).

The Logical Link is the basic unit of this framework. At a minimum, some set of nodes must be grouped together into a ˝Logical

Link. These logical links provide an environment which allows ˝all IPv6 and ND functions to be used exactly as they would be used over a broadcast media LAN. Logical links could then be interconnected via routers, thus providing exactly the same on-link/off-link semantics as broadcast LANs.

Multiple LLs can exist on the same ATM network, and will function as disjoint logical networks, requiring routers for nodes on different LLs to communicate. Using ATM-to-ATM routers is expensive in terms of network utilization (data must be repeated over the ATM cloud for each hop) and doesn't provide the level of QoS guarantees provided by direct VCs. Being able to directly connect nodes is different Logical Links is desirable for ATM network utilization, and necessary to provide full QoS based services between any two nodes.

If two nodes which are not on-link need to communicate they must do so via a router (at least initially). Again, this follows the IPv6 semantics. If a node needs to communicate with a node which is not on-link it will attempt to send the packets to the ˝off-link node via one of the on-link default routers (note that in IPv6 nodes communicate with routers using only link-local addresses, ˝so the router **MUST** be on the node's Logical Link). The router can then forward the packet to the next hop, or can redirect the sending host to another router. In the case of ATM, the router could also redirect the sender directly to the destination node ˝if the destination is physically on-link. If the routers ran a protocol such as NHRP the local router could determine the link-layer address of the destination, and then send an ND Redirect message to the sending node.

While communications through routers or via router redirects completely follows the IPv6 link semantics, it may not be the ˝most optimal use of ATM. This is especially true since nodes which ˝are logically off-link may still be physically on-link. In cases where remote nodes are still physically on-link and it may be desirable to let these nodes directly connect by default. In these cases it would be more efficient if the IPv6 notion of on-link could be made to more closely match the ATM notion of ˝on-link without creating extremely large and unwieldy LLs. The framework described here allows multiple Logical Links to be grouped together so that nodes on each LL are allowed to directly connect to nodes on other LLs in the group as long as an IPv6 address of appropriate scope is used. This grouping of LLs utilizes ND protocols and the IPv6 determination of "on-link" to permit nodes on separate LLs to directly connect. Doing this extends the concept of "link" somewhat for ATM while still retaining the semantics of links and the ND protocols. However, IPv6 will continue to use the same "on-link" determination process as it does for other LANs, "on-link" nodes will always be permitted to directly connect, and communications with off-link nodes will be handled via routers (or direct connections resulting from redirects from routers).

To form these larger groupings of nodes, Logical Links may be grouped together to form larger node groupings, call Sites. A Site can consist of any number of LLs. Each LL in a Site functions just like an isolated LL, with the exception that the Site grouping can manage logical "on-link" determination between LLs. When LLs are grouped into Sites, nodes on each LL are allowed to directly connect to nodes on any other LL via the ATM network. However, logical reachability is managed by IPv6 ˝address scoping. That is, nodes on one LL can only reach nodes on ˝another

LL within the same site only by using a site-local or global address.

Sites can be further grouped into larger units as needed to provide managed connectivity between any set of nodes. As with Sites, logical reachability between nodes in different Sites is managed by IPv6 address scoping. That is, nodes within one site can only reach nodes on other sites by using global addresses.

The specification of the NDS server hierarchy proceeds from the IPv6 address scoping rules. IPv6 addresses form a simple hierarchy due to their scoping rules. The IPv6 hierarchy has three levels:

- **link-local** — the lowest level which is limited to the local network link or Logical Link.
- **Site-local** — the "middle" level which spans multiple links, but is not globally visible.
- **Global** — the highest level which has unlimited scope.

By maintaining these concepts in this framework, IPv6 concepts ˝and addressing structure can be mapped onto ATM

In its simplest form, the NDS tree has one NDS server for each level of IPv6 addressing hierarchy. That is, there is a server ˝at the link-local level, a server at the Site level, and a server at the Global level. Servers at lower levels all connect to the server at the next higher level. All nodes are connected at the link-local level. The NDS servers control the flow of Neighbor Discovery messages between levels based on the addressing scope ˝of the message's target address. This then defines three NDS servers, each belonging to a specific logical level (denoted by the notation NDS(*level*) as follows) :

- **NDS(LL)** — The NDS server which serves the Logical Link and handles all ND messages with link-local scope.
- **NDS(Site)** — The NDS server which serves the Site level and handles all messages with site-local scope.
- **NDS(Root)** — The NDS server which serves the global address space and handles all ND messages with global scope.

Additionally, a fourth server is defined, which will be explained in the following sections. This is

- **NDS(N)** — The NDS neighborhood server which serves a subset of a logical link. Only NDS(N) servers can accept connections from individual nodes.

Each NDS server is a logical services on the ATM network, not necessarily a single physical entity. That is, one physical node (a host or router) could contain services for various levels of the hierarchy, each logically independent. Also, the NDS(Root) server need not be a separate logical service, but is only the logical top of the hierarchy. Any NDS server **MAY** also be configured as the NDS(Root) server if the hierarchy does not extend above the root. For example, for an isolated LL, the NDS(LL) server is also logically the NDS(Root) server.

While IPv6 specifies three addressing scopes, having only three layers in the NDS hierarchy may be insufficient on large ATM networks. On large networks, the number of nodes that may need ˝to

be connected to a specific NDS may be larger than the system running the NDS can handle. Also, it may be desirable to distribute the NDS tasks to more servers to better distribute the load on the NDS servers. To accommodate this, and to allow the tree hierarchy to scale, the number of NDS levels in the tree ˝does not need to correspond to the number of levels in the IPv6 addressing hierarchy. Instead, the NDS tree is specified in ˝terms of both the number of logical levels (which is always three or fewer), and the number of physical levels (which can be any ˝number as necessary to achieve the scalability and connectivity requirements of a specific ATM network). The logical levels always correspond exactly to the IPv6 addressing scopes. Thus, within the link-local logical level (the Logical Link), there could be any number of physical NDS server levels (the height of the subtree), each composed of any number of NDS servers as ˝needed (the width of a level of the subtree).

Because each level of NDS server outside the LL filters out ND packets based on target address (only passing up packets with higher addressing scope) the amount of traffic flowing between servers should be lower at the higher levels of the tree. This should enable the NDS tree to scale to almost any size, even global.

Note that all nodes in an ATM network need not be connected via the same tree. Multiple NDS trees can exist in any ATM network ˝so that the nodes in each tree are logically isolated from each other. Nodes within a single NDS tree are always allowed to directly connect (directly establish VCs for exchange of data). This is because these nodes are always considered "on-link". On-link nodes always resolved their addresses via ND, perform ˝address configuration via ND, and perform router discovery via ND. Nodes which are not on-link can communicate via a router or can ˝directly connect.

Finally, the NDS tree is meant to handle only Neighbor Discovery protocols and address configuration protocols. It is not meant, and **MUST NOT** be used to carry general multicast and broadcast data. The NDS tree handles only multicast data to a specific set of multicast addresses. A separate multicast mechanism must be used to carry generic IPv6 multicast data. Multicast is ˝described later on the Multicast section. Packets to the following multicast addresses **MUST** be sent via the NDS tree:

- all-nodes multicast address
- all-routers multicast address
- solicited node multicast address
- DHCP server multicast address

These multicast address differ from more generic multicast functions in that they are used as part of some discovery mechanism which usually involves a unicast response to the multicast. That is, the intended receivers of the data must join the multicast group, but the sender of the data does not join ˝the group since the sender expects no reply or a unicast reply. ˝Other messages that follow this model could also be carried by the ˝tree. For example, router protocols such as RIP, or service location protocol messages [IP-SRVLOC]. While this framework does not explicitly describe the transmission of such packets through the NDS hierarchy, it does not prohibit the use of the hierarchy by protocols other than ND protocols. Using the NDS tree to carry any other multicast protocol **MUST** be done only with IETF agreement for specific purposes. Nodes **MUST NOT** use the NDS tree to carry

multicast traffic not explicitly specified for the NDS tree by some IETF standard.

The framework described here requires no modification to the existing Neighbor Discovery as described in IPV6-ND. They will continue to operate as they do on legacy LANs requiring no modifications of the network layer to accommodate ATM. However, some additional protocols are required for node and NDS ˝server-to-server registration, but these are isolated to the layer between IPv6 and ATM. Also, this framework does not depend on any ˝central repository of information for the network to operate. While some network state and reachability information is maintained by the servers, this state relates to subnet reachability rather than node specific information, and is used only to optimize the relaying of ND packets on the NDS tree and is not required for proper operation of the tree. All node specific information remains on the end-nodes and is obtained through the existing Neighbor Discovery mechanisms rather than from some server's cache. This makes the NDS easier to implement, more efficient, makes failover easier, and maintains the full semantics of Neighbor Discovery as expected by IPv6. For example, this allows host aliasing and node load-balancing between interfaces by allowing each node to choose what response to return to each solicitation.

The basic idea of the NDS tree is to provide a mechanism for moving Neighbor Discovery packets (and perhaps other types of packets) between nodes. Only packets which are sent to specific multicast addresses are moved by the NDS tree. Packets sent to unicast addresses **MUST** be sent over a VC connecting the source and destination nodes (or via a router). The movement of specific packet types are described in greater detail later, but is described briefly below:

- **Router Advertisements** — The usual case described in IPV6-ND is that RA messages are sent to the all-nodes multicast address. Routers on LLs will continue to do this. The RA packets will be relayed up the tree to the NDS(LL), which will then "broadcast" them back down to all nodes on the LL. Thus all nodes will receive the RA messages. Unicast RA messages are sent via a VC between the router and the destination node. RA messages are also "leaked" out of the LL to advertise prefix information to other LLs.
- **Router Solicitations** — RS messages are sent by soliciting nodes to the all-routers multicast address. As with RA messages these are relayed up to the NDS(LL), which then "broadcasts" them down to all nodes on the LL. Each node on the LL must filter incoming packets so that only routers process the RS messages. The usual case will be for a router to respond with a multicast to the all-nodes multicast address as described above. If the router unicasts the response it must first create a VC to the soliciting node. RS messages are never relayed beyond the LL of the soliciting node.
- **Neighbor Solicitations** — NS messages are sent by a node when it needs to resolve an IPv6 address to a datalink address. NS messages are sent to the solicited node's solicited-node multicast address. NS messages are relayed by the NDS tree to the NDS(LL) server. The server first determines the scope of the solicited address. If the scope is link-local then the NDS(LL) "broadcasts" the NS back to all nodes on its LL. Each

node must filter incoming packets from the NDS tree and only admit those packets addressed to it. If the solicited address is site-local or global then the NDS(LL) server determines if the address prefix is on its local LL or another LL on the NDS tree. If it's a local prefix then the NS is "broadcast" on the local tree. If not, then the NS message is relayed up the tree and is eventually delivered to one or more NDS(LL) servers for other LLs. These servers then "broadcast" the NS to all nodes on their respective LLs. Thus, the solicited node will receive the NS packet to it can respond to it.

- **Neighbor Advertisements** — Solicited NA messages are sent on a VC connecting the solicited and soliciting nodes. When a node receives a valid NS message it creates a VC to the soliciting node and then unicasts the NA message back to the soliciting node over this VC.
- **Redirects** — Router redirects are always sent over the VC which connects a node with the router. When a node sends a packet to an off-link node via a router it must first open a VC to the router. This is done by the sending node if the router provided a Datalink Address option in the RA messages. If the router did not provide this option then the sending node must first perform Neighbor Discovery on the router and the router will open a VC to the soliciting node in response to the node's NS message. When the router detects that the sending node should redirect its packets the router simply unicasts the Redirect message over the VC connecting the node and router. The sending node will then take whatever actions are necessary to connect to the target of the redirect.

Note that in all cases above no changes are made to the network layer. Also, complete semantics of Neighbor Discovery are maintained since it is always the solicited node that chooses ˝what (and even if) to reply to the solicitation. This also maintains complete IPv6 security semantics and features since the concept ˝of the security association is not being changed for ATM. This will allow nodes to choose their responses to solicitations based on security information as is done with other datalinks. Thus, ATM will be transparent to the network layer except in cases where extra services (such as QoS VCs) are offered.

3.1. NDS Trees

Implementing the physical NDS levels under UNI 3.0/3.1 and UNI ˝4.0 will require that each entity (either a node or NDS server) at a specific level in the tree be configured with the ATM link-layer address of the NDS server (or servers if redundant or distributed NDS severs are used) immediately above it. Of course, this will require entering a great deal of link-layer addresses into every entity on the tree. Not only is this time consuming and error prone, but since ATM link-layer addresses are not fixed, some changes in ATM network topology could cause some NDS server's address to change, requiring reconfiguration of some number of nodes or servers.

The UNI 4.0 [ATM-UNI40] specification defines an ATM anycast capability which can be applied to this framework to the NDS tree to be completely self configuring. The UNI 4.0 anycast ˝capability allows ATM nodes to register group addresses (anycast addresses) with the network. A group address is registered with a specific

reachability scope which determines which nodes will be able to connect to the registering node using the registered group address. This reachability is determined by the ATM network hierarchy and the PNNI [ATM-PNNI] call routing mechanisms. The UNI 4.0 (and PNNI) specification define two types of group ˝address scoping. One type is referred to as PNNI routing scope and is directly related to the PNNI network routing hierarchy. This scope has 104 level, but is very closely tied to the network topology. The second defined scope is called the organization scope. This scoping is a logical scoping which has been designed to more naturally map on to human organizations. This type of scope has only 16 levels. This framework uses organizational scope for the NDS autoconfiguration functions because it represents a much more manageable number of levels, and because ˝it not as dependent on ATM network topology as PNNI scoping. The UNI 4.0 and PNNI specification define the following scopes:

- Local Network (LN)
- Local Network plus one (LN+1)
- Local Network plus two (LL)
- Site Minus two (S-2)
- Site Minus one (S-1)
- Site (S)
- Site Plus one (S+1)
- Organization minus one (O-1)
- Intra-Organization (O)
- Organization Plus one (O+1)
- Community minus one (C-1)
- Intra-community (C)
- Community plus one (C+1)
- Regional (R)
- Inter-regional (IR)
- Global (G)

When an ATM node places a call to a group address it must specify the highest scope through which the call will be routed. The caller can limit the scope within which a call will be routed, ˝and thus, what set of target nodes can be reached. A call will only be delivered to a called node if the scope of the advertised ˝group address includes the calling node, and if the routing scope specified in the call includes the target node's group address scope. A call to a group address is routed to the "nearest" node based on the ATM network topology.

Using UNI 4.0 anycast addressing, it will be possible to make the NDS tree autoconfiguring by having every entity (either a node or an NDS server depending on the level) at one level in the tree connect to the NDS server above it simply by placing a call to ˝the server's group address using the appropriate scope in the call. Using this mechanism, an NDS tree would form based largely on the topology of the ATM network and the placement of servers within that topology. In this way, only a small number of well known addresses would be required to configure a very large tree, or even multiple trees. Effectively, only one well known address at each of the 16 possible levels of the physical tree would be required.

In UNI 4.0 networks, where group addresses are utilized, there **MUST** be one unique group address at each level of the hierarchy which servers at each level register (for a total of 16 assigned

group addresses). Multiple group addresses are required since ˝the routing of group address calls is always to the "nearest" (in terms of the ATM network topology) node which has registered that address with the network. Because of the way ATM anycast call routing is specified, it would not be possible to autoconfigure the hierarchy using only a single group address and relying on ˝the anycast routing scope rules to direct the call to the intended recipient. Effectively, the "nearest" node that registers a group address "masks" the presence of servers who advertised the same group address with a greater scope. The group address used by each level of NDS server is denoted by GA(*level*) where *level* is one of the organizational scopes described above. Each GA(*level*) is registered with the network with the scope corresponding to ˝the level (that is, GA(C) would be registered with the community scope).

Group address are assigned by the ATM Forum. To support this framework with IETF would request a set of group addresses from the ATM Forum and assign them for use by IPv6 over ATM. These group addresses would be well-known addresses permanently ˝assigned to IPv6 over ATM.

The following sections describe how this hierarchical tree of NDS servers is organized and configured both manually and automatically.

3.2. NDS Tree Configuration

An NDS server tree or subtree is formed by interconnecting some number of physical NDS servers is a specific manner. The connections between nodes and NDS servers at any level of the ˝tree use the same procedures, and set up the same type and number of connections. This results in a tree in which each level is organized exactly like any other level, and in which only one procedure for interconnection of the levels must be used. The manner is which the NDS tree is connected is done to permit a specific type of dataflow within the tree.

The NDS tree is composed of a number of physical NDSs at some number of levels connected with the following ATM VCs:

- Connected to the NDS at the next higher level with a PtP VC.
- Connected to the NDS at the next higher level by a PtM VC which is rooted in the NDS at the next higher level.

With these connections, each NDS has a PtP connection over which it can communicate with each individual NDS below it, and a PtM ˝VC over which it can "broadcast" to each NDS below it. This connection structure, when extended to some arbitrary number of levels, provides a way for packets to be "broadcast" to a set of NDSs or nodes on the tree (or to all nodes).

The general dataflow in the tree is that packets are always unicast up the tree towards the root. At some point an NDS can "broadcast" the packets back down the tree as required by the protocols. An NDS can also choose to unicast down the tree to a specific server under it according to the processes described later in this document.

The physical NDS tree can be of any height or width as needed. However, at some point on the tree the boundaries for a Logical Link and Site must be defined. The specification of the NDS(LL) or NDS(Site) servers can be at any layer in the tree based on ˝node connectivity requirements. Thus, the establishment of an NDS(LL) or NDS(Site) server is done entirely by configuration rather than placing these servers at some arbitrary level in the tree. Once ˝a physical NDS has been designated as an NDS(LL) or NDS(Site), then that NDS must perform additional special functions. Also, all ˝NDS servers above the NDS(LL) servers must be configured as being outside a LL so they can perform additional ND message routing functions as described later (servers below the NDS(LL) do not perform these functions).

The part of the NDS tree below the NDS(LL) server is referred to as the LL subtree. The part of the NDS tree below the NDS(Site) server is referred to as the site subtree.

In each LL there **MUST** be exactly one active NDS(LL) server (there can be multiple "hot" standby servers). In each Site, there **MUST** be exactly one active NDS(Site) server. In the whole NDS tree there **MUST** be exactly one active NDS(Root) server. The NDS(Root) server can be the same as the NDS(LL) or NDS(Site) servers if the tree terminates at that level.

3.2.1. Manual NDS Tree Configuration

In UNI 3.0/3.1 or UNI 4.0 environments where manual configuration is used, each node and NDS must be configured with at least one ATM link-layer address of the NDS to which it must connect. An NDS that has no configured NDS address **MUST** be configured as the NDS(Root) server.

Each node or NDS will be configured with the ATM link-layer addresses of a set NDS servers to which it can connect. Each NDS **MUST** connect to exactly one NDS server above it unless it is the NDS(Root) server (which by definition has no NDS server ˝above it). Nodes may connect into multiple LLs. A separate table of NDS server addresses **MUST** be maintained for each LL. Each node **MAY** also make more than one connection to a given LL.

All nodes **MUST** create and bind to at least one ATM NSAP for each LL connection they will make. This NSAP is used to connect to ˝the NDS tree, to receive the PtM connection from the called NDS server, and for receiving data connections from other nodes. The node **MAY** bind to separate NSAPs for each of these connections as long as it keeps track of which NSAPs it will use for each type ˝of connection. All NDS servers, other than the root server, **MUST** also create this (these) binding(s).

When a node or server is configured to connect to more than one NDS, the ATM link-layer addresses configured in the NDS server table **SHOULD** be placed in order to prioritize the order in which NDS servers should be used. This will provide load distribution between multiple NDS servers at the same level, while still permitting tree function and failover in the event one server fails.

All NDS servers **MUST** bind to an ATM NSAP over which they will accept calls from nodes or NDS servers below them. This NSAP **SHOULD** be as deterministic as possible since it **MUST** be entered in

the table of NDS addresses for servers or nodes below. If this NSAP changes all servers or nodes which connect to this NDS must be reconfigured.

The process for connecting any node or NDS into the NDS tree is the same.

- Starting with the first valid ATM link-layer address in its table, the node or NDS places a PtP call to that address. If the call fails then the next address is tried. If calls to all addresses fail then the node or NDS backs off for some random period of time [TBD] and repeats the process until a call succeeds.
- When the PtP VC is established the node or NDS will perform a registration process [TBD] with the called NDS.
- When the registration process is completed successfully, the called NDS server will add the calling party to a PtM connection which it uses to "broadcast" packets to all connected nodes or NDSs.

This process is performed by every node and NDS on the tree until the entire tree is formed.

The tree **MAY** also be statically configured using PVCs.

If an NDS fails the nodes or NDSs to which it is connected will immediately receive an ATM RELEASE notification, and all the server's VCs will be torn down by the network. When the nodes or servers below the failed server detect this condition **MUST** initiate the initial connection process described above. The NDS server above the failed NDS server or node **MUST** destroy any cached information associated with the failed server or node, but **MUST NOT** take any action to re-establish the connection.

3.2.2. Autoconfiguration of Trees

Autoconfiguration of the NDS tree is only possible on UNI 4.0 ATM networks which implement anycast addressing. In fully autoconfiguring NDS trees, the number of physical levels in the tree is limited to 16 by the UNI 4.0 organizational scoping. If more levels are needed they must be manually configured.

Not all 16 levels need exist in an autoconfigured tree. Only those levels necessary to provide the needed level of ATM connectivity and scalability have to exist.

Note that autoconfiguration only works on private ATM networks since anycast addressing has not been defined for public ˝networks. An NDS tree on a public network would have to be configured manually (probably via PVCs for Internet providers).

In autoconfiguration environments, each NDS registers GA(X) for its level in the hierarchy. For example, the NDS at level LL would register GA(LL). The scope of the registered GA(X) depends on the height of the tree and the connection scope required by each NDS. Each NDS can register any reachability scope with ˝GA(X) as required to provide the needed connectivity. The only restrictions in group address scoping are that those NDS servers below the NDS(LL) server **MUST NOT** register their GA(X) with a scope wider than the GA(X) of the NDS(LL) server. Similarly, NDS servers below the NDS(Site) server **MUST NOT** register their GA(X)

addresses with a scope wider than the NDS(Site) server or less than the NDS(LL) server. Finally, NDS servers between the NDS(Site) server and the NDS(Root) server (if they are not the same) **MUST NOT** register their GA(X) addresses with a scope wider than that of the NDS(Root) server, or lower than the NDS(Site) server.

These scoping requirements permit some backup and redundancy in various levels of the tree. For example, if all nodes in an LL are connecting through NDS servers at the LN level, each NDS(LN) server could register GA(LN) with a scope greater than LN and ˝less than or equal to the scope used by the NDS(LL) server (which ˝would have to be at level LN+1 or higher). With this type of scoping, nodes would normally connect to the closest NDS(LN) server based on the ATM network topology. However, if that server were to fail, nodes would the automatically reconnect to the next nearest NDS(LN) server.

To perform autoconfiguration all nodes and servers create and ˝bind to all NSAPs as specified for manual configuration. NDS servers **MUST** always have unicast NSAPs through which they can be reached. In addition to the NSAPs described for manual configuration, all NDSs **MUST** register GA(X) for its level with appropriate scope and bind to that NSAP to receive calls.

Each node and NDS on the tree **MUST** be configured with its level on the tree. These levels are numbered 0 through 16, with 0 being a node, and 16 being the highest level NDS server permitted.

To configure the tree, the following process is used. Each node and NDS **MUST** have a table with the list of GA(X) for each level X.

- The node or NDS at level X places a call to GA(X+1) with a call routing scope of X+1. If the connection fails then a call to GA(X+2) with call routing scope of X+2 is tried. If the end of the GA table is reached the node or NDS backs off for some random period of time [TBD] and then repeats the procedure.
- If the call succeeds, the calling party performs a registration process with the called NDS. Part of this process involves providing the called NDS with the calling party's level in the hierarchy. If the calling party's level in the hierarchy is not appropriate for the called party (i.e., a node reaching an NDS above the NDS(LL) server), the connection is released, the index into the caller's GA table is reset, the process is repeated until an NDS at the appropriate level is reached.
- If the called NDS learns through the registration process that it has connections from more than one level in the hierarchy, it will release the connection from the node or NDS at the lowest level. NDS servers **MUST NOT** permit connections from more than a single level below (otherwise the tree structure would not be regular).
- Once the registration process completes successfully, the called NDS server adds the calling party to its PtM connection which it uses to "broadcast" to all calling parties.

If an NDS server fails, the nodes or NDSs to which it is ˝connected will immediately receive an ATM RELEASE notification, and all the server's VCs will be torn down by the network. When the nodes or

servers below the failed server detect this condition they **MUST** initiate the initial connection process described above. The NDS server above the failed NDS server or node **MUST** destroy any cached information associated with the failed server or node, but **MUST NOT** take any action to re-establish the connection.

As stated earlier, there **MUST** be exactly one NDS server at the NDS(LL), NDS(Site), and NDS(Root) levels. This means that only one physical NDS(X) server which functions as one of these ˝special servers is allowed to register GA(X) at any given time. This restriction is required to avoid fragmenting a LL or Site, which is what would occur if there were more than one server which registers a specific GA with the same scope. By restricting ˝these levels to a single NDS sever, the same procedures used achieve redundancy and failover at other levels (where multiple servers are permitted) can not be used. Thus, some new procedure must be defined to achieve failover at these levels. This procedure is only necessary for the servers which function as the NDS(LL), NDS(Site), and NDS(Root) servers.

This procedure is TBD.

3.3. Forming Logical Links

Logical Links are the basic unit of the NDS tree. Like LAN segments, they form the basis of an IPv6 ATM "internet". Logical Links act very much like LAN segments, and are the point to which all nodes connect.

At a minimum, a Logical Link consists of one NDS(LL) server and one or more NDS(N) servers to which one or more nodes are connected. Minimally, the NDS(LL) server can also function ˝as the NDS(N) server. In a Logical Link there is always exactly one active NDS(LL) server. This server directs the ND message ˝traffic flow within the LL, controls which messages are forwarded outside the LL to the Site, and controls what messages are admitted into the LL from the Site.

Physically, the NDS(LL) server does not have to be the server to which all nodes are connected, nor does it have to be the only server on the LL. All that is required is that the single ˝NDS(LL) server be located at the root of the LL physical NDS subtree. ˝The LL physical NDS subtree can be as wide or deep as needed to achieve the desired node connectivity and LL scalability. This means that there could be any number of NDSs within the LL, but that the NDS which is the root of the LL subtree must be explicitly identified as the NDS(LL) since it must perform ˝special functions.

The underlying characteristic of an LL is that all nodes on the ˝LL are members of the same IP subnets (they use the same prefixes to configure their addresses), and that any node on the LL can use link-local addresses to reach any other node on the LL. Link-local addresses are limited in scope to Intra-LL communications and are not visible outside the LL. Finally, an LL forms a "broadcast" domain in which any node can "broadcast" data to all other nodes on the same LL (but not to other LLs on the same ATM network).

The following figure shows an example of a four level LL. The nodes are all connected to NDS(0) servers. These are connect to

NDS(1) servers, which are then connected to NDS(2) server. The NDS(2) servers are finally connected to an NDS(3) server. The NDS(3) server is also configured as the NDS(LL) server. In UNI 4.0 environments where autoconfiguration is enabled, the scope of the group addresses GA(0), GA(1), and GA(2) must be less than or equal to the scope of address GA(3). Thus, the scope of the ˝LL on the ATM network would be the scope with which the GA(3) ˝address is registered. All nodes within this scope would connect to this ˝LL.

A Logical Link has the following properties:

- All nodes in a LL can establish direct ATM virtual circuits to any other node on the LL
- All nodes in a LL can perform Neighbor Solicitation to discover the ATM address of any other node on the LL using link-local addresses or global addresses
- Host on one LL can not be discovered or contacted using link-local addresses by nodes on an other LL

Node can connect to multiple LLs. This places one level of abstraction between the IP layer and the ATM layer, essentially virtualizing the ATM layer into multiple logical links. In manually configured environments the ATM NSAP of an NDS server to which a node connects for each LL that a node could join would have to be configured on the node. In a UNI 4.0 ˝autoconfiguration environment the node could join only the closest LL using autoconfiguration, but could join other LLs using manual configuration. However, if a node wanted to join a specific subnet (as opposed to a specific LL) this could be handled via a special protocol which will be described later.

In most cases, each LL will be associated with a single IP ˝subnet. In this case a LL will equate to an LIS. However, just as it is permissible to configure multiple subnets on a physical LAN segment, it will also be possible to configure multiple subnets ˝on an LL. This framework places no restrictions on how IP subnet ˝can be formed when compared to existing LAN technologies.

Just as with legacy LANs, IP subnets can not span multiple LLs. With legacy LANs, IP subnets can not be spread across multiple, physically disjoint LAN segments (without further subnetting). This restriction remains when using LLs. Multiple LLs can not be aggregated at the link level. Aggregation occurs at a higher level which does not permit the sharing of subnet addresses by ˝two LLs.

3.3.1. Neighborhoods

There is one last special class of NDS server. This is called ˝the Neighborhood server, or NDS(N). The NDS(N) server is the server to which individual nodes **MUST** connect. NDS(N) server **MUST NOT** accept connections from other NDS servers. NDS servers other ˝than the NDS(N) server **MUST NOT** accept connections from nodes. Thus, the registration process used by the calling party connecting to the NDS will provide the information the NDS needs to determine the identity of the connecting party (this registration ˝process is also used for security to control which nodes can connect to the LL as described in the Security section).

In the figure above, the NDS(N) servers are the NDS(0) ˝servers.

3.4. Forming Sites

A site is simply a collection of a set of LLs whose nodes can connect using IPv6 site-local addresses. A site is built from multiple LLs by connecting the NDS(LL) servers to an NDS(Site) through zero or more intermediate NDS servers. Each site terminates at exactly one NDS(S) server. The NDS(Site) level server must exist to establish a site. The number of levels in the Site tree can be extended to provide the needed levels of connectivity and scalability.

The NDS(S) server performs filtering of ND messages entering and leaving the site. It will not pass any site-local solicitations outside the site, thus limiting the scope of site local ˝addresses.

The site layers of the hierarchy are illustrated in the following figure. This figure illustrates a site composed of three levels of NDS servers above the LL layer. The site terminates at a single NDS(Site) server three levels above the NDS(LL) servers. The NDS(X+2) server is configured as the NDS(Site) server. ˝In UNI 4.0 autoconfiguration environments, the NDS server between the NDS(LL) servers and NDS(Site) server **MUST** register their respective global addresses with a scope greater than the scope ˝of the NDS(LL) servers, and less than or equal to the scope of the NDS(Site) server. Thus, when using autoconfiguration, the scope of the site is the scope with which the NDS(Site) server ˝registers its group address. The site will include all nodes within this scope as determined by the ATM network topology. The scope of ˝the Site NDS(Site) server can be selected so that it includes all the nodes which are to be members of the Site.

A site has the following properties:

- All nodes in the site can connect to all other nodes in the site by creating an ATM virtual circuit
- All nodes in a site can perform Neighbor Solicitation to discover the ATM address of any other node on the site using site-local address or global addresses
- Hosts in one site can not be discovered or contacted using site-local addresses by nodes in another site

3.5. Beyond Sites

The tree structure can be extended by more levels as needed. Sites can be grouped together to create larger groupings of ˝nodes. However, only global IPv6 address can be used between nodes of different Sites. The tree can be extended up as far as is necessary and can be made as wide as necessary to include all the Sites which will be allowed to communicate.

A full tree is used to manage reachability of a set of nodes on ˝an ATM network. These nodes are logically grouped into logical ˝links and sites, but are capable of directly establishing a connection to each other. Multiple trees can exist within the same ATM network. In this case, nodes managed by different trees would be

required to go through a router to communicate even if they were capable of creating direct connections. A tree consists entirely of nodes which are allowed to directly connect. A node can not ˝be part of a tree if it can not connect to every other node on the tree.

There is no limit to the number of LLs they may join, or restrictions on which LLs they may join. A node can connect to many different LLs, either within the same NDS tree, or in different NDS trees. Hosts that must belong to a specific LL, ˝but which are not connected to the ATM network within the scope of an NDS(N) server with that LL, can connect to specific LL in two ways:

1. Connect directly to an NDS(N) in the target LL by manually configuring the unicast address of the NDS(N) server(s) in that LL.
2. Allow the NDS tree to relocate the node using the protocol described under manual IPv6 address configuration. For the node to be automatically relocated, the node must be assigned an IPv6 address (via either DHCP or manual IPv6 address configuration), which belongs to another LL.

In UNI 4.0 autoconfiguration environments the configuration of ˝the tree will be determined by the ATM network topology. That is, ˝the topology of the network will determine which systems can interconnect using group addresses of a given scope. Thus, NDS server systems should be chosen taking their location on the ATM network into account. Having various NDS servers register their address with a higher scope will also relieve some of the topological restrictions.

Finally, any system can perform NDS server functions at multiple levels. There is no restriction (other than possibly ATM topological restrictions) on how many levels can be served by a single physical system. All that has to be done is to configure the system appropriately. The system would then register ˝multiple GAs and the appropriate scopes. However, there is no provision made in this specification for providing Intra-system cut-through in cases where the same system is serving adjacent levels. In this case, all communications between NDSs at adjacent levels ˝will be carried across the ATM network (though only to the server's local switch).

4. NDS Operations

Once formed, the NDS tree provides a way for reachability information to be circulated between LLs as well as a way of "broadcasting" or "multicasting" certain information. IPv6 addressing scopes are mapped on to the layers of the tree by designating specific NDS servers a Logical Link, Site, and root servers. These NDS servers perform the specific functions necessary to implement IPv6 addressing semantics within the tree. All NDS servers work together to maintain IPv6 Neighbor Discovery semantics within the tree. Thus, the tree structure maps IPV6 onto ATM in a consistent manner.

The flow of data through the tree has the following characteristics:

- data from the leaves of the tree is unicast through the NDSs towards the root
- data coming from the root can be "broadcast" towards the leaves of the tree
- data from the root can also be unicast towards specific leaves or branches (this is also referred to as "directing" the data down the tree)

This dataflow provides full ability to "broadcast" data from any node to all other nodes at any level in the hierarchy. Because ˝of this feature, it is possible to use the existing IPv6 ND and DHCPv6 protocols with no modifications. Also, since the ability to "broadcast" solicitation and other packets is provided, there is no need for any central server to maintain node reachability information. Host reachability information can continue to ˝reside in each node, and can be queried by any other node connected to the tree. This enables the NDS tree to provide the same ˝semantics as ND expects on legacy LANs.

The IPv6 ND and DHCPv6 protocols require that nodes join various multicast groups. In the NDS tree, multicasting reception of ND packets is handled simply by filtering incoming packets which arrive on the node's PtM connection from the NDS(N) server. That is, all multicast packets are really "broadcast", and each node must perform appropriate address filtering. For example, all ˝non-router nodes must discard any packets received which are ˝addressed to the all-routers multicast address. Since this framework specifies only a specific set of multicast addresses which are to be treated in this fashion, and since packets addressed to these addresses will only arrive on specific VCs, this filtering can be provided without affecting general multicast functions ˝[IP-ATMMC].

The tree can be used to send packets to the following multicast addresses:

- The all-nodes address
- The all-routers address
- The DHCPv6 server/Relay Agent address
- The solicited-node address

The tree is not intended to handle the general multicast case and must not be used for this purpose (it would be undesirable to use it for this purpose since the load on the NDSs and latencies ˝would be unacceptable for sending data). The broadcast mechanism of ˝the LL should be used only for sending broadcast data that is used ˝to determine node reachability, service discovery, or routing protocol information.

The IPv6 ND solicitation/advertisements mechanisms work as follows:

- all solicitations and advertisements sent to a multicast address are sent to all nodes on an LL
- Unicast advertisements in reply to a solicitation are send via a VC which the solicited node creates to the soliciting node. Multicast replies are sent to all nodes on the LL.

The following sections describe how the existing IPv6 protocols are applied and how any new protocols will be used.

4.1. Router Advertisements

Router Advertisements are sent by routers to the all-nodes multicast address by all routers at periodic intervals or in response to a Router Solicitation message. These advertisements not only contain information about the router being advertised, but also address configuration information which is used by each node to perform stateless address configuration and to determine which prefixes are on-link [IPv6-ADDRCONF and IPv6-ND]. The Router Advertisement can also contain information about the network such as what the default MTU size should be for communications on the subnet, and whether stateful address configuration should be used.

Router Advertisements are handled specially by the NDSs since ˝they contain information about the local network (LL) configuration. There are no required changes to either router or nodes for ATM. They will continue to use Router Advertisements as described in IPv6-ND. The special work is done entirely by the NDSs and is transparent to IPv6 entities.

Like any node, a router must first connect to an NDS(N) server to become part of a Logical Link. Part of the process of joining a logical link is the registration of the node with the NDS(N) server. When a router connect to the NDS(N) server and ˝registers, it will simply identify it's self as a normal node.

When a router sends out an advertisement (either solicited or unsolicited), it **SHOULD** send it to the all-nodes multicast address. To do this, it must send it advertisement out on its ˝PtP VC that connects to its NDS(N) server. The NDS(N) server will forward the message up the tree. The message will continue up ˝the tree until it arrives at the NDS(LL) server (which could be the same as the NDS(N) server). The message is moved up the tree by each NDS server receiving the message on one of its PtP VCs from the server below, then forwarding the message to its PtP VC to ˝the server above. When the NDS(LL) server receives the Router Advertisement it **MUST** record all prefix information contained in the router advertisement. This information is used by the ˝NDS(LL) to determine which prefixes are on it's LL.

All the prefix information contained in local Router Advertisements is placed in a cache in which each entry **MUST** age using the same algorithm used by nodes to age prefix information. This will assure that prefix information is aged consistently throughout the LL. Each time an NDS(LL)server receives updated prefix information it **MUST** update the aging information in it's prefix cache based on the latest lifetime values.

After saving the prefix information, the NDS(LL) server **MUST** forward the unmodified Router Advertisement message out its PtM connection which connects to all NDS servers immediately below ˝it. This will result in the multicasting of the Router Advertisement to all nodes on the LL.

If the NDS(LL) server is not connected to a Site then it is not required to take any more actions to process a Router Advertisement.

4.1.1. Exporting Prefix Information

If an NDS(LL) server is connected to a Site, then it must export all Router Advertisement messages so that its local prefixes are advertised to other LLs on the tree. To do this, the NDS(LL) ˝**MUST** copy each router advertisement it receives and forward it on the PtP VC which connects the NDS(LL) server to the Site. The ˝NDS(LL) server **MUST NOT** make any modification to the Router Advertisement messages.

4.1.2. Routerless Logical Links

Under this proposal, each Logical Link need not have a router. ˝In fact, since nodes on-linked LLs can communicate directly, routers would only be needed when nodes need to communicate off-link.

In cases where no routers are present, nodes will have no entries in their default routers table. When a node's default router table is empty a node considers all prefixes on-link. Thus, an NDS tree in which no LLs have routers will operate as long as hosts are manually configured (or use NHCPv6) with site-local or global addresses. The NDS(LL) and NDS(Site) servers will enforce the IPv6 addressing scope rules.

However, it will probably be more convenient (and efficient) for each LL to advertise it's prefixes, but still be configured without a router. In this case, some node on the LL **MUST** be configured as a non-forwarding router so that it will distribute prefix information in its router advertisements. Local nodes can then use the prefixes for autoconfiguration, and the NDS(LL) can export the prefixes to other LLs.

4.1.3. Movement of Prefix Information Outside Logical Links

When an NDS(LL) exports a Router Advertisement message, the message will be sent towards the root of the tree by each NDS server relaying the information out its PtP VC that connects to the next higher NDS server. When the RA message arrives at the root of the tree the NDS(Root) will relay the message down ˝via its PtM connection which connects all the NDS servers immediately below it. These server will in turn multicast the RA to all servers immediately below them. This process is repeated at each level of the tree until the RA message arrives at an NDS(LL) server. The NDS(LL) servers then take the RA messages and ˝process them as described later.

As the Router Advertisement message travels up the NDS tree each NDS server above the NDS(LL) server **MUST** record all the prefix information contained in the RA message in a prefix cache. The prefix entries in this cache **MUST** be aged using the same algorithm nodes use to age prefix information. Along with the prefix information, each cache entry **MUST** contain a reference to the PtP

VC on which the RA message arrived. Thus, each NDS server will develop a table associating specific prefixes with specific PtP VCs. This table is used to direct Neighbor Solicitation message through the tree.

Note that the prefix cache is not required for proper operation ˝of the tree. It is used to optimize message flow through the tree ˝to reduce traffic and NDS server load. Thus, the NDS tree is ˝capable of being fully operational when all or some NDS servers have no state. State is used only to optimize performance of the tree. This also makes failover and recovery of an NDS server very easy. That is, if an NDS server fails and is rebooted (or replaced) it does not have to update its prefix cache right away. It can "learn" prefixes through normal traffic on the tree. This means that a recovered NDS server can be fully operational immediately on connection into the NDS tree and does not have to take any explicit action to fill its cache.

NDS servers below the NDS(LL) servers **MUST NOT** maintain any prefix cache information. The **MUST** only relay ND packets up or down the tree. The Logical Link its self is completely stateless ˝requiring no internal state for relaying ND packets.

4.1.4. Importing Prefix Information

To provide each local node with a list of prefixes which are available on the NDS tree, each NDS(LL) server will import prefix information received in the Router Advertisement messages other NDS(LL) servers export. When an NDS(LL) server receives an RA message from the Site it **MUST** extract prefix information from the RA message so that this information can be sent to its local nodes. The NDS(LL) servers **MUST NOT** simply relay the RA messages to their local LL. The NDS(LL) servers **MUST NOT** import prefix information which they exported.

The NDS(LL) does not relay any received Router Advertisements to the nodes on its LL. Instead, it creates its own Router Advertisements using the prefix information it receives from remote routers. Thus, the NDS(LL) server is a source of Router Advertisements on an LL, but it is not a router. The NDS(LL) becomes a source of prefix information, but not router information.

To create the Router Advertisement messages for the LL, the NDS(LL) server extracts prefix information from all RA messages received from external LLs. The following information **MUST** be extracted and placed in a table of external prefixes:

- The prefix length.
- The prefix value.
- The valid lifetime value.
- The preferred lifetime value.

All other information in the RA message **SHOULD** be discarded. The NDS(LL) servers **MUST** age each entry in the table according to the prefix valid lifetime value. The NDS(LL) servers **MUST** update the prefix table as new RA messages arrive.

The NDS(LL) servers uses the information in the external prefix table to build Router Advertisement message which are sent out on

its logical link. These RA messages **MUST NOT** be exported to other LLs. These RA messages **MUST** be sent to the local LL on a periodic basis or in response to locally generated Router Solicitation messages. The algorithm specified in IPV6-ND **SHOULD** be used to determine when these RA messages are sent.

The NDS(LL) servers will compose the Router Advertisement ˝messages using the following information:

- Source address of the unspecified address.
- Destination address of the all-nodes address.
- Router Lifetime of 0.
- Reachable Time of 0.
- Retrans Timer of 0.
- Max Hop Limit of 0.
- M bit set to be consistent with local router advertisements
- O bit set to be consistent with local router advertisements
- No Source Link-layer Address.
- MTU option set to be consistent with local router advertisements.
- Prefix information taken from the external prefix table:
    - Prefix length as received in the external RA.
    - Prefix value as received in the external RA.
    - The L bit is set.
    - The A bit is cleared.
    - Valid Lifetime adjusted to account for the time difference between the time the external RA message was received and the time the internal RA message is being sent. This will assure that all nodes will expire the prefix at the same time.
    - Preferred Lifetime adjusted to account for the time difference between the time the external RA message was received and the time the internal RA message is being sent. This will assure that all nodes will deprecate the prefix at the same time.

The values of the M and O bits and the MTU option **SHOULD** be set via a local configuration option, but **MAY** be obtained from local Router Advertisement messages. The NDS(LL) servers **MUST NOT** create and send the local RA messages unless these values have been explicitly obtained from a configuration option or from a local RA. That is, defaults are not permitted.

Finally, any needed security information **MUST** be added when a security association between the source and destination. All nodes on a LL **SHOULD** accept Router Advertisements from the NDS(LL) server when performing authentication or other security ˝checking.

Using this method of routing NDS traffic flow in both the up and down directions, messages only travel up the NDS tree until they arrive on the first NDS server that recognizes the target's prefix. From there, the message is directed to a specific LL.

All NDS servers outside an LL **MUST** time out all router and prefix information based on information received in the original Router

Advertisement messages. This **MUST** be done to assure consistency of prefix information used and advertised throughout the tree.

4.2. Router Solicitations

Router solicitations are sent by nodes to discover information about routers or to obtain a list of prefixes for stateless address configuration and on-link determination. Nodes send Router Solicitation messages in the normal way as described in IPV6-ND. The RS messages **MUST** be sent on the node's PtP VC which connects it to an NDS(N) server. The NS message is then relayed up the tree until it arrives at the NDS(LL) server. The NDS(LL) server then relays the RS message out via its PtM connection ˝which connects all NDS servers below. These servers in turn relay the RS message out their PtM connects until the NS message arrives at all nodes on the LL. All nodes will receive the RS message, but only routers will filter it and admit it.

The NDS(LL) servers **MUST NOT** relay any Router Solicitation messages outside their LL.

If an NDS(LL) server is maintaining a table of external prefixes it **MUST** respond to the RS message and compose and transmit an RA message as described above. The NDS(LL) server **SHOULD** use the algorithm described in IPV6-ND to limit the number of RA messages sent on the LL.

4.3. Neighbor Solicitation

Neighbor Solicitation are sent by nodes to discover the ˝link-layer address of other nodes. Neighbor Solicitation is normally done the first time data packets are to be exchanged between nodes. The information provided by the Neighbor Solicitation mechanism enables each node to unicast data to the other node. In an ATM environment, the result of successful Neighbor Solicitation ˝should be the establishment of a PtP VC between the soliciting and solicited nodes over which data can be exchanged. By default, this VC will carry only best effort traffic. On ATM networks, nodes **SHOULD** establish multiple VCs with traffic parameters and QoS setting to carry differing classes of traffic. The initial ˝VC setup between nodes **SHOULD** be used to carry all traffic with no QoS requirements.

Neighbor Solicitation is used by nodes with no modifications for ATM (other than the format of the link-layer address). Any node on a LL can reach any other node on the LL using link-local addresses. Any node in a site can reach any other node in the site using site-local addresses. Any node on the tree can reach any other node on the tree using global addresses. Because all nodes on the tree appear to be both physically and logically "on-link", any two nodes can communicate using IPv6 addresses of the appropriate scope.

All communications with nodes that are not "on-link" (i.e., not on the same NDS tree) will result in the sending node sending ˝packets to the destination through a default router. Default routers **MUST** exist within the same LL as the node sending packets to the external network since link-local addresses are used to establish

connectivity to the router. Routers can either forward the packets to the next hop, or redirect the sending node to another router or directly to the destination node. To redirect a node, the routers could be running some routing protocol such as NHRP which could locate and redirect to off-link destinations.

When a node needs to communicate with another node it will create a Neighbor Solicitation message and send it via the PtP VC to its attached NDS(N) server. The Neighbor Solicitation message will be forwarded up the tree to the NDS(LL) server. The NDS(LL) server **MUST** examine the target address of each Neighbor Solicitation messages which it receives to determine how the solicitation message is to be routed. The NDS(LL) server examines the target address of each solicitation message and determines if the scope of the target address is link-local. If the target address scope is link-local, the NDS(LL) server **MUST** relay the Neighbor Solicitation message to all NDS servers in it LL via its PtM connection. This "broadcasts" the solicitation down the tree until it reaches all nodes on the tree. The node who's address matches the target address reply to the solicitation as described in the next section.

If the scope of the target address in a Neighbor Solicitation message is site-local or global, the NDS(LL) server **MUST** compare the prefix of the target address with its list of prefixes which are on its LL. If a match is found then the NDS(LL) server will relay the message down its LL subtree as described above. If no match is found then the NDS(LL) server **MUST** send the Neighbor Solicitation message to the NDS server above it via its PtP VC to that server.

As each Neighbor Solicitation message travels up the tree towards the root, each NDS server **MUST** compare the prefix of the target address in the message with its list of prefixes in it's prefix cache. If a match is found, the NDS server **MUST** relay the Neighbor Discovery message down the tree on the VC associated ˝with the matched prefix. If no match is found then the Neighbor Discovery message **MUST** be sent up to the next level of the tree towards the root.

As each Neighbor Solicitation message moves down the tree, each NDS server at the LL level an above **MUST** compare the prefix of the target address in the message with the list of prefixes in its prefix cache. If a match is found, then the NDS server **MUST** relay the Neighbor Solicitation message out the PtP VC associated with the matched prefix. If no match is found then the NDS server **MUST** transmit the message on the PtM VC connecting to all servers below.

4.4. Neighbor Advertisements

There are two types of Neighbor Advertisements: solicited and unsolicited. The handling of these two types of advertisements ˝by the NDS tree is very different.

When a node receives a Neighbor Solicitation which contains its address in the target address field, it must send a reply to the soliciting node so that each node will be aware of the other's link-layer address. However, unicasting a message through the tree structure is very difficult, and probably unnecessary. When a node receives a Neighbor Solicitation it can assume that its being solicited because the soliciting node is ready to

communicate with it. Thus, establishing a connection to the soliciting node is an optimization that will save traffic and ˝work on the tree. Thus, when any node **MUST** respond to a Neighbor Solicitation it will do so by placing an ATM call to the soliciting node and sending the Neighbor Advertisement via the newly established VC. This VC will be of the default type (best effort), with the default MTU (9188) unless the two nodes negotiate a different MTU using UNI signaling when the call is placed.

Unsolicited Neighbor Advertisements are sent by nodes when they change their link-layer addresses. Unsolicited Neighbor Advertisements are sent to the all-nodes multicast address. ˝These messages **MUST** be sent via the VC which connects the node to its NDS(N). When a Neighbor Advertisement message arrives at the NDS(LL) server, the server **MUST** send the advertisement back down the LL via its PtM connection to NDS servers below. This will "broadcast" the unsolicited advertisement to all nodes on the sender's LL. Additionally, the advertisement **MUST** be sent out on every PtP VC which connects the sender to other nodes. This must be done to assure that every node the sender is in contact with receives the sender's updated link-layer so that any cached link-layer data can be updated.

4.5. Anycast Addresses

The current IPv6 address architecture draft [IPV6-ADDR] restricts the use of anycast addresses to routers. This is done because ˝the use and routing of anycast messages is not yet well defined or understood for legacy LANs. However, the ATM environment and NDS tree offer a way to properly handle anycast addresses. Thus, ˝this framework includes a specification for the handling of anycast addresses within the NDS tree

Anycast addressing is handled in a hierarchical fashion, and follows the layout of the tree. Anycast addresses will have a "scope" within the tree, just as any other address. However, anycast addresses scope can be as small as a neighborhood. ˝When a node needs to connect to an anycast address it will reach the ˝node "closest" to it, based on the tree hierarchy.

When a node registers its anycast address on a node, the node **MUST** be configured to know that the address is an anycast address. This information **MUST** be passed to the IPv6 ATM interface so that the ATM code can take special actions. When an anycast address ˝is assigned to an IPv6 ATM interface, a special anycast ˝advertisement message **MUST** be sent out to the NDS tree. This message will follow the format of all other new inter-NDS messages. This is a required extension for ATM only to enable anycast address ˝handling to work efficiently within the NDS tree. All anycast advertisements travel up the tree as far as permitted by the IPv6 scope of the anycast address type (anycast addresses use unicast address formats). As they travel up the tree ALL NDSs **MUST** record the anycast address along with the VC of the server (or node) ˝from which the anycast advertisement arrived. When the root NDS ˝server receives the anycast advertisement it **MUST** record the anycast address, and **MUST NOT** resend the Neighbor Advertisement back down the tree.

When a node needs to connect to an anycast address it **MUST** send a Neighbor Solicitation with the anycast address as the target address. This message is relayed up towards the root. Because IPv6 anycast addresses are assigned from the normal unicast address space the NDS servers will not be able to distinguish a solicitation for an anycast address from one for a unicast address. Thus, ALL Neighbor Solicitation messages **MUST** be processed using the following procedure. Since some environments will not need to have (or want) anycast capabilities. To allow anycast capabilities to be disabled, every NDS server **MUST** have a configuration variable that can be set to disable anycast functions. Anycast capabilities **MAY** be disabled on a per-LL or per-subtree basis by disabling NDS server anycast processing on the appropriate set of NDS servers.

When a Neighbor Solicitation arrives at each NDS server, each server **MUST** examine the target address of the solicitation and compare the solicited address with the list of anycast advertisements it has received. If the solicitation's target address matches an anycast address in the NDS server's list, then the NDS server **MUST** relay the Neighbor Solicitation message to the VC from which the anycast advertisement was received. If the target address does not match an anycast address in the NDS server's table, the server **MUST** then perform the normal Neighbor Solicitation message processing as described earlier. Thus, the Neighbor Solicitation is sent up through the NDS tree normally, only going up the tree hierarchy as far as is necessary.

Nodes which advertise anycast addresses **MUST** resend their advertisements on a periodic basis.

If an NDS fails and recovers, it will loose any cached anycast data. In this case the server **MUST** forward all solicitations as it normally would during the server recovery process.

It may be possible that two nodes will advertise an anycast address within the same scope of the NDS tree. That is, some NDS server will receive two or more advertisements for the same anycast address, but from different VCs. When this occurs, it means that there is more than one "nearest" node advertising the anycast address within the subtree below the server. In such cases, the server **SHOULD** direct Neighbor Solicitations to alternate branches of the subtree so as to distribute the connections to the nodes which registered the anycast address. The method for doing this is up to specific implementations, but the minimum of distributing solicitations in a round-robin ˝fashion **MUST** be implemented.

As the result of NDS server failures or other conditions, it may be possible for a Neighbor Solicitation to reach more than one node that has advertised an anycast address. In this case, the soliciting node will receive more than one VC and Neighbor Advertisement from the solicited nodes. It is then up to the soliciting node to select one of the responding nodes for communications, either with or without the co-operation of the responding nodes. The choice of a responding node depends on the specific protocol or service which is connecting, and it is ˝beyond the scope of this document to specify how responding nodes are to be selected.

When a node that has advertised an anycast address fails, the NDS(N) server to which it has connected will detect the failure when it receives the VC RELEASE notification from the ATM ˝network.

When an NDS(N) server receives a RELEASE for a VC for which it ˝has recorded anycast advertisement data, it **MUST** send an anycast flush message up the NDS tree towards the root. This message **MUST** contain the anycast address being flushed. All nodes that ˝receive the anycast flush message **MUST** remove the entry in their anycast table for the address being flushed associated with the VC on which the flush is received.

If a node de-configures an anycast address on its IPv6 ATM interface, it **MUST** send an anycast flush message to the NDS tree.

Using this method, "closest" means a node which can be reached by ascending the minimum number of levels up the tree. For example, if one system in every neighborhood advertises the same anycast address anycast solicitations to that address would never leave the neighborhood since the NDS(LL) server would know how to ˝direct the anycast Neighbor Solicitation. If one node in each LL advertised the same anycast address all other nodes in the same ˝LL would reach them and their solicitations would never leave the neighborhood (under normal conditions, except when the NDS(LL) server had just recovered).

4.6. Neighbor Unreachability Detection

Neighbor Unreachability detection will work with no changes in behavior other than nodes will get immediate notification when communications with other nodes fail because of a network failure (i.e., a RELEASE on the VC(s) will be received). Nodes can then attempt to re-call the other node using cached link-layer addresses, or the can re-solicit the remote nodes. In the case ˝of re-calling, the other node should be considered unreachable after some number of retries. In the re-solicitation case, the other node is considered unreachable as specified in the Neighbor Discovery spec.

5. Address Configuration

There are three types of address configuration that may be used ˝in IPv6: stateless, stateful, and manual. These methods of address configuration are used to assign IPv6 addresses with scope beyond the local link to IPv6 end-system nodes. If nodes do not require connectivity beyond the local link then only link-local address are needed. In stateless address configuration, nodes receive a list of prefixes which may be used for address configuration in Router Advertisement messages. Nodes then create address by combining the prefixes with the node's host-token. In stateful address configuration nodes are assigned addresses by a central address administration authority via DHCPv6. In manual configuration, nodes are configured with IPv6 addresses in much the same way as IPv4 networks are now. IPv6 over ATM must ˝support all these address configuration methods.

IPv6 nodes can use any combination of address configuration methods.

5.1. Link-Local Addresses

A link-local address is an address which can be used only to ˝reach nodes on the same local link (or LL for ATM). Link-local addresses are never routed and are unique only within the same local link. All nodes have link local addresses. Nodes can create link-local addresses independent of any other network entity. Link-local addresses are completely autonomous.

All nodes within an LL can generate link-local addresses and use them in the normal way. Link advertisements and solicitations ˝for link-local addresses are never relayed outside the LL by the NDS(LL) server. Nodes on one LL cannot reach any node on another LL using link local addresses. These addresses must only be unique within an LL.

A link local address is formed by appending the node's host-token to the link-local address prefix [IPV6-PVCs]. For ATM, a host-token format has not yet been agreed to. This is an area that will require further work and discussion.

5.2. Stateless Address Configuration

Stateless address configuration requires that there be at least one router on the local-link which has been configured with a ˝list of prefixes which may be used by local nodes to create IPv6 addresses. Generally, these prefixes are "loaned" for a specific period of time, after which nodes must stop using the addresses created from the prefixes unless the loan time is extended. In IPv6, the prefix list is carried in Router Advertisement ˝messages. If multiple routers are present on a local-link then the must advertise consistent information, otherwise nodes may not be able to configure addresses and create connections off the ˝local-link.

Router Advertisement messages contain two types of prefixes. One type is marked as available for address configuration, the other is marked as not available for address configuration. Both may ˝be used in determining if an address is "on-link" (i.e., the address is reachable without going through a router).

As described in the section on Router Advertisements, information advertisements send by local routers (routers on the LL) will always be preserved and sent to all nodes by the NDS(LL) server. Thus, stateless address configuration on an LL will continue to work just as it does on legacy LANs. Router Advertisements created by the NDS(LL) servers from external RA messages will contain a prefix list, but these prefixes **MUST NOT** be marked as being valid for address autoconfiguration. They will only be marked as being usable for on-link determination. This will not affect stateless address configuration, but will provide the information used by nodes in determining which prefixes on the tree are considered "on-link".

5.3. Stateful Address Configuration

Stateful address configuration uses an address management authority to assign IPv6 addresses to nodes using a registration protocol [IPV6-DHCP]. This requires that there be a registration

agent on the local-link that either assigns addresses directly, ˝or is capable of contacting the entity which can assign addresses.

To use stateful address configuration, there must be at least one DHCPv6 server or DHCPv6 relay agent on the Logical Link. To use stateful address configuration, a node must first contact the local DHCPv6 agent on the LL. To do this, nodes transmit a message to the DHCP multicast address. Messages broadcast to the DHCPv6 multicast address **MUST** be carried by the NDS tree hierarchy. Nodes which are functioning as DHCPv6 agents **MUST** accept packets received for the DHCPv6 multicast address. All other nodes **MUST** discard these packets.

5.4. Manual Address Configuration

Manual address configuration requires no special processing. ˝When nodes are brought on-line they are manually configured with their IPv6 addresses. These nodes then use the standard IPv6 duplicate address discovery mechanisms to make sure that they have been assigned a unique unicast address.

5.5. Node Relocation

In cases of stateful or manual address configuration, nodes may ˝be assigned an address whose prefix is not valid for the LL to which the node is connected. This is more likely to happen under UNI 4.0 autoconfiguring environments since a change in ATM network topology could change the LL to which a node automatically connects. In manually configured environments, the only cause of node joining the incorrect LL will be configuration errors.

If a node joins an incorrect LL, this can be dealt with in two ways. First, nothing can be done and the node can be allowed to connect to the LL. This case is analogous to the same case in legacy LAN environments. The node will be able to communicate using link-local addresses, but may not be able to communicate using site-local or global address. The second way of handling this is to automatically detect that a node has registered with the wrong LL, and then have the NDS relocate the node to the correct LL.

When a node joins a LL, it must first register with the NDS(N) server to which it connects, and then it must perform the normal ND processes to configure its addresses on the interface. Part ˝of the address configuration process is to perform duplicate address discovery. At the time duplicate address discovery is performed, the node has not yet fully configured the address on the ˝interface (the address is tentative). While the address is tentative, the node **MAY** also require relocation to another LL. Each node **MUST** have a configuration variable which enables and disables the relocation request function. If enabled, relocation request is done at the same time duplicate address discovery takes place.

To request a possible relocation, a node **MUST** send a relocation request message containing its tentative address and the link-layer address of the interface. Only one address **MUST** be included in each message. When an NDS(N) server receives a relocation request message it **MUST** forward it to the NDS(LL) server. When the NDS(LL) server receives a relocation request message it **MUST**

compare the prefix of the address in the message to the list on local prefixes (i.e., those received in local Router ˝Advertisement messages). If a match is found then the node requires no relocation and the relocation request is discarded. If no match is found then the node has been configured with an address which does not belong on the local LL. In this case, the NDS(LL) **MUST** forward the relocation request message up towards the root of the NDS tree. The routing of this message through the NDS tree will follow the same process as the routing of Neighbor Solicitation message. Thus, the message will be directed to the LL on which the node's address would be valid.

When a relocation request message arrives at an NDS(LL) server from the server above, the NDS(LL) server **MUST** compare the prefix of the address in the relocation request message to its list of local prefixes. If no match is found then the NDS(LL) server ˝**MUST** discard the address. If the prefix is determined to be a local prefix the NDS(LL) server **MUST** forward the relocation request message to exactly one NDS server below it (or process the ˝massage its self if there are no lower NDS servers). The NDS(LL) server **SHOULD** distribute the relocation request messages among all NDS server below it to distribute the connection loads on those servers.

As the relocation request travels down the LL subtree from the NDS(LL) server towards an NDS(N) server, all intermediate NDS servers **MUST** send the message to only one NDS server below them. This must be done to distribute the load and assure that the relocation request arrives at only one NDS(N) server on the ˝LL.

When the relocation message arrives at an NDS(N) server, the NDS(N) server **MUST** place a call to the node that sent the relocation request. The relocating node can be called via the link-layer address included in the relocation request message. When the NDS(N) server connects to the relocating node it **MUST** send a relocation reply message consisting of the IPv6 address that the relocating node sent in the relocation request message, along with the link-layer address that the relocating node can ˝use to reach the NDS(N) server to connect and register. The relocating node can use this information to determine to which relocation message the NDS(N) server is replying, and to ˝cache the new NDS(N) server's address in the event the connection to the server is broken. The relocating node **MUST** then initiate the registration process using the PtP VC from the new NDS(N) server as the initial connection to the new LL. The relocating node **MUST** also release all VCs to the original LL. The relocating node **MUST** de-configure any addresses (link-local, stateful and stateless) which it registered when connected to the old LL.

Since nodes can join multiple LLs, nodes **MUST** have virtual interfaces, one for each LL they join. Each virtual interface **MUST** have a unique ATM link-layer address so that each virtual interface can act autonomously. This will make it possible for nodes to join multiple LLs by connecting to a single LL with all virtual interfaces and then permitting those virtual interfaces which must relocate to do so automatically. This will reduce the amount of configuration necessary in both manual and autoconfiguration environments.

Relocation request messages **MUST** be resent three (3) times at one second intervals to make sure at least one reaches the possible target LL. If no response is received after a three tries, the sender can assume it is connected to the correct LL.

5.6. Duplicate Address Detection

All nodes must perform duplicate address detection for all ˝unicast addresses assigned to an interface (or a virtual interface to a ˝LL in the case of ATM). This is done by transmitting special Neighbor Solicitation messages to determine if another node is also using the same unicast address. A unicast address can not ˝be assigned to an interface until its uniqueness has been verified. A unicast address in the process of duplicate address detection, but not assigned to a node, is referred to as a tentative ˝address.

To perform duplicate address detection, a node **MUST** first join the all-nodes multicast group and then the solicited-node multicast address for the tentative address. When a node joins a LL it becomes a leaf of a single PtM connection which is used by the NDS(N) server to "broadcast" to all nodes to which it is serving. There is not a separate PtM connection for all multicast groups. Instead, nodes **MUST** filter incoming packets from the PtM circuit based on destination address to assure that only packets that should be delivered to that node are passed up to the IPv6 input routine. Thus, joining any of the all-nodes, all-routers, or solicited-node multicast groups only requires application of node packet filtering. Packets to all these multicast groups are treated specially by the NDSs but are generally "broadcast" to all nodes within some given scope.

Once the node has enabled the correct filtering for its Solicited Node multicast address it can then perform DAD in the normal manner simply by sending the NS message with the unspecified address in the source address field. The soliciting node will receive its own NS messages, and **MUST** discard these. An indication that a duplicate node exists is indicated by the duplicate node creating a VC back to the soliciting node and replying with an NA message over this VC.

Duplicate Address Discovery NS messages are treated just like any other NS messages by the NDS(LL) and other NDS servers. That is, their flow through the NDS tree will be based on the destination address and target addresses in the NS message.

6. Multicasting

The NDS tree is not intended to provide a general multicasting facility. It is only intended to provide connectionless forwarding of multicast packets, generally for uses where one ˝node is attempting to locate some node or service on the tree.

A general multicast service must be provided so that hosts can directly exchange multicast data across VCs just as they do with unicast data. The NDS tree could be used to handle multicast configuration messages, but not to carry multicast data.

This is an area that still requires some work and will be ˝examined for future revisions of this draft.

7. VC Characteristics

There are two classes of VCs specified in this framework: the VCs used to interconnects elements of the NDS tree (including nodes), and VCs used for data communications between nodes. These two types of VCs should share as many characteristics as possible, particularly packet framing. Unless otherwise specified, framing should be LLC/SNAP encapsulation as specified in IP-ATMND. Optionally, each party of a call **MAY** negotiate null encapsulation.

7.1. NDS VCs

VCs between elements of the NDS tree, including the VCs between nodes and the NDS(N) servers, should all be set up with the same characteristics. Since these VCs are not used to carry data between nodes, they have no QoS or bandwidth requirements.

The characteristics of the PtP VCs interconnecting elements of ˝the NDS tree are:

- Best effort traffic type in UNI 3.X environments, and ABR in UNI 4.0 environments.
- An MTU of 9188 bytes. The MTU **MAY** be negotiated at call setup time but **MUST NOT** be negotiated to a value of less than 1508 bytes (Ethernet MTU plus 8 bytes of LLC/SNAP encapsulation).
- LLC/SNAP encapsulation as specified in IP-ATMND. Optionally, null encapsulation **MAY** be negotiated at call setup time.

The characteristics of the PtM connections which connect NDS servers to the NDS servers or nodes below them are:

- Best effort traffic type in UNI 3.X environments, and ABR in UNI 4.0 environments.
- An MTU of 9188. This **MUST NOT** be negotiated since all leaves of the PtM connection may not be able to handle an MTU negotiated by the root of the PtM VC and the initial leaf. All entities **MUST** handle an MTU of 9188 on this connection.
- LLC/SNAP encapsulation. Encapsulation **MUST NOT** be negotiated since all leaves of the PtM connection may not be able to handle alternate encapsulation formats negotiated by the root of the PtM call and the first leaf. All entities **MUST** handle LLC/SNAP encapsulation.

7.2. Data VCs

The VCs used for data communications between nodes can be set up according to QoS requirements of the application driving the connection. Since the purpose of this framework is to provide an environment where nodes on the same ATM network can directly communicate to utilize ATM QoS features, the use of multiple VCs with different traffic parameters is encouraged. However, the default circuits which are established between nodes (for ˝example, the VC established in response to a Neighbor Solicitation) should have the following characteristics:

- Best effort traffic type in UNI 3.X environments, and ABR in UNI 4.0 environments.

- An MTU of 9188 bytes. The MTU **MAY** be negotiated at call setup time to any value permitted by IPv6. The MTU **MUST NOT** be negotiated to a value less than that supported by IPv6.
- LLC/SNAP encapsulation as specified in IP-ATMND. Optionally, null encapsulation **MAY** be negotiated at call setup time.

Additional VCs can be created as needed with any characteristics as needed as long as both communicating nodes agree on all VC characteristics, and the VC parameters do no violate any IPv6 or ATM specification.

8. Security

The security implications of the NDS tree is still undetermined. This will be part of ongoing work for later revisions of this document.

Acknowledgments

The author wishes to gratefully acknowledge the help and suggestions provided by Markus Jork, Geraldine Harter, Gerd Hoelzing, and Jim Bound.

Authors' Address

Peter A. Schulter
Digital Equipment Corporation
ZK03-3/U14
110 Spit Brook Road
Nashua, NH 03062
Phone: +1 603 881 2920
FAX: +1 603 881 2257
email: schulter@zk3.dec.com

References

[IPV6-SPEC]
S. Deering and R. Hinden, "Internet Protocol Version 6 [IPv6] Specification Internet Draft, June 1995 <draft-ietf-ipngwg-ipv6-spec-02.txt>

[IPV6-ADDR]
R. Hinden, S. Deering, Editors, "IP Version 6 Addressing Architecture" Internet Draft, November 1995 <draft-ietf-ipngwg-ipv6-addr-arch-03.txt>

[IPV6-ADDRCONF]
S. Thomson, T. Narten, "IPv6 Stateless Address Autoconfiguration" Internet Draft, November 1995 <draft-ietf-addrconf-ipv6-auto-05.txt>

[IPV6-ND]
T. Narten, E. Nordmark, and W. A. Simpson, "IPv6 Neighbor Discovery" Internet Draft, February 1, 1996

<draft-ietf-ipngwg-discovery-04.txt>

[IPV6-ETHER]
M. Crawford, "A Method for the transmission of IPv6 Packets over Ethernet Networks", Internet Draft, October 1995 <draft-ietf-ipngwg-ethernet-ntwrks-01.txt>

[IPV6-SA]
R. Atkinson, "Security Architecture for the Internet Protocol" RFC 1825, August 1995

[IPV6-AUTH]
R. Atkinson, "IP Authentication Header" RFC 1826, August 1995

[IPV6-DHCP]
J. Bound, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)" Internet Draft, November 1995 <draft-ietf-dhc-dhcpv6-03.txt>

[IP-FRAME]
R. Cole, D. Shur, "IP Over ATM: A Framework Document" Internet Draft, October 1995 <draft-ietf-ipatm-framework-doc-06.txt>

[IP-ATM]
M. Laubach, "Classical IP and ARP over ATM" RFC 1577, January 1993

[IP-ATMU]
M. Laubach, "Classical IP and ARP over ATM Update (Part Deux)" Internet Draft, August 1995 <draft-ietf-ipatm-classic2-00.txt>

[IP-ATMMC]
G. Armitage, "Support for Multicast over UNI 3.0/3.1 based ATM Networks" Internet Draft, February 9, 1996 <draft-ietf-ipatm-ipmc-11.txt>

[NHRP]
D. Katz, D. Piscitello, B. Cole, J. Luciani, "NBMA Next Hop Resolution Protocol (NHRP)" Internet Draft, November 1995 <draft-ietf-rolc-nhrp-06.txt>

[IP-IPV6ND]
G. Armitage, "IPv6 and Neighbor Discovery over ATM" Internet Draft, August 1995 <draft-ietf-ipatm-ipv6nd-00.txt>

[ATM-UNI30]
ATM Forum, "ATM User Network Interface Specification Version 3.0" ISBN 0-13-225863-3, Prentice Hall, Englewood Cliffs, NJ, 1993

[ATM-UNI31]
ATM Forum, "ATM User Network Interface Specification (UNI) Version 3.1" ISBN 0-13-393828-X, Prentice Hall, Englewood Cliffs, NJ, June 1995

[ATM-UNI40]
ATM Forum, "ATM User-Network Interface (UNI) Signalling Specification Version 4.0", ATM Forum/95-1434R6, December 1995

[ATM-PNNI]
ATM Forum, "PNNI Draft Specification", ATM Forum 94-0471R14

This Internet Draft expires July 15, 1996.