Network Working Group
INTERNET-DRAFT V1

P. Tsuchiya
Bellcore
June 1991

Discovery and Routing over the SMDS Service

Status of this Memo

The difference between this version (1) and the previous version (0), besides the different formatting, is that this version introduces the use of ARP for routers using OSPF to discover the address of other OSPF routers, even though those other routers are not on the same group address.

For now, this memo is an internet-draft, and in fact this version of it is very rough. I'm sure much of the language will need extensive work, especially the musts, shoulds, and mays. In addition, parts of this memo are currently under-specified. There are some relatively complex protocol mechanisms described in this memo, which need extensive critical review. In particular, there are some (hopefully minor) departures from the traditional use of IP addresses. The techniques described in this memo need to be implemented as soon as possible.

Please send comments to the IP Over Large Public Data Networks working group, iplpdn@nri.reston.va.us, or if the comments are particularly humiliating to the author, send them to tsuchiya@thumper.bellcore.com. The above paragraphs of course won't appear in the final RFC. The following paragraph will, but for now please ignore it.

This memo defines a protocol for both intra- and inter-domain discovery and routing over the Switched Multi-megabit Data Service Network. This RFC specifies an IAB standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "IAB Official Protocol Standards" for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Abstract

RFC 1209 [1] describes the encapsulation of IP over the SMDS service generally, and the use of ARP over the SMDS service configured as a Logical IP Subnetwork (LIS). This memo expands on RFC 1209 by describing how to do discovery and routing, both for private (intra-domain) and public (inter-domain) applications, over the SDMS

service. As such, this memo considers cases where a private network is spread over multiple LISs. In particular, this memo allows hosts or routers in different LISs to exchange packets directly--without going through an intervening router.

Introduction

RFC 1209 gives an overview of SMDS. This memo assumes an understanding of the material in RFC 1209.

SMDS can be configured to appear to the user to be both a private, usually multicast network and a public, usually non-multicast network. The former is achieved mainly through the use of group addresses and address filtering. The latter is possible because SMDS uses globally unique addressing at its interfaces.

The purpose of the private/multicast configuration is to emulate to the extent possible a multicast LAN. As a result, SMDS is easily integrated into a LAN-based, private network. Unfortunately, because of the scope of the SMDS service, it is impossible and indeed undesirable to emulate every aspect of a multicast LAN. In particular, the membership of a single SMDS group address, which forms a Logical IP Subnetwork (LIS), must be limited (currently, to 128 members). Therefore, a private network that has more than 128 systems (hosts or routers) to connect to SMDS cannot treat the SMDS service as a single LIS, and must, in some respects, view the SMDS service as multiple LISs connected by routers.

This multiple LIS configuration can result in a path whereby a packet enters and exits the SMDS service more than once. For instance, consider a packet transmission between two hosts X and Y on different LISs. Since the two hosts don't view each other as being on the same LIS, they will use a router that belongs to both LISs to send traffic between them. The SMDS service is crossed twice when strictly speaking it only need be crossed once.

> Note: The extent to which this is a serious problem depends on many factors, such as how often it occurs, and how non-optimal the two-hop path is. For instance, it is much worse if both hosts are on the east coast and the router is on the west coast than if the router is close to one of the hosts. In other words, these multi-hop paths may or may not be acceptable.

A system attached to the SMDS service, therefore, may exchange packets with one or more of the following:

- a private system on one of its LISs--private-local

- a private system not on one of its LISs--private-remote

- a public system (which by definition is not on one of its LISs)--public

RFC 1209 describes how a system can discover the SMDS address of private-local system. This memo describes how systems can discover the SMDS addresses of, and therefore exchange packets directly with, private-remote and public systems. It describes how this should be done both with existing protocol mechanisms and with new protocol mechanisms. In the former case, static configuration is the primary means of discovery. In the latter, a new protocol mechanism, the unsolicited ARP Reply, is used to avoid the burden of static configuration for hosts.

Addressing Considerations

Except for the special case of two routers connected by a point-to-point link, all systems have one or more IP address/subnet mask pairs associated with every network interface.

> Note: This may not be true for pre-subnetting implementations. In this memo we ignore such implementations on the basis that if one is willing to equip a system with an SMDS interface, then one should also be willing to equip it with up-to-date software.

The subnet mask, when applied to the IP address, gives the network/subnet number of the attached network [2]. The following conditions apply to the use of IP network/subnet numbers [3,4]:

1. If the network/subnet number of an IP address matches that of the connected network, then the system with that IP address is directly reachable over the connected network.

2. If the network/subnet number of an IP address (excluding those of neighbor routers that are explicitly configured) does not match that of the connected network, then the system with that IP address can only be reached through a router.

3. Routers do not necessarily require that a neighbor router on the same network share a network number with it. This depends on the routing protocol. Both BGP [5] and certain OSPF [6] configurations (un-numbered point-to-point links and virtual links) do not require a shared network number.

4. Hosts may or may not require that routers reachable over the connected network share a network number with them. RFC 1122 is not

clear on this point [3]. In any event, it seems likely that many hosts will require that even explicitly configured routers share a network/subnet number.

Given the ambiguity of 4, it is safer to assume that all hosts require that routers reachable over the connected network share a network number with them.

With these conditions, routers won't be able to exchange packets with hosts, and hosts won't be able to exchange packets with any system, that does not share a network/subnet number.

However, it is not possible for a system to have one network/subnet number for packet exchanges with both private-local and private-remote systems. This is because the system will use ARP to discover the SMDS address of private-local systems, and will use some other mechanism (static configuration or unsolicited ARP Reply) to discover the SMDS address of private-remote systems. But the only way a system can distinguish between private-local and private-remote systems is by comparing the IP address against two network/subnet numbers, one for private-local and one for private-remote.

This leads to the following requirements for systems attached to the SMDS service.

- An address is considered private-local if ARP is enabled for the network/subnet number that the address matches. Each system must have some local means for determining whether or not ARP is enabled for a network/subnet number. From the perspective of SNMP, however, ARP is considered disabled if there is no ipOverSMDSAddressEntry entry in the SMDS MIB [7] for the IP address associated with that network/subnet number.

- All systems that are members of the same LIS (i.e., are private-local with respect to each other) must share a network/subnet number (this requirement is stated in RFC 1209).

- Further, systems that are not members of a LIS must not have the network/subnet number for that LIS.

- If two systems, one or both of which are not routers, are private-remote or public with respect to each other, and those two systems wish to exchange packets directly, then those two systems must share a network/subnet number. (Below we show that public network/subnet numbers are difficult to form and of potentially limited use.)

A convenient IP network/subnet numbering arrangement for a private

network that spans multiple LISs would be to assign a subnetted part of a class B network number to all systems of the private network that are attached to the SMDS service. Each LIS would be further subnetted. Therefore, each system could have two addresses and masks, as follows:

	address (hex)	mask
private-local:	80.1d.42.c9	ff.ff.ff.80
private-remote:	80.1d.42.ca	ff.ff.f0.00

The private-local mask allows for 126 addresses (excluding -1 and 0), which just about fills up the 128 maximum on group address members. The private-remote mask allows for five bits to distinguish the various LISs, for a total of (2**5) - 2 = 30 LISs. The remaining values of the class B could be used for networks "behind" the SMDS service (LANs and such).

To continue the example, if the system with address 80.1d.42.c9 received an IP packet with destination address 80.1d.48.9e, it would see that the address was not on its LIS:

```
((80.1d.42.c9 & ff.ff.ff.80 = 80.1d.42.80) !=
(80.1d.48.9e & ff.ff.ff.80 = 80.1d.48.80)).
```

It would therefore not ARP for the SMDS address.

If the system with address 80.1d.42.c9 received an IP packet with destination address 80.1d.42.e1, then there is an ambiguity, because this address matches both the private-local and private-remote masks. In this case, the system should know to take the "more specific" match, which is the one where the mask has the most 1's. If it doesn't then the packet may take an extra hop.

This results in the following requirement.

- If a system arranges its addresses so that its private-local network/subnet number is a subnetted portion of its private-remote network/subnet number, or if its private-remote network/subnet number is a subnetted portion of its public network/subnet number, then it should match on the most specific mask.

- If the system is not capable of picking the most specific match, then its private-local, private-remote, and public network/subnet numbers should not overlap.

Note: The above addressing arrangement resulted in the system having two separate IP addresses (for its two logical

interfaces) even though it only has one physical interface. While it would of course be possible to build a system that could handle having one IP address with multiple logical interfaces with different masks, this seems to be too much of a departure from IP address fundamentals. Therefore, systems configured with multiple logical interfaces must have multiple IP addresses. Fortunately, most systems should have no more than two such addresses.

Forming public network/subnet numbers is problematic, and potentially not very useful. To form a public network/subnet number, a large network number, almost certainly a class A, is needed. This number would be assigned to all systems attached to the SMDS service, thus increasing the complexity and overhead of all systems.

But the only case where the public network/subnet number is needed is where one or both of the systems communicating are hosts (because routers are able to directly exchange packets without sharing a network/subnet number). Host to host or host to router public packet exchange is likely to be the least common type of packet exchange over an SMDS network. Router-to-router packet exchange, both public and private, should be much more common, and directly attached hosts will more likely exchange packets privately than publicly.

And, a host can directly exchange packets publicly without a public network/subnet number by configuring itself as a router and running a scaled down version of BGP (this is discussed later). Or, a host can always send public packets by going through one of its private routers, thus suffering multi-hops.

For the above reasons, it seems unnecessary to have a public network/subnet number.

Router Configurations

Before discussing the mechanisms of discovery and routing over SMDS, we discuss some issues concerning router configuration.

Any two routers that have routing table entries for each other and can forward packets to each other are called neighbors. Depending on the routing protocol, neighbor routers may or may not exchange routing updates. For instance, with OSPF, because of designated routers, it is possible for two routers to learn of each other and forward IP packets to each other without ever directly exchanging routing updates. The IP address of neighbor routers is learned from the designated routers in OSPF packets, and the SMDS address of neighbor routers is learned from the designated routers acting as ARP servers (see section "Router Operation"). With other protocols, such as RIP

[9], two routers can only be neighbors if they exchange routing updates directly.

A private domain will have some number R of routers connected to the SMDS service. If all R routers are neighbors of each other (each router has R-1 neighbors), then a packet will almost never traverse more than two routers. The worst-case multi-hop would be three: host-router, router-router, router-host. We call this configuration of routers the all-neighbors configuration.

The partial-neighbors configuration, then, is one where not all routers in a private domain are neighbors of each other. With the partial-neighbors configuration, a packet may take any number of hops across the SMDS service, depending on how sparse the neighbor connectivity is. If there are 5 routers, A, B, C, D, and E, and the neighbor relationships are A-B, B-C, C-D, and D-E, then a packet entering at A and exiting at E will cross the SMDS service four times.

The advantage of the all-neighbors configuration is shorter paths across the SMDS service, and this memo recommends it whenever possible. Depending on the routing protocol used, and the group address configuration, the disadvantage may be in increased routing traffic over the SMDS service. Another disadvantage of the all-neighbors configuration is in the amount of state each router has to keep, but this is unlikely to be a problem except perhaps for extremely large configurations (say many hundreds of routers directly attached to the SMDS service). In some cases (discussed in section ROUTING PROTOCOL OPERATION), the amount of configuration needed to maintain an all-neighbors configuration may be prohibitive.

Because common network/subnet addresses are not necessarily available for public systems, it is necessary to find a means of discovering SMDS addresses purely in the context of the public routing protocol, which is BGP. Since there will be many hundreds or thousands of BGP routers on the SMDS service, it is critical that BGP can be operated with minimal configuration, traffic, or memory overhead.

This memo defines three modes for BGP operation:

Mode 1: Two BGP peers exchange BGP information directly

Mode 2: Two BGP peers exchange BGP information via a "BGP server"

  Mode 2a: Full router--the BGP router maintains complete BGP information

  Mode 2b: Partial router--the BGP router maintains only

what it needs

The purpose of Mode 2 is to minimize configuration, traffic, and memory overhead when there are a large number of BGP routers connected.

With BGP servers, BGP routers need only configure a handful of BGP servers, not every other BGP router. The BGP servers, then, act as a distribution point for BGP routing information. (Note that a BGP server runs standard BGP. It is a server by virtue of its configuration parameters.)

While a full router must receive and store roughly the same amount of information whether it peers with a BGP server or directly with every other BGP router, the KEEPALIVE traffic is substantially reduced with BPG servers. For instance, 5000 domains and a KEEPALIVE of 5 minutes results in an average of 16 KEEPALIVEs per second per router.

Moreover, BGP servers allow for a "partial router" (Mode 2b). This is a router that maintains partial or no permanent routing information. Instead, the partial router sends its IP packets to a BGP server, which forwards the packet appropriately, and sends the partial router "on-demand" BGP Update information for only the destination in the IP packet.

Finally, the use of BGP servers eases the configuration problem. There is no automatic way to configure BGP peers. Therefore, the more BGP peers a BGP router has, the more manual configuration necessary.

A BGP router can mix Modes 1 and 2. In other words, it can peer directly with some BGP routers, but otherwise receive its information from BGP servers.

Routing and Discovery over SMDS

All hosts and routers have an IP-to-physical address translation table. (For the purposes of this memo, the physical address corresponds to the SMDS address.) For a system to send a packet directly to another system, it must be able to translate the IP address to a physical address, either by indexing the table or through an algorithmic manipulation of the IP address. (The latter does not apply to SMDS.)

Hosts can learn IP-to-physical address translations by only one of two ways: static configuration of the IP-to-physical address translation table, or reception of an ARP Reply. Routers can learn IP-to-physical address translations in the same two ways as the hosts, plus

via the BGP attribute NEXT_HOP_SNPA [8] (the latter applies mainly to public internetworking).

While it is always possible to avoid multi-hops by staticly configuring the IP-to-physical address translation table, it is preferable to do so automatically via the reception of ARP replies for hosts, or BGP Updates for routers. The NEXT_HOP_SNPA information in the BGP Update is adequate for conveying SMDS addresses for public internetworking. Since ARP requests cannot be sent for systems that are private-remote, we define a new mechanism for learning SMDS addresses, which is the Unsolicited ARP Reply (UARP Reply).

- The reception of a UARP Reply is handled exactly the same as the reception of an (requested) ARP Reply.

- Hosts must never send UARP Replies.

- When a router F receives an IP packet P from a system S over its SMDS interface for which the next hop system N on the path to the destination is back over the SMDS interface, the router F forwards the IP packet P to N, and may send S a UARP Reply or a "on-demand" BGP UPDATE depending on the following.

  The router F searches its routing databases for a neighbor whose SMDS address matches the source address of the received packet P. The address may match nothing, in which the packet will have been received from a host, or the address may match an entry for a BGP (public) neighbor, or an entry for a private neighbor.

  If the source SMDS address in packet P matches nothing, then an ICMP Redirect followed immediately by a UARP Reply is sent. The UARP Reply contains the SMDS address of the next hop system N (in ar$sha), and the IP address of the next hop system N given in the Redirect (in ar$spa). The source IP address in the IP header of the UARP Reply should contain the IP address of F. The destination IP address in the IP header of the UARP Reply contains the source IP address of the received packet P.

  Otherwise, if the source SMDS address in packet P matches that of a private router neighbor R, and the next hop system N is not a router neighbor of F (which is determined by comparing the SMDS address of the next hop system with those of the router neighbors), and the next hop system N is either private-local or private-remote, then a UARP Reply is sent to R. The packet fields arsha, arspa, and the source IP address are set as in the previous paragraph. However, the destination IP address in the IP header of the UARP Reply contains the IP address of router neighbor R. Note that router neighbor R might also be a BGP neighbor.

However, since R is private, it would be an internal BGP neighbor, and will have therefore already received all of the BGP information that F has (internal BGP neighbors are always full routers).

Otherwise, if the source SMDS address in packet P matches that of a BGP neighbor R, then a BGP Update is sent to R. It contains the IP address of the next hop router F in the NEXT_HOP attribute, the SMDS address of F in the NEXT_HOP_SNPA attribute, and the network of the destination address in packet P must be one of the networks listed in the BGP Update. Note that F should not have received packet P if BGP neighbor R is a full router. F may therefore wish to check to make sure that R is a partial router, and if not, to report an error to system management.

The reason for sending the ICMP Redirect in the case of sending a UARP Reply to a host is to give the host an IP number to relate the UARP Reply with. If the IP address of destination of packet P is not on the SMDS service, the the host will not recognize that it can reach the destination directly and may not accept the UARP Reply. With the ICMP Redirect, the host knows that it is routing to a router on the attached network.

With the technique of UARP Replies, if either the SMDS entry or exit system is a host, then a direct path will be found across the SMDS service even if the partial-router configuration is used. However, if both the SMDS entry and exit systems are routers, and a multi-hop path is found by routing, then that path will persist. The reason for this is that it just doesn't work to have a router try to redirect another router to still another router. This memo doesn't go into detail about this except to say that the IP architecture is such that routers fundamentally expect to know everything they need to know from the beginning, and getting them to cache things on the fly generally mucks things up. Only by putting limitations on the spreading of BGP routing information, can we get away with "on-demand" BGP updates for partial routers.

- o A router must have some mechanism to prevent its sending an excessive number of the same UARP Replies or BGP Updates. This might happen if the system receiving the UARP Replies or BGP Updates did not honor them, for instance in the case of UARP Replies because its mask was not configured correctly.

One such algorithm would be to establish three variables associated with a particular UARP Reply or BGP Update; arpEnabled, arpRate and arpPersistance. When the "first" UARP Reply or BGP Update is sent, create the three variables, and set arpEnabled to ON, set arpRate to some constant, say 20, and set arpPersistance to some other constant, say 5. Each time a packet is received that should result in sending

the UARP Reply or BGP Update, check arpEnabled. If it is OFF, then do nothing (that is, don't send the UARP Reply or BGP Update). If arpEnabled is ON, decrement arpRate. If arpRate does not decrement to 0, then do nothing. If arpRate decrements to 0, then send a UARP Reply (preceded by the ICMP Redirect if necessary) or BGP Update, and decrement arpPersistance. If arpPersistance does not decrement to 0, reset arpRate to its constant (20). If arpPersistance does decrement to 0, then set arpEnabled to OFF. After some timeout period, destroy the three variables (so that another identical UARP Reply or BGP Update will be considered the "first" one). This algorithm has the effect of constraining the rate at which UARP Replies or BGP Updates will be sent, and of giving up on sending them for a period of time if the recipient seems to be ignoring them.

- Routers and Hosts must time-out the information learned from UARP Replies and on-demand BGP Updates, just as they do for ARP Replies. Routers and Hosts should refresh the time-out period upon reception of a packet with an SMDS source address and IP source address matching the information in the UARP Reply or BGP Update. Note that in the case of the BGP Update, the source IP address will be compared against a masked IP address.

Routing Protocol Operations

In what follows, we discuss the operation of specific routing protocols over SMDS. In some cases, the routing protocol takes advantage of multicasting over SMDS.

- In such cases, the routing protocol must use the same group address as that defined for sending ARP requests. In the SMDS MIB [7], this is the object-type smdsARPReq, which is a member of ipOverSMDSAddressEntry.

We assume that the reader is familier with the protocols discussed.

OSPF and RIP All-Neighbors Configurations:

OSPF and RIP can operate both in multicast and non-multicast modes. Multicast is preferable because it requires less configuration.

If there are less than 128 routers in a private network attached to the SMDS service, then those routers can form a single LIS (group address) that includes just themselves. We call this the router LIS. They can multicast OSPF or RIP packets over the LIS as they would over a multicast LAN. The only configuration necessary is that of the addresses (IP, SMDS group, and SMDS single). Designated routers are elected in order to reduce overhead.

If there are also hosts attached to the SMDS service, and the total number of hosts and routers exceeds 128, or for some other reason all hosts and routers cannot join the same LIS (for instance because group addressing is not yet available over LATA boundaries), then LISs in addition to the router LIS must be formed for the hosts. Each of these LISs must have at least one router as a member. Note that this configuration essentially forms a 2-level hierarchy of LISs. The "core" LIS is the router LIS. The "leaf" LISs are the host LISs, and attach to the core LIS by virtue of routers that belong to both LISs. One "level" of UARP Replies are required to allow two hosts on different LISs to exchange packets directly.

OSPF Designated Router Operation:

In some cases, it may not be possible to put all routers on the same LIS. Using the designated router election feature of OSPF, it is still possible to get an all-neighbors configuration of routers without requiring N**2 configuration of routers.

The operation of designated router election is as follows. Some or all routers are configured as eligible. This means that they may become designated routers. Some or no routers are configured as ineligible. The eligible routers have a means of sending OSPF packets to all other routers (either using ARP or by static configuration). The ineligible routers do not need to be configured with information on how to reach any other routers.

The eligible routers establish each others as neighbors. Of these, one is chosen as the designated. The designated router then becomes neighbors with all routers, forming a star configuration. The designated router then tells all routers of all other routers in its link state advertisements.

At this point, the ineligible routers know the IP addresses of all other routers, but not the SMDS addresses. Therefore, when a packet arrives that must be routed to another router, the SMDS address for the other router must be learned. This is done by sending ARP Requests to the designated router. To make this work, the following is required:

- A separate network/subnet number is required for all routers. This network/subnet number must be distinguishable from private-remote network/subnet numbers. This is because the private-remote systems are marked as not ARP-able, whereas other routers can be ARPed for (ineligible routers only) by virtue of the designated router.

- When an ineligible router becomes neighbors with the designated router, it must install the SMDS address of the designated router as the ARP address for the network/subnet number representing all routers (smdsARPReq in the SMDS MIB).

- Eligible routers must be able to respond to ARP Requests from neighbor routers about neighbor routers.

OSPF and RIP Partial-Neighbors Configurations:

The following configurations are not recommended for OSPF, in lieu of all-neighbors configuration using designated router election. They may be necessary for RIP, however.

Even if all of the routers of a private network cannot join a single LIS, it is still possible to have automatic configuration. This can be done by forming multiple router LISs, where some number of routers on each router LIS belong to more than one router LIS (multi-homed router) in such a way that a connected graph is formed. By connected, we mean that there is a path from any router LIS to any other router LIS through a series of zero or more LISs connected by multi-homed routers. The number of "levels" of UARP Replies is equal to the diameter of the graph formed by multi-homed routers (as nodes) and LISs (as links). The only configuration necessary is that of the addresses (IP, SMDS group, and SMDS single). Within each LIS, designated routers are elected in order to reduce overhead (OSPF only).

Especially in the early stages of SMDS deployment, there may be cases where two LISs cannot be joined by a multi-homed router. In this case, routers in each LIS must configure a logical point-to-point link with each other. In the worst case, there may be no LISs at all (for instance, because inter-LATA group addressing is not yet available, and there is one router in each LIS). Even in this case, however, logical point-to-point configuration with all other routers can be avoided. This can be done by configuring each router with logical links to a subset of the other routers such that the resulting graph is connected.

Note also that the above router operation applies to any routing protocol that can broadcast its routing updates.

BGP:

Mode 1 BGP Router Operation:

Operation of a BGP Router in Mode 1 is straight-forward. External

BGP is used, and is operated as normal [5], using IP encapsulation as described in [1]. The IP and SMDS addresses of the BGP peer are manually configured, and are obtained via means not specified by this memo.

Mode 2a Operation:

A BGP Router peers with a BGP server exactly as it would with another BGP router (i.e., Mode 1 operation). The difference between Mode 1 and Mode 2 BGP router operation is in how BPG peers are established. In Mode 2, a BGP router must keep a list of BGP servers. Since substantially the same information will be received from all of them, it is only necessary to peer with one BGP server at a time, or two if a hot backup is desired. Therefore, a Mode 2a (and Mode 2b) BGP router needs the ability to choose active peers from its list of BGP servers.

Mode 2b Operation:

As with the Mode 2a (full) BGP router, a Mode 2b (partial) BGP router must keep a list of BGP servers, and must have an algorithm for choosing active BGP servers. The partial BGP router must additionally treat its active BGP server(s) as its default route. In other words, the partial BGP router will send any packets that it doesn't have explicit routing information for to a BGP server. Marking the BGP server as a default is a matter of local configuration. That is, the BGP server will not send the BGP router any indication that it is a default router. Also, the partial router must be viewed as a default router by systems "behind" the BGP router (in the subscriber's network). The partial router must not send routing information it learns from a BGP server to any other routers.

The partial router can elect to either receive all BGP information from the BGP servers and choose not to keep it, or the BGP server can be configured (locally) to not send the partial BGP router any routing information at all (except of course for the "on-demand" BGP Updates already described). In any event, the BGP router must send the BGP server its own BGP routing updates. This way, the BGP server can further distribute it to other BGP routers (and servers).

BGP Server Operation:

Much of the operation of BGP servers is given in the previous paragraphs. In this section, the exchange of BGP information between BGP servers is discussed.

There will be more than one BGP server. The reasons are both to spread the load over multiple servers, and to provide backup servers.

Every BGP server is expected to have knowledge of all destinations advertised to all BGP servers. Therefore, the BGP servers must exchange BGP information with each other.

To do this, the BGP servers should behave as though they all belong to a single autonomous system. (Strictly speaking, an SMDS service is not an autonomous system, because IP packets can transit the SMDS service without going through a router). That is, they should use external BGP to exchange information with BGP routers, and use internal BGP to exchange information with each other. This means that every BGP server maintains a peer relationship with every other.

> Note: This requires N**2 BGP server internal relationships. For the near term, and perhaps even long term, I don't think this will be a problem. I think even several hundred BGP servers could be handled.

BGP servers are configured to pass the NEXT_HOP and SNPA_NEXT_HOP fields untouched, both when they advertise updates internally and externally. When BGP servers advertise updates externally, they should append an AS number representing the SMDS service to the AS_PATH.

General Discussion

While I have taken my best shot at coming up with clean and efficient solutions to the problems of discovery and routing over SMDS, there are several possible options to the techniques discussed in this memo that should be considered. This discussion is not meant to be included in the final RFC.

The UARP Reply is used as a redirect mechanism mainly because it contains the hardware address. It might be better to use a whole new ICMP message to convey this information. The new message would contain the same information as the UARP Reply, but would have a different ICMP message number, and therefore would be distinguishable from the ARP Reply.

There is an interesting mechanism for discovery over SMDS that I considered but chose not to incorporate. With SMDS, it is possible for a system to send messages to group addresses that it is not a member of. This means that if a system were configured with a list of the group addresses for each LIS on its private network, it could ARP for things not on its own LIS.

I decided not to do this for several reasons. First, it didn't eliminate the need for the UARP Reply mechanism, because initially one may not be able to do group addressing across LATA boundaries.

Second, the idea of non-symmetric ARP groups makes me very uncomfortable. For instance, it seems that it would require that each system have a logical interface and associated IP address and mask for each LIS that it might ARP on. But these addresses would be unusable for sending and receiving packets, and in fact things would break if those addresses were known outside the system that owned them. Either that, or it would be necessary to modify the systems so that they could ARP over something that they could not match up against one of their interfaces' network/subnet numbers. But this again is drifting too far from the fundamental meaning of IP addresses for my comfort.

In general I am not completely comfortable with the material in this memo. To me, there are too many kludges in it--kludging addresses over logical interfaces so that a system knows that something else is reachable over the network, and kludging the ARP Reply so that a system can efficiently learn the SMDS addresses of other systems. Another approach to this whole problem would be to create an SMDS-wide ARP service, or at least an ARP service that could handle all ARPing for a private network. However, this solution required a whole new distributed algorithm for the purposes of collecting and disseminating the ARP requests and replies. Since the routing algorithm already has most of the information needed at hand, it seems overly expensive to create a new algorithm to handle ARPing. Also, many of the addressing weirdness didn't seem to get completely resolved even with an ARP service (unless the use of ARPing over group addresses was limited to ARP servers only).

REFERENCES

[1] Piscitello, D., Lawrence, J., "The Transmission of IP Datagrams over the SMDS Service", RFC 1209, USC/Information Sciences Institute, March 1991.

[2] Mogul, J., Postel, J., "Internet Standard Subnetting Procedure", RFC-950, USC/Information Sciences Institute, August, 1985.

[3] Braden, R.T.,ed., "Requirements for Internet hosts - communication layers", RFC-1122, USC/Information Sciences Institute, October, 1989.

[4] Braden, R.T., Postel, J.B., "Requirements for Internet gateways",

RFC-1009, USC/Information Sciences Institute, June, 1987.

[5] Lougheed, K.; Rekhter, Y., "A Border Gateway Protocol 3 (BGP-3)", Internet-draft, January 1991.

[6] Moy, J., "OSPF specification", RFC-1131, USC/Information Sciences Institute, October, 1989.

[7] Tesink, K. ed., "Definitions of Managed Objects for the SIP Interface Type", Internet-draft, March 1991.

[8] Tsuchiya, P.T., "Border Gateway Protocol NEXT_HOP_SNPA Attribute", Internet-draft, March 1991.

[9] Hedrick, C., "Routing Information Protocol", RFC-1058, USC/Information Sciences Institute, June, 1988.