TUNNELING

A commonly encountered problem in internetworking is that of routing packets from source to destination where both hosts are on the same type of network. E.g. TCP/IP-based Ethernet, but the packets must traverse a non-IP WAN e.g. ATM or X.25. The solution to this general problem involves the use of tunneling (see fig 5-47)-\*-. This technique requires the use of MP routers (at each end of the tunnel) which treat the IP packet as the payload of a non-IP WAN packet, during the time that the packet is transmitted across the WAN. Only the sending and receiving MP routers need to understand both the IP and WAN-specific network protocols; the hosts need not.

CONNECTION-LESS DELIVERY OF IP PACKETS (MTU, Fragmentation, Reassembly)

Although the maximum permissible size of IP packet is 65,536 bytes, in practice, this is much larger than the size of a datalink layer frame, within which it must be encapsulated for transportation across the physical network (e.g. Ethernet). Each packet switching technology places a fixed upper bound on the amount of data that can be transferred in a physical frame (e.g. 1500 bytes for Ethernet & 4470 bytes for FDDI), and is often called the MTU (Maximum Transfer/Transmitting Unit). The solution used for transmitting large IP packets using underlying datalink layer protocols with smaller MTU’s is a technique called Fragmentation. This involves dividing an IP packet into smaller pieces whenever they must traverse a network that cannot handle the original packet size. Each piece (fragment) has the same format as a packet; fields in the IP header specify that a packet is actually a fragment and provide the offset of the fragment’s data in the original packet. The size of each fragment is a multiple of 8 bytes and the value of the offset ranges from 0 to 8191, expressed in multiples of 8 bytes. The IP protocol does not limit packets to a small size (to avoid fragmentation), nor does it guarantee that large packets will be delivered without fragmentation. The sending host, at application layer, can choose any packet size it considers appropriate; fragmentation of the packet and its reassembly occur automatically, transparent to the sender and receiver. Fragments must be reassembled to produce a copy of the original packet before handling it over to the destination host’s application layer.

In general routers and hosts must handle fragments or packets of at least 576 bytes without need for further fragmentation.

Example:

Host B

Host A

Network 3

Network 1

MTU = 1500

MTU = 1500

internetcloud1

MTU = 620

Network 2

R2

R1

R1 fragments packets A → B; reassembled by R2

R2 fragments packets B → A; reassembled by R1

DATA 2 (600 bytes)

DATA 3 (200 bytes)

DATA 1 (600 bytes)

Original

PACKET HEADER

MF=1, Offset = 0/8 = 0

20

DATA 1 (600 bytes)

FRAGMENT 1 HEADER (20)

What is sent

MF=1, Offset = 600/8 = 75

20

DATA 2 (600 bytes)

FRAGMENT 2 HEADER (20)

MF=0, Offset = 1200/8 = 150

DATA 3 (200 bytes)

FRAGMENT 3 HEADER (20)

Note: MF = 0 and offset = 0 → Not a Fragment (Full Packet)

Re-assembly Issues:

In a TCP/IP internet, once a packet has been fragmented, the fragments travel as separate packets all the way to their final destination (host) where they are reassembled fig 5-20 -\*-). This approach has two disadvantages:

1. Even if the physical networks have larger MTU, only small fragments traverse them; this leads to inefficient use of some intervening physical networks.
2. Reassembly of the packet at the final destination, instead of the intervening routers, requires that all fragments successfully traverse all intervening physical networks, within a specified (by the destination host) duration of time. Since loss of a single fragment results in the loss of the entire packet, this approach increases the probability of packet loss.

However, the approach has two major advantages:

* 1. Each fragment can be routed independently right up to its final destination – faster then reassembling and re-fragmenting many times.
  2. Intervening routers need not have storage for accumulating and reassembling fragments – simplifies router hardware and software.

INTERNET CONTROL PROTOCOLS

These protocols are used in the TCP/IP network layer in addition to IP which only handles data transfer.

ICMP (Internet Control Message Protocol)

This protocol is used by routers and hosts to communicate error and control information to other routers and hosts. ICMP is a required part of TCP/IP and must be included in every TCP/IP implementation (stack). Whenever a router cannot route a packet because: the router does not have a route to the destination, the destination host is disconnected from the network, the TTL counter expires, the intermediate routers are congested, or, the packet header is incorrect – the router sends an ICMP message to the host where the packet originated. Intermediate routers are not informed. The originating host must relate the error to an individual program and take corrective action; this portion is not governed by ICMP. Routers also use ICMP to request other routers for their subnet masks, or, to advertise their presence (periodically) on an attached LAN. Hosts use ICMP to solicit “advertisements” from the routers connected to their LAN by broadcasting an ICMP message. This enables a host to find out the IP address of routers which it can use right after booting.

Hosts or routers may also use ICMP to verify end-to-end connectivity (i.e. IP and ICMP are running on source and destination, and intermediate routers have correct routes) by sending ICMP messages of Echo-Request/Reply type (a.k.a PING), or, to measure network performance. In order to handle a variety of message types, the ICMP message has an ICMP header containing a TYPE field which indicates the type of message (see fig 5-61). Following the header, the message (optionally) has the IP header and first 64 bits of data copied from the IP packet that caused the error to be reported. The entire ICMP message is itself transported within the data portion of an IP packet whose Protocol field is set to ICMP (value = 1). The routing of ICMP messages is thus handled just like IP packets carrying data. IP software on a host, receiving an IP packet, with Protocol = ICMP, hands it up to ICMP-handling software, instead of data-handling software, instead of data-handling (TCP, UDP) software. However, if a router fails to deliver IP packets containing ICMP messages (e.g. due to congestion), it does not respond by sending new ICMP messages to the sender (to avoid further congestion).

ARP (Address Resolution Protocol)

Every machine (host, router) which has an IP address assigned to it must also have a Datalink layer (Ethernet, FDDI) address because an IP packet must eventually be encapsulated within a datalink frame before it can be transmitted and received. Datalink addresses are pre-assigned (uniquely) by the interface-card manufacturer, while IP addresses are (uniquely) assigned by Network Administrator, based on the network class, using an organization’s internal conventions.

A mechanism is needed for a machine to find out the Datalink (DL) address of another, based upon its IP address. ARP (and its variant Proxy ARP) provides this capability by broadcasting the IP address and getting the mapping as a reply from the machine with the target IP address. ARP eliminates the need to permanently store IP ↔ DL address mappings at routers, or a central database by providing a dynamic address binding technique. Whenever a host uses an IP address for the first time, it broadcasts an ARP message with this IP address, the destination host responds with its mapping which is used and saved by the sending host. Each host maintains an ARP cache containing mappings for the IP addresses it has used in the past and any others of which it has been informed. At boot time, a host broadcasts its own IP/DL address pair within an RP message so that all existing hosts on the LAN can save it in their cache. Thus, a host has the mappings for hosts that connected to the LAN after it, and, for hosts to which it sent packets in the past.

Sending ARP messages to hosts belonging to a remote LAN requires participation of the local router. (Self study 450-452 -\*-).

An ARP message is carried within the data portion of a DL frame, with the type filed in the frame header indicating an ARP message instead of IP packet. Thus, ARP is considered part of the physical network system, not the TCP/IP protocol set.

RARP (Reverse Address Resolution Protocol)

RARP enables a host that has a DL address (Ethernet) to obtain its IP address. Machines that have attached disks store their assigned IP addresses locally. This address is available to the binary OS-image when the machine boots. For diskless machines, RARP provides a mechanism to get their IP address before they boot, so that they can use standard TCP/IP FTP to obtain their initial boot image from a shared file server on the network. Such machines broadcast a RARP-request message, embedded within a DL frame’s data area, on their LAN. The format of a RARP message reassembles, and the requesting machine’s DL address is provided in the message instead of any (source, target) IP addresses. At least one machine on the LAN functions as a RARP server, maintains DL ↔ IP address mappings, and responds to DL broadcast frames containing RARP-request message. This reply contains the IP address of the requesting host, directed at the requesting host’s DL address. If multiple RARP servers respond, the host uses the first response.

Since RARP broadcasts are not forwarded by routers, a dedicated RARP server is needed for each LAN. Newer mechanisms for getting DL → IP address mappings (manually prefixed or dynamically assigned), such as BOOTP and DHCP, do not rely solely on local broadcast techniques. Hence they do not require the server to be on the same LAN as the requesting host. Like ARP, RARP is considered part of the physical network system, as it does not use the TCP/IP protocol stack.

RARP

ARP

Directed Reply (local IP)

Broadcast Request (local DL)

Broadcast Request (remote IP)

Directed Reply (remote DL)

DL

DL

RARP SERVER

HOST

REMOTE HOST

LOCAL HOST

IP

IP

ROUTING ALGORITHMS

In a large internet consisting of several ASes, two protocols are used; OSPF (Open Shortest Path First, an interior gateway protocol) for routing within an AS, and, BGP (Border Gateway Protocol, an exterior gateway protocol) for routing across ASes.

OSPF

This protocol is a specific implementation of the Link State Routing technique (see 5.2.5 for details) and was standardized (1990) as a replacement for RIP (Routing Information Protocol). Universally supported by all routers, OSPF is a public routing protocol, which supports adaptive dynamic routing, handles multiple service types (realtime, non-realtime), exploits and uses alternate routes, enables hierarchical routing and provides basic security mechanisms. OSPF functions by abstracting the actual network (routers and lines) into a directed graph, wherein the weight of an edge reflects distance and delay attributes of the connection. It then computes the shortest path between routers. To make these computations manageable and limited to smaller subsets of routers, OSPF divides each AS into numbered areas, which are similar to WAN-subnets in the sense that the detailed topology of an area is invisible outside it. An area interfaces with other areas via an Area Border router. One specific area (number 0) is designated a Backbone Area and has Backbone routers which provide routing across other areas within the AS. Each AS designates one (or more) routers in its Backbone Area to serve as the AS Boundary or Border router that enables routing between this and other ASes, using BGP.

An Internet

Area Border Router

area 1

area 0

BGP

area 0

area 1

AS1

area 2

AS2

Internal Router

AS Boundary Router

area 0

area 1

Backbone Router

AS3

OSPF identifies five types of messages which are exchanged between routers within an AS (inside an area and across areas, i.e. intra-area and inter-area). See fig 5-46 5-66 and -\*-. And associated text for details.

BGP

This protocol handles inter-AS traffic by routing packets between ASes and allows enforcement of many types of routing policies, which may stem from political, economic or security considerations. Such routing policies, which are not a part of BGP, are manually configured into each router. From BGP’s perspective, an internet is viewed as a collection of three types of interconnected ASes: stubs ASes having a single connection to the BGP network, multi-connected ASes that disallow pass-through packets, and, transit ASes that handle pass-through third-party packets at a cost. To ensure reliable exchange of routing information and to conserve capacity, BGP exchange this information in an incremental manner (using deltas) and by using a reliable underlying protocol i.e. TCP.

OSPF

OSPF

BGP

R1

R2

OSPF

OSPF

AS2

AS1

R1 and R2 are BGP peers that use BGP to advertise networks in their ASes after collecting their information from other routers internally. Organizations usually designate routers at the outer edge of their AS as BGP routers. BGP uses a blend of ideas from distance-vector-protocol and link-state-protocol techniques. Currently, BGP-4 is widely used in TCP/IP internets.