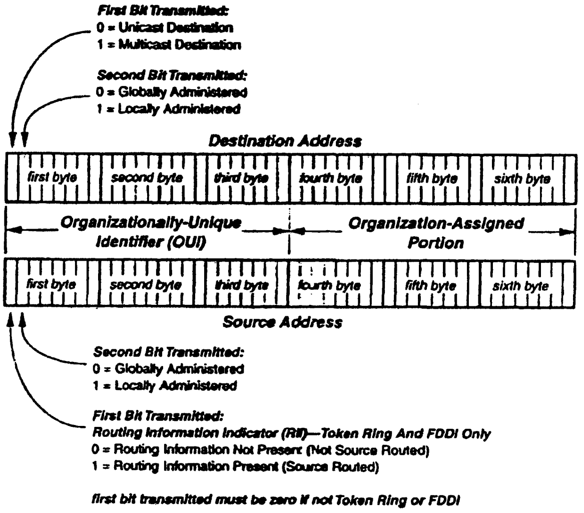
4B15 05 cribs

Q1 a) The MAC address can identify both the transmitting station (*source address*) and the receiving station (*destination address*) and are usually referred to as the *MAC address* of the station on the LAN.

The 48bitMAC address space is split into two halves:

- A *unicast address* identifies a single device or network interface. When frames are sent to a single station, the unicast address is used as the destination address. The source address is always unicast. Often referred to as *individual, physical* or *hardware addresses*.
- A *multicast address* identifies a group of logically related devices. They provide a means of one-to-many communication allowing multiple destinations to be addressed by a single communication. Often referred to as *group* or *logical addresses*.


First Bit Transmitted:
0 = Unicast Destination
1 = Multicast Destination
Second Bit Transmitted:
0 = Globally Administered
1 = Locally Administered
Destination Address
first byte
second byte
third byte
fourth byte
fifth byte
sixth byte
Organizationally-Unique Identifier (OUI)
Organization-Assigned Portion
first byte
second byte
third byte
fourth byte
fifth byte
sixth byte
Source Address
Second Bit Transmitted:
0 = Globally Administered
1 = Locally Administered
First Bit Transmitted:
Routing Information Indicator (RII)—Token Ring And FDDI Only
0 = Routing Information Not Present (Not Source Routed)
1 = Routing Information Present (Source Routed)
first bit transmitted must be zero if not Token Ring or FDDI

The first bit of the destination address defines if it is a unicast (0) or a multicast (1) address. Source addresses are always unicast so the first bit is zero except in token ring or FDDI, where is describes how the packet is routed. The second bit of the address defines whether the address is globally unique (0) or locally unique (1) to the LAN. Globally unique addresses are assigned by the manufacturer, whereas locally unique addresses are assigned by the LAN administrator, carefully! The 48-bit MAC address is divided into two parts. The first 24 bits constitute the organisationally unique identifier (OUI), which indicates which organisation (typically the manufacturer) is responsible for assigning the remaining 24 bits of the address. A MAC address is usually expressed in canonical byte form *aa-bb-cc-dd-ee-ff* where the first 3 pairs of bytes are the OUI and the last 3 pairs of bytes are set by the manufacturer.

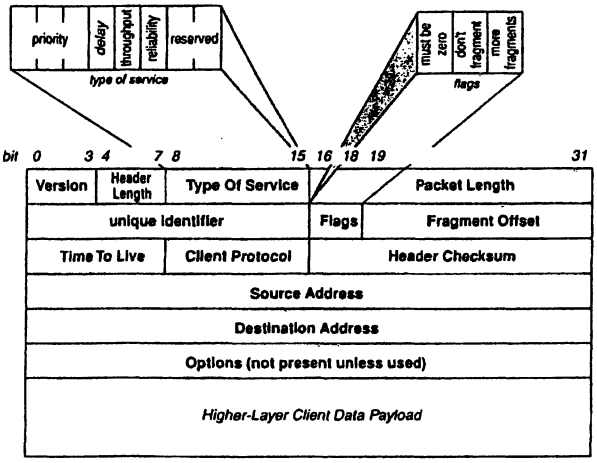
The MAC address has become a vital part of the popularity of LAN protocols and their extension into MANs and WANs. The fact that they are globally unique means that a piece of hardware can be identified uniquely anywhere in the world.

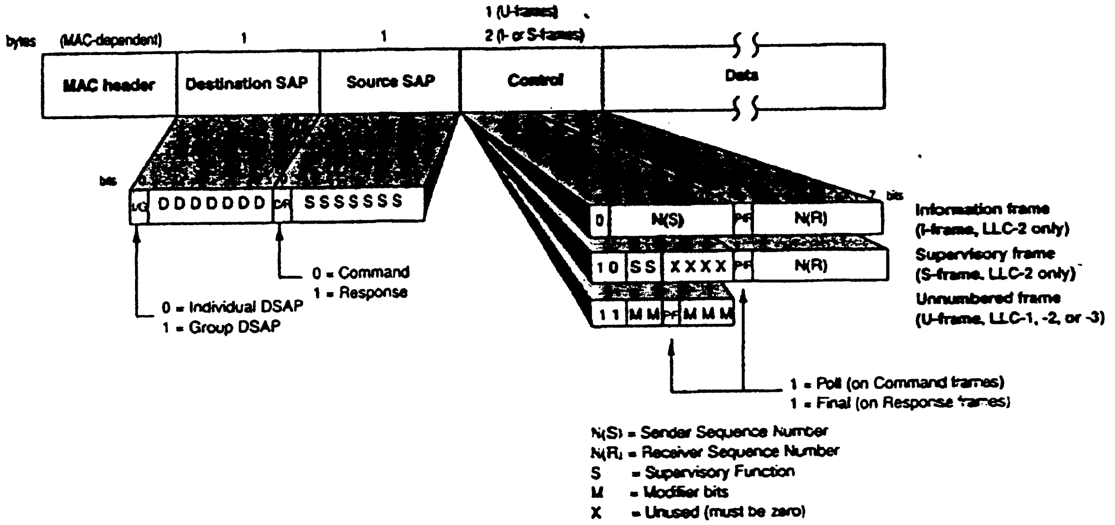
b) The data segment is encapsulated with a TCP datagram as its payload. The fields are added to give TCP a basic layer 4 flow control function and set up virtual circuits or ports.

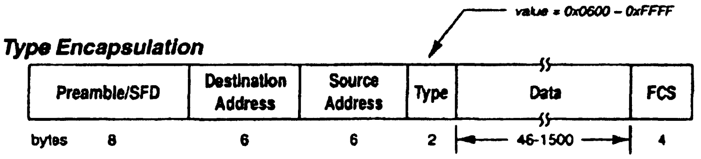
Source port	Dest. port	Sequence number	Ack number	Control flags	Window field	Checksum	Urgent pointer	Data
2	2	4	4	2	2	2	2	

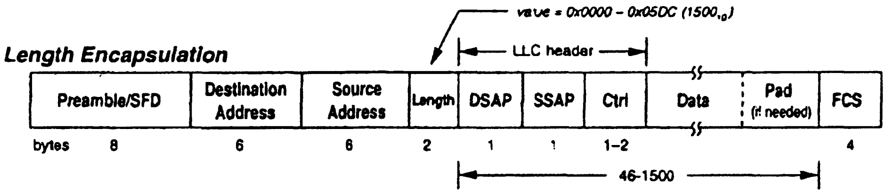
The TCP datagram is then encapsulated within the IP packet where the header and FCS is added to give the packet a layer 3 identity (address) and functionality. The IP address is added along with several control and option fields to manage interconnection and transmission.

The IP packet is then sent to the layer 2 process in order to convert it into a frame suitable for transmission across a LAN. Layer 2 is split into the LLC and MAC sublayers. As the LAN here is an ethernet and uses LLC, then it must use the length encapsulation where the IP packet is placed within the LLC frame and SAPs are added to identify higher layer processes. The frame is also checked for MTU. Which must be < 1500 bytes for ethernet.



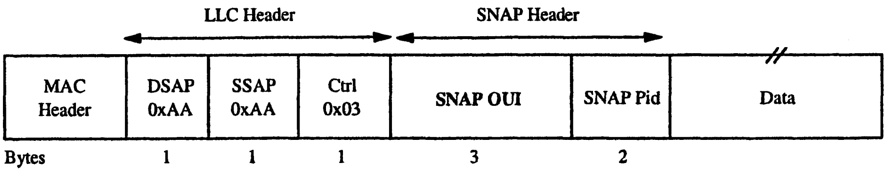

bytes (MAC-dependent) 1 1
1 (U-frames)
2 (I- or S-frames)
MAC header
Destination SAP
Source SAP
Control
Data
bits
I/G DDDDDDD C/R SSSSSSS
0 = Command
1 = Response
0 = Individual DSAP
1 = Group DSAP
7 bits
0 N(S) P/F N(R)
1 0 S S X X X X P/F N(R)
1 1 M M P/F M M M
Information frame (I-frame, LLC-2 only)
Supervisory frame (S-frame, LLC-2 only)
Unnumbered frame (U-frame, LLC-1, -2, or -3)
1 = Poll (on Command frames)
1 = Final (on Response frames)
N(S) = Sender Sequence Number
N(R) = Receiver Sequence Number
S = Supervisory Function
M = Modifier bits
X = Unused (must be zero)



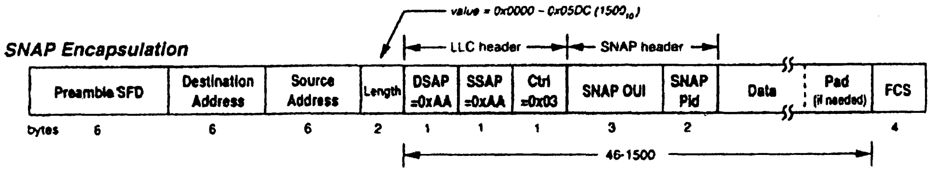


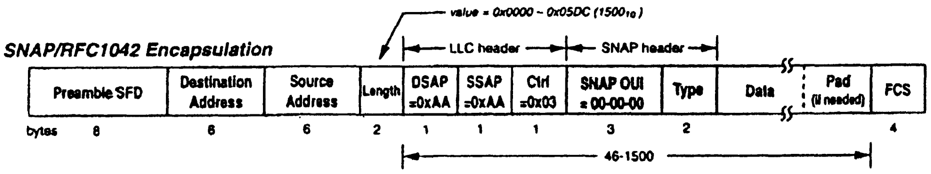
The role of the SAP is to identify at layer 2 the higher layer processes that may (or may not) involve the used of this particular frame. The SAP is used to identify which layer 3 protocol is being used such as IP or IPX etc. This way the next layer format handling stack can be set up. An example of the use of a SAP is in the spanning tree protocol where a SAP of 0x42 is used to indicate that in this case the frame is only for use at layer 2.

c) There are a limited number of SAPs that can be addressed with the single LLC byte, restricting the multiplex to a maximum of 256 clients, half of which are reserved for multicasts. This is even further reduced, as the second bit is often reserved in the SAP as well. The multiplex can be expanded with a further header sub-division known as sub-network access protocol (SNAP) encapsulation. The DSAP and SSAP are filled with the value OxAA followed by the organisationally-unique identifier (OUI) which indicates the organisation for which the protocol identifier (Pid) is significant. This allows any organisation to set up to 65536 private higher layer identifiers. See the MAC section for the OUI and Pid.



A frame encapsulating LLC data could use the LLC SNAP SAP! This means that a LAN can provide type encapsulated data fields even if the native Ethernet MAC doesn't support it through higher layers such as AppleTalk and TCP/IP.

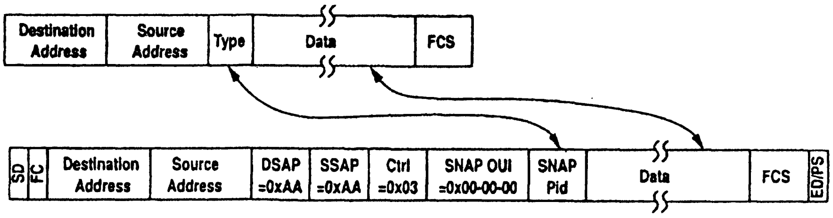




d) The basic function of the translational bridge is to translate frames between different LAN protocols. This process must map between the MAC specific fields in the frames including source and destination addresses, access control and frame control, protocol type etc. The bridge must also map between the different methods used to encapsulate user data such as the Ethernet type field encoding and the LLC/SNAP encoding used in token ring and FDDI. Some of the common data fields are listed below.

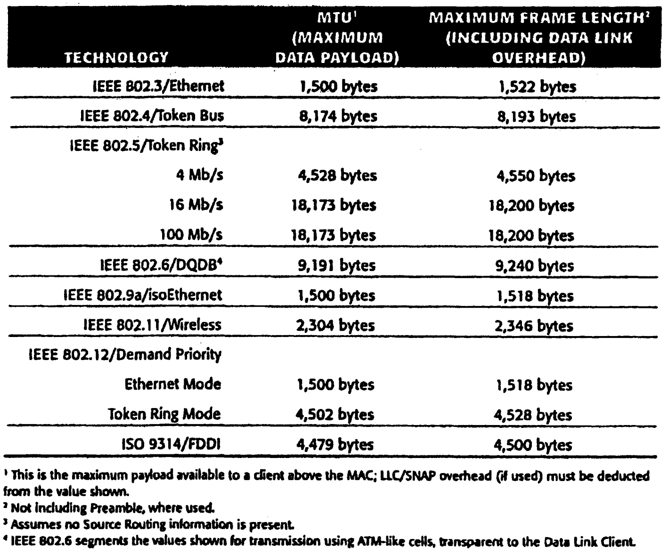
Ethernet	Token Ring	FDDI
	Access Control	
	Frame Control	Frame Control
Destination Addr	Destination Addr	Destination Addr
Source Addr	Source Addr	Source Addr
Length/Type		
Data	Data	Data
FCS	FCS	FCS
	End delimiter/ Frame status	End delimiter/ Frame status

As can be seen in the above table, when data is passed from one MAC to another by the bridge, it must discard and add certain fields which will be expected by the new MAC protocol. Eg when forwarding to a token ring an access control field must be added with token and monitor bits set to zero. Priority is often ignored, however it is possible to create a new high priority token on the token ring network based on higher level functions set within the bridge. In a similar fashion with the token ring and FDDI MACs, the frame access fields and the end delimiter/frame status fields must be created.



Notes: Arrows indicate that the literal contents are copied to the appropriate field(s).
Fields without arrows are created as needed for the transformation.
The Frame Check Sequence must be recalculated after the transformation.

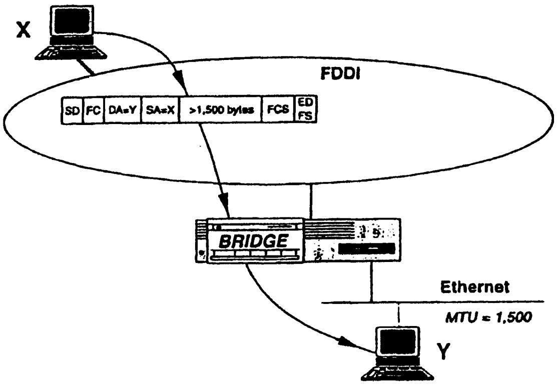
e) One of the biggest problems in transmitting data between different LAN types is the different maximum frame size allowable. These are referred to as maximum transmission units (MTUs). The table below shows a few different MTUs for different LAN types.



TECHNOLOGY	MTU[1] (MAXIMUM DATA PAYLOAD)	MAXIMUM FRAME LENGTH[2] (INCLUDING DATA LINK OVERHEAD)
IEEE 802.3/Ethernet	1,500 bytes	1,522 bytes
IEEE 802.4/Token Bus	8,174 bytes	8,193 bytes
IEEE 802.5/Token Ring[3]		
4 Mb/s	4,528 bytes	4,550 bytes
16 Mb/s	18,173 bytes	18,200 bytes
100 Mb/s	18,173 bytes	18,200 bytes
IEEE 802.6/DQDB[4]	9,191 bytes	9,240 bytes
IEEE 802.9a/isoEthernet	1,500 bytes	1,518 bytes
IEEE 802.11/Wireless	2,304 bytes	2,346 bytes
IEEE 802.12/Demand Priority		
Ethernet Mode	1,500 bytes	1,518 bytes
Token Ring Mode	4,502 bytes	4,528 bytes
ISO 9314/FDDI	4,479 bytes	4,500 bytes

[1] This is the maximum payload available to a client above the MAC; LLC/SNAP overhead (if used) must be deducted from the value shown.
[2] Not including Preamble, where used.
[3] Assumes no Source Routing information is present.
[4] IEEE 802.6 segments the values shown for transmission using ATM-like cells, transparent to the Data Link Client.

The main problems occur when connecting an FDDI LAN with an MTU of 4479 bytes to an Ethernet with an MTU of 1500 bytes. A transparent bridge would either discard the frame at source or try and transmit it where it would be discarded at the receiver.



There are three basic solutions to this problem.

1. Set the largest possible frame size to 1500 bytes on the FDDI system.
2. FDDI is often used as a backbone for interconnecting Ethernets. In this case there will never be an FDDI data frame generated as all sources will be Ethernets.
3. Deal with the problem at the network layer by fragmenting long frames.

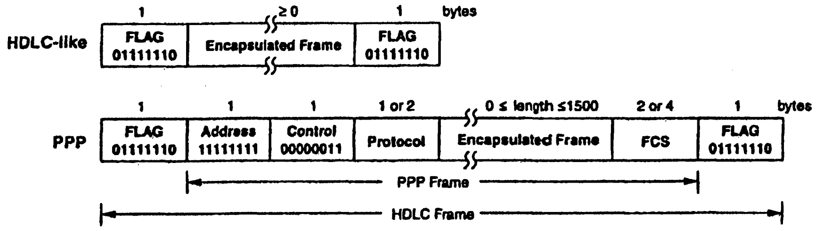
Q2 a) There are some key differences between the technologies use in LANs and WANs. Important factors which will influence technology decisions include:

- ✓ Data rate. In general, WANs operate at a much lower data rate than LANs. LANs can operate at 10, 100 and 1000Mbit/sec whereas a T1/E1 WAN is more likely to operate at 1 or 2Mbit/sec
- ✓ Error rate. Error rates in WANs are typically an order of magnitude less than in a good quality LAN.
- ✓ Cost. LAN bandwidth is basically free. Once installed, there is little recurring cost. A WAN link is normally leased from a service provider, which will incur running costs.

The process of transmitting a LAN frame over a WAN link is rarely done by translation, as it would involve a loss of data structure at either end and there is no direct mechanism for using a 48bit globally unique MAC address in a wide area connection protocol. The commonest method used is to encapsulate the MAC frame within a frame format which suites the wide area protocol used.

- ✓ Support for a single device encapsulating multiple protocols across a WAN link.
- ✓ Vendor (ISP) interoperability across the WAN

- ✓ Error detection (FCS)



PPP encapsulation shown using default HDLC framing and maximum length

The problem with protocols like PPP, frame relay and X.25 is that there is limited flexibility in the way in which a WAN can be set up. PPP normally assumes a mesh network of connections or channels which cannot be changed dynamically or quickly. Hence as the networks expanded there was a need for a more flexible mechanism for interconnecting networks. At the same time, the DARPA experimental network which was mostly between the military and universities was taking off using TCP/IP. It became obvious that the key was in going to a higher, more sophisticated layer in the OSI model giving layer 3 switching using IP addresses. Initially this was limited by the technology to a few large nodes or routers, which knew, learned and adapted the structure of the network. As the number of nodes grew, so did the technology and switching became localised at layer 3. Many LANs are now run entirely through layer 3 with ARP to convert IP to MAC addresses.

b) A very important rule in establishing a layer 3 switching system is the concept of the fast path functionality, as it would be impossible to design a cheap layer 3 switch for a LAN which could cope with the myriad of minutiae which go into the IP protocol stack. Hence a fast path is established, where the majority of packets in a data stream are dealt with in a common fashion, making a hardware implementation feasible. Luckily the more complex features in a IP packet are rarely used, and can be dealt with in a series of much slower boundary options within the router as these cases comprise only a tiny fraction of overall traffic. Similarly there is no need to support housekeeping functions such as ICMP and SNMP. Hence the majority of the packets are split off to form the fast path. What functions are in the fast path? This will vary, but we can consider unicast IP traffic as a good example:

Packet parsing and validation – The router need to separate certain fields to determine the type of handling required.

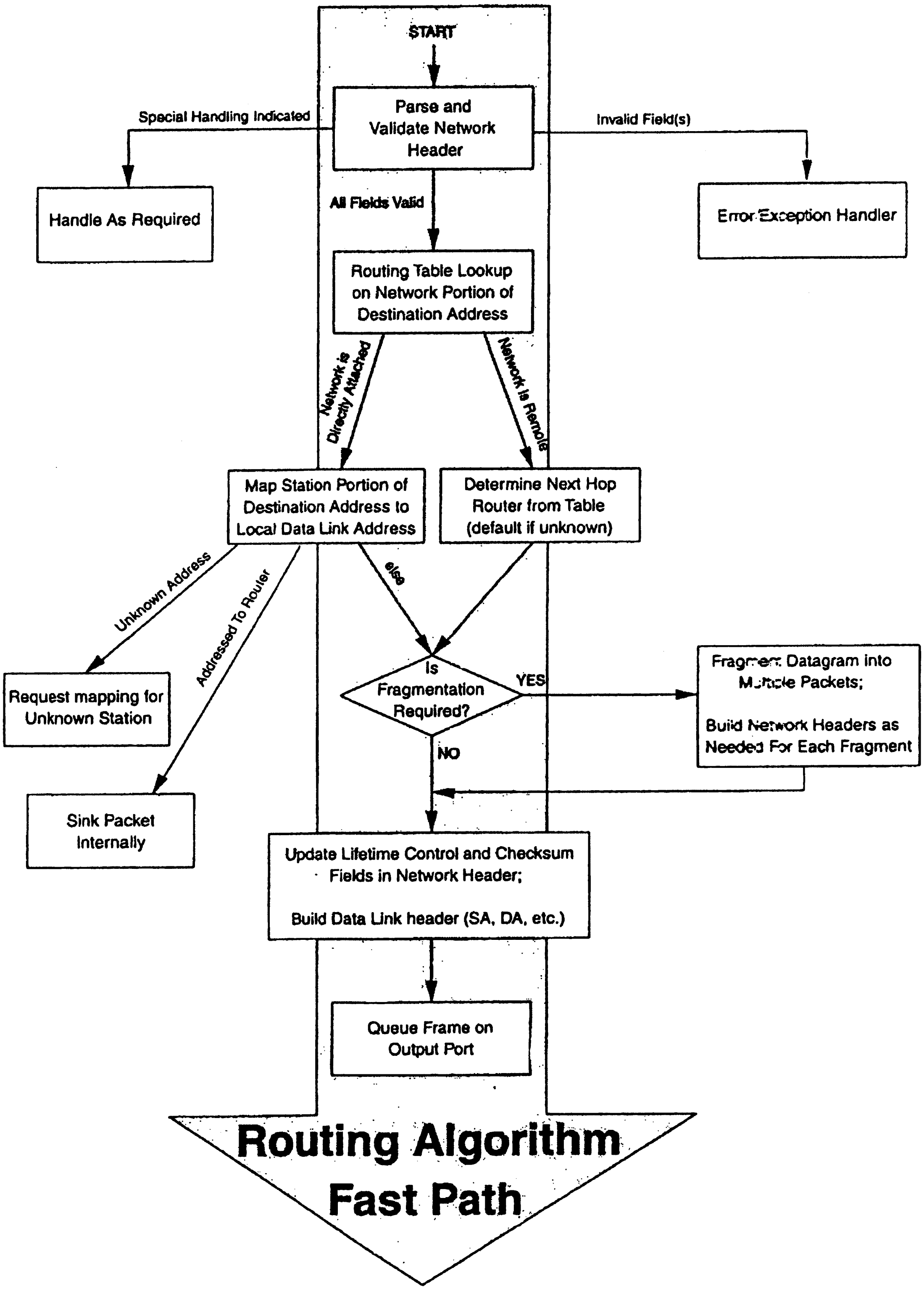
- ✓ Check the IP version number
- ✓ Check the header length field (>20bytes means routing information present)
- ✓ Calculate the header checksum
- ✓ Validating the source address

Routing table look up – The router performs a table look up to determine the output port to direct the packet to, based on the network identifier of the IP address. The result of this will be that either:

- ✓ The destination network is reachable only by forwarding the packet to another router (remote network). This may occur due to a match with the network identifier or due to the selection of a route for an unknown destination network. Either way the look up will return the address of the next router and the port on which it can be reached.
- ✓ The destination network is known to be directly attached to the router. The station identifier portion of the IP address must also be mapped onto the layer 2 address using the address resolution protocol (ARP) cache.

Fragmentation – Each available output port will have an associated maximum transmission unit (MTU), which is the largest frame length permitted. It is generally a function of network technology or MAC used (Ethernet, Token ring, PPP etc). If the packet is larger than the MTU then it must be fragmented. Layer 2 protocols have no mechanism for fragmentation, however layer 3 protocols such as IP do. This is often a mixed blessing as the processing burden is high. Hence fragmentation is usually avoided or dealt with outside the fast path.

Update lifetime control and checksum – The router adjusts the time to live (TTL) field in the packet which is used to prevent packets from endlessly bouncing around internetworks. Packets have their TTL decremented when routed and are discarded once the TTL expires. Finally the header checksum is recalculated for the new TTL.


START
Parse and Validate Network Header
Special Handling Indicated
Handle As Required
Invalid Field(s)
Error Exception Handler
All Fields Valid
Routing Table Lookup on Network Portion of Destination Address
Network is Directly Attached
Network is Remote
Map Station Portion of Destination Address to Local Data Link Address
Determine Next Hop Router from Table (default if unknown)
Unknown Address
Addressed To Router
else
Request mapping for Unknown Station
Sink Packet Internally
Is Fragmentation Required?
YES
NO
Fragment Datagram into Multiple Packets;
Build Network Headers as Needed For Each Fragment
Update Lifetime Control and Checksum Fields in Network Header;
Build Data Link header (SA, DA, etc.)
Queue Frame on Output Port
Routing Algorithm
Fast Path

The majority of packets routed will undergo this exact process. A few specialised processes will occur off the fast path:

- ✓ Fragmentation and assembly
- ✓ Source routing option
- ✓ Route recording option
- ✓ Timestamp option
- ✓ ICMP message generation
- ✓ Routing protocols (RIP, OSPF, BGP)
- ✓ Network management (SNMP)
- ✓ Configuration (BOOTP, DHCP)

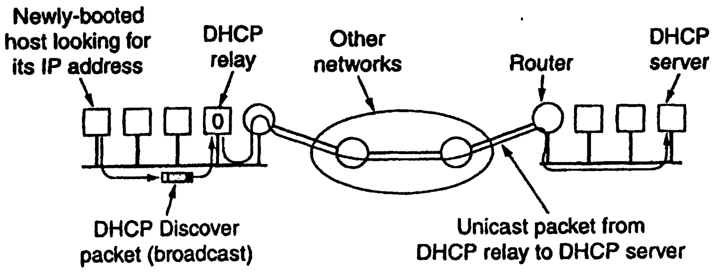
c) The IP version 4 address is 32 bits long, total number of addresses = 2^{31} = 4,294,967,296 addresses. Unfortunately the origin of the IP address is through DARPA and they defined a series of IP address classes which limits the number of unicast addresses which are available.

Class A: Sets the first bit as 0, bits 1 to 7 as the network ID and bits 8 to 31 as the host ID. This gives 126 networks and approx 16 million devices on each network.
Class B: Sets the first two bits as 10, bits 2 to 15 as the network ID and bits 16 to 31 as the host ID. This gives 16382 networks and 65134 devices on each network.
Class C: Sets the first three bits as 110, bits 2 to 23 as the network ID and bits 24 to 31 as the host ID. This gives 2 million networks and 254 devices on each network.
Class D: Sets the first four bits as 1110 and is used for broadcasting and multi-cast addressing
Class E: Sets the first five bits as 11110 and is for future use.

An extreme estimate of the unicast address space which can be uniquely allocated using the IP class system would be 126 + 16382 + 2097150 = 2113662 (note this excludes address portions which are all 0's or all 1's as they are reserved for internal procedures). This is a severe limit on the number of nodes which can be set up on the internet. It is not totally pessimistic as all three classes of addresses can define multiple stations. Hence a station total is 126*16777216+16383*65134+2097152*254 = 3.7billion addresses (86% of total).

There have been two mechanisms proposed to prevent the limited number of IP addresses from throttling the internet. The first was to go to IP version 6 which along with a host of more sophisticated link management procedures also includes a 64 bit back compatible with version 4, address space. The problem has been how to get users on the internet to adopt it?

A second solution was to use the an updated version of the bootstrap protocol (BOOTP) which manually assigned IP addresses, called the dynamic host configuration protocol (DCHP).

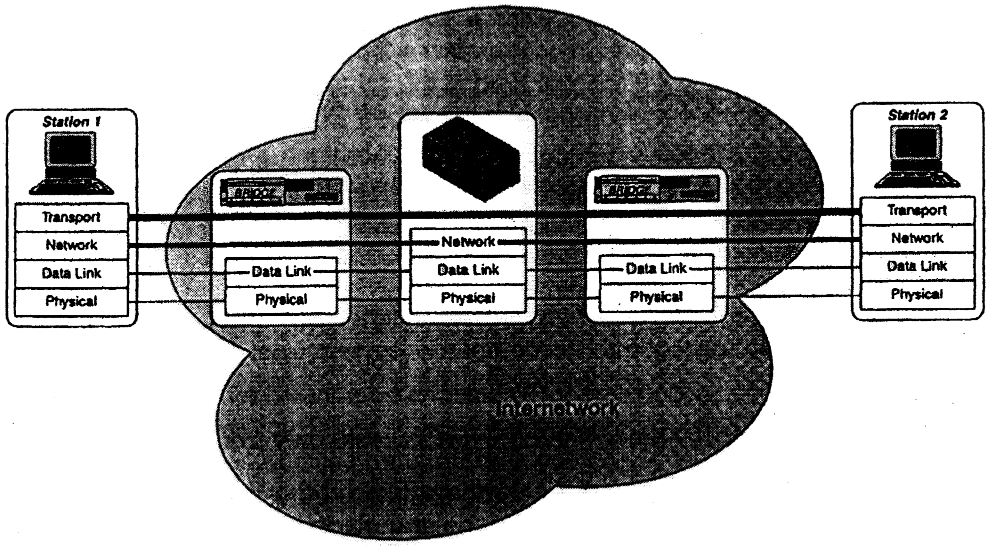

Newly-booted host looking for its IP address
DHCP relay
Other networks
Router
DHCP server
DHCP Discover packet (broadcast)
Unicast packet from DHCP relay to DHCP server

DCHP allows both manual and automatic IP address assignment and has largely replaced both BOOTP and the reverse address resolution protocol (RARP). DCHP is based on a special server which assigns IP addresses to stations asking for one. In order to allow the server to be accessible from all stations, a relay agent is used at the edge of each LAN to forward DCHP requests. An IP address can be allocated for the time the station is connected or it can be leased for a fixed period of time. The problem with this system is the extra complexity and extra hardware and software needed. Plus security issues.

d) TCP can offer very complex services including acknowledged connection oriented service and even transmission via a form of virtual circuit to minimise delay and maximise quality of service (QoS).
The basic structure of the TCP (or UDP) header contains several important data fields. The *source* and *destination ports* are used to identify well known application processes such as FTP or SMTP and can be used in the same way a logical connections or virtual circuits are identified in frame relay or X.25. Some protocols such as FTP use 2 ports, 21 is defined form control and 20 for data transfer. Ports can also be scanned for software vulnerabilities in higher layers and are common targets for hackers.

Remember that this on top of IP which is a connectionless protocol where packets are routed as they enter each network node. Hence it is the IP layer which controls the actual flow of the packets across the network and offers no guarantee of service from end to end. An IP packet cannot easily control the delay it encounters which is a problem for services such as video or VoIP.

e) Manufacturers also actively market layer 4 switches which operate on the transport layer, normally providing end-station to end-station services across an underlying internetwork. TCP operates only between end stations and underlying bridges and routers are not involved.



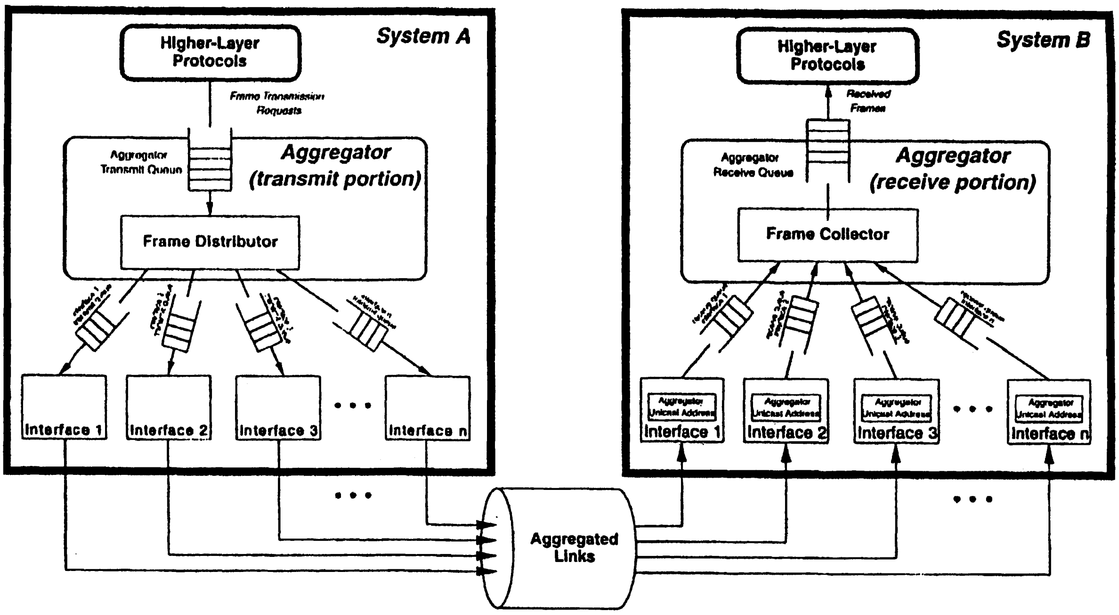
Strictly speaking a layer 4 switch is not possible as there is no means of identification (packet ID or network address) at this level. However higher layer policies such as specific address or domain filtering and security could be implemented. Many higher level features can be specified for the switch to operate on such as network management, congestion control and delay sensitive traffic priority such as video streaming. The control of layer 4 operation using protocols such as TCP is often done on a stream of packets referred to as an *application flow*. But in summary, a layer 4 switch does not implement layer 4 functionality, it is still a layer 3 device.

Q3 a) Link aggregation is a technique where multiple links are combined to make a single link with a higher bandwidth. This is often referred to as trunking or bonding. This technique provides higher bandwidth connections and it allows a network to be expanded in multiples rather than factors of 10. It also provide a form of redundancy should one part of the aggregate link fail, the rest will keep working. Of course nothing comes for free.

- Additional interfaces will be needed at both ends.
- Additional slots may be used up
- Additional complexity is needed in device drivers
- Additional controls will be needed for smooth operation.

In a traditional non aggregated link, each network interface controller (NIC) has a globally unique 48bit MAC address. This is used as the source and destination address for the station. When an aggregated link is set up, the link should appear to higher levels to have a single MAC address, however this is not case in hardware as each NIC has its own address. Hence the software driver which is controlling the aggregated NICs must take a single address and assign it to an aggregated group of links. This can be done by overwriting the MAC address register in software.

The key question in an aggregate link is how to assign the data to each link in the aggregate. A *striping* technique could be used where a frame is split between links and sent in parallel, however this is not possible with LANs. Hence whole frames are sent along each link and the problem is how to manage this process without breaking the LAN hard invariants. This is particularly difficult when trying to maintain frame sequence.



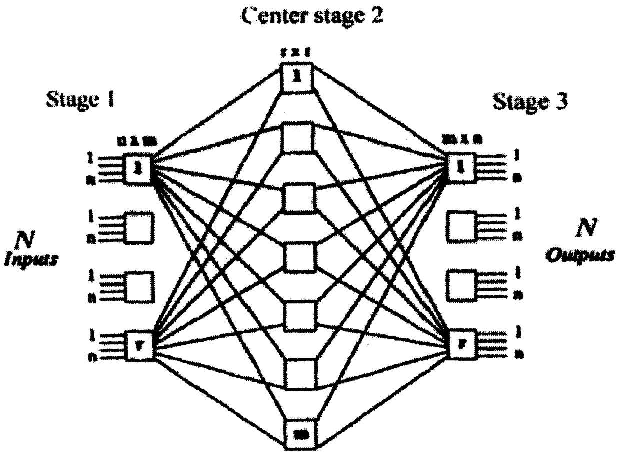
The key question in an aggregate link is how to assign the data to each link in the aggregate. This is most efficiently done at layer 2 as there is direct access to hardware addresses for each link in the aggregate. A link aggregation process does not want to spend unnecessary time deciding ARP lookup etch on layer 3. A *striping* technique could be used where a frame is split between links and sent in parallel, however this is not possible with LANs. Hence whole frames are sent along each link and the problem is how to manage this process without breaking the LAN hard invariants. The secret to maintaining frame sequence is to find why frame order must be maintained and which frame need to be kept in order. We can then relax the strict invariant requirement of LAN transmission. Not all traffic across the same LAN link will be from the same stations or applications so not all order is essential. In an aggregate link, a *conversation* (sometimes called flows) is defined in traffic when order must be maintained. Hence the distributors job means that frames from the same conversations must be sent down the same link. An example could be in a switch to switch link aggregate where destination MAC addresses make a very good means of determining conversations. This technique does not work well in switch to server connections as all frames will carry the same MAC destination address. A better system might be to use MAC source addresses.

b) Pure electronic switching is limited in two main ways. It is a 2D technology which makes crossovers a complex issue and it is a planar technology which means that there is always an inherent capacitance within the electronic circuits. This capacitance is the limiting factor on link length, speed and power dissipation and is overtaken by optical interconnect once data rates exceed 1GHz. Optics is a true 3D medium with no interference between optical beams in either free space or in waveguides. Hence a crossover is simple to make. The limits on speed are normally at the interface between the optical and electronics at either the laser (modulator) or photodetector interface electronics where capacitance becomes an issue again. Hence, the longer data can remain in the optical domain, the more efficient it becomes. This is particularly relevant in optical switching where crossovers are key to the successful operation of a high speed data switch. The fact that the majority of transmission networks are now optically based is another incentive to be able to implement all optical switching systems.

The limiting factors on the size of in optical switch are mostly due to physical constraints in building the system.

- As the number of ports in creases, so does the physical size of the switch which puts pressure on the design of optical components (large angles) , opto-mechanics and waveguide arrays.
- The loss and crosstalk in an optical switch are also normally related to the number of ports, hence they will become a limiting factor.
- The control and arbitration is also a problem as this has to be done in electronics and becomes very complex when managing buffers and headers.

c)



A 16 × 16 three-stage Clos network.

N= no. of inputs and outputs (N=n x r)

n = no. of inputs per switch in stage 1 & no. of outputs in stage 3.

m = no. of outputs per switch in stage 1 & no. of inputs in stage 2.

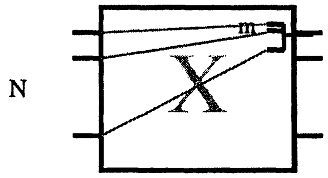
r = no. of switches per stage in stages 1 & 3.

A Clos network is strictly non-blocking if m is greater than or equal to 2n-1. In the above example the number of crosspoints = 2*4*4*4 + 4*4*4 = 192 compared to 16*16 = 256 for a full crossbar. m is less then 2n-1 hence the switch is not strictly non-blocking but it is re-arrangeably non blocking. For a Clos network the advantage in reduction of crosspoints occurs between N = 15 and N = 16 ports. 7 centre stages are needed to make it strictly non-blocking but number of crosspoints now = 567

d) A packet switch has the requirement that all inputs should be able to be connected to a single output, hence to perform this task an NxN switch would need N^3 crosspoints if it were a crossbar. A Clos network is designed to be centre symmetric, hence it cannot easily cope with the full packet switch requirement without an enormous number of crosspoints and some rather complex buffering in the centre stages.

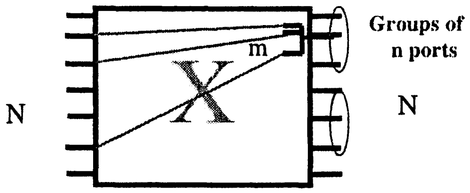
There are two ways in which a Clos network could be improved under the conditions of packet switching.

1) Speed-up: Add extra paths through the centre stages to allow for multiple access to output ports. Output ports must now be capable of queuing the multiple packets at the output. Surprisingly only a speed up of 2 is needed which is much better than n^3 for the crossbar.



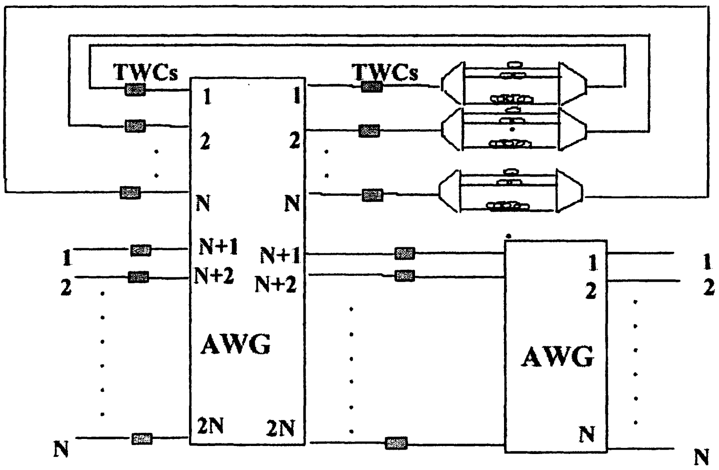
2) Grouping: The nearest neighbours to the selected output ports are grouped together so if a port is busy it will send the packet to the next nearest port. Typical grouping would be 4 output ports.

Speed up = m/n



Switch dimensions
N x n/mN

e) Wavelengths can be used to route data instead of packet headers. Each route to an output port is designated by a particular wavelength. The route corresponds to a channel in a passive AWG structure which maps wavelengths to particular locations.



The diagram represents a single layer where a single wavelength enters through each port 1 to N. The components are as follows:

AWG – this is the fixed waveguide device which maps a wavelength to a particular output port.

TWC – are tuneable wavelength converters which take in a single wavelength of data an convert it to another wavelength.

Fibre buffers – this is a series of selectable delay lines (fixed length of fibre) which can delay a single packet while waiting for an output port to become free.