PART IIB 2011 4B15 ADVANCED TELECOMMUNICATIONS NETWORKS DR TIM WILKINSON

4B15 2011 exam cribs (note they are more verbose than expected from candidates)

Q1 a) A bridge is a piece of OSI layer 2 hardware, which allows frames to be passed between LANs that have different geographical locations and even different LAN protocols (MACs). A bridge is a data-link layer device and any interconnection of LANs via bridges is often referred to as a catenet. Some basic bridge principles:

- Each station has a globally unique 48 bit unicast address.
- There is a table within the bridge which maps station addresses to bridge ports.
- The bridge acts in promiscuous mode, it receives (or attempts to) every frame on every port.

When a frame is received on any port, the bridge extracts the destination address from the frame, looks it up in the table and determines the port to which the address maps. If the looked up port from the table is the same as the one on which the frame arrived then the frame is discarded, as it assumes that the station on that port will have already received the intended frame (filtering). If the frame received by the bridge is mapped to a different port than the one it arrived on, it then forwards the frame to the port and onto the appropriate LAN.

A bridge tries to make a catenet appear transparent to end stations, as if it were a single LAN. Hence higher layer services will expect a LAN-like performance from the catenet below it. A LAN data-link must exhibit certain properties. Hard Invariants, Non duplication of frames, Sequential delivery of frames. These invariants and the complexity of the address table limit the scaling on the catenet, hence a bridge is not a suitable device to maintain a global network such as the internet.

As with the technological evolution of the bridge into the switch, the advance of technology has allowed the use of network layer routing or layer 3 switching to be implemented as part of a LAN. Using modern silicon integration and application specific ICs (ASICs) it is now possible to build routers for similar costs as layer 2 devices. Wire speed devices now marketed as 'layer 3 switches' and the difference between layer 2 and 3 switching depends on the needs of the station and administration. There are a whole host of layer 3 protocols to choose from including IP, IPX, DECnet, Phase IV, AppleTalk and the OSI CLNP, however the majority of networks and network vendors have migrated towards the internet protocol (IP) as the preferred protocol of choice. Hence most routers are based on IP, with a few offering limited IPX functionality. The global acceptance of the layer 3 IP address has meant that it is an ideal method for making a global network structure.

The router is not hampered by the restrictions of hard invariants, hence the layer 3 functionality allows it to learn network structures and routing paths as well as optimise its operation through a whole host of criteria such as latency, number of hops or error rate. The cost of this is in operational complexity.

b) The address table is built automatically by considering the source address of frames received by the bridge. The bridge will look up the destination address in order to assign a port, and at the same time it will look up the source address to see if it has ever heard from that particular station before. When an entry is not found for the source address in the table, then a new entry is created. If there is already an entry, then it is updated to the port from which the frame was originally received. Over time the bridge will learn a port mapping for all active stations on the LANs.

If a bridge only ever learned address to port mappings, then two problems would occur: If old entries are not removed from the table, then its size will increase and will eventually take too long to search through. If a station moves from one port to another, then frames will be sent to the incorrect port until the moved station decides to transmit itself. This could take forever with a poorly designed upper layer structure.

The simple solution to both these issues is to age entries on the address table until they become stale entries, which have expired and are removed from the table. The definition of activity is based on appearance of the source address only. In a typical bridge address table, there will be a series of bits stored, which indicate the age, and status of a table entry. The commonest way is to use 3 bits, the valid bit (V) and the hit bit (H) dictate how old the entry is. The V bit indicates if an entry is currently valid in the table, and the H bit indicates that a source address has appeared within the last ageing cycle. The typical length of the ageing cycle is around 300 seconds (5 minutes). Some tables use a third bit as a static bit to indicate an address entry which cannot be aged or changed.

A router undergoes a similar process to prevent stagnation of its table. It will either periodically download routes from other sites or time out in active entries from its table.

c) The spanning tree protocol uses accelerated table ageing to prevent instability in its structure, especially when events such a link failure lead to changes in the STO topology. If the topology changes it is usually due to component or link failure, management intervention or network evolution. STP treats these as boundary events and can be slow to react to changes. Hence a smoothing procedure is employed to prevent possible shocks.

- Provides for explicit topology change and notification.
- Provides for acknowledgement of the changes.
- Uses timers to:

1. Prevent rapid transition between blocking and forwarding states. Hence minimise transient loops
2. Allow bridges to participate in the election of new designated bridges or ports. This prevents recursive changes

Before transitioning to the forwarding state as a result of a topology change, a port will wait for the topology change information to propagate through the catenet. Weird things can happen when a topology changes, hence when the topology bit is set in a BPDU the ageing timers for address tables are set to a much shorter duration, which means they will purge old values much quicker. This is important because, when topologies change, huge chunks of addresses will appear to shift from port to port.

d) The roles of the station and bridge are completely reversed in source routing when compared to the operation of transparent bridges. The end stations are totally responsible for the transportation of traffic across the catenet and the bridges are totally unaware. Hence an address table only exists for the local ports on each bridge, however if the STP is used, then more information is required to be stored about the structure and status locally of the ST.

Each end station in the catenet connection must learn all about the available paths in the catenet through the process of route discovery. There may be multiple paths through a catenet and the stations must select the route to use. Normally both stations will use the same route forward and in reverse, with the route being stored in cache until the connection is over. During route discovery, the station may learn that the end station is on the same ring in which case normal token procedure is invoked and source routing not used. During the discovery process the end stations will also learn the maximum transmission unit (MTU) which can be handled across the path, which will then set the MTU used in data transmission.

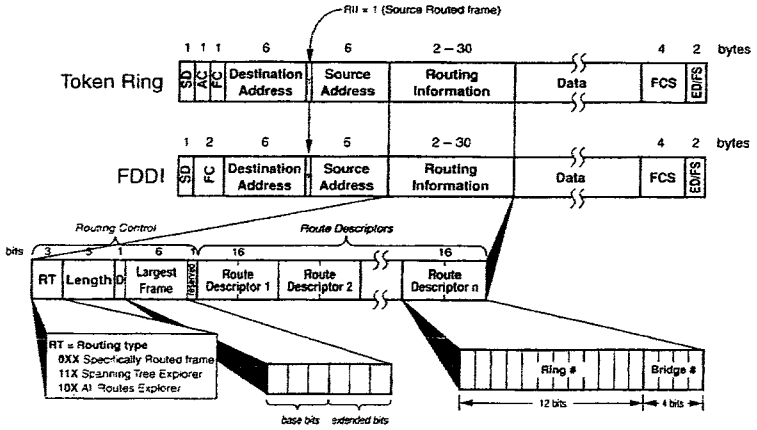
Source routing inserts extra fields into the MAC frame, hence it is necessary to indicate presence of a source routed frame. This is done by using the first bit of the 48bit source address in the MAC frame, which normally defines unicast or multicast address format. This bit is normally wasted as multicast source address is meaningless, hence it is used to indicate (set to 1) a source routed frame. The use of a address table in source routing is quite different as there is no need to have a look-up table in the strict sense, just a mapping of the nearest links which lead to other networks. Hence in the pure form, ageing would not be helpful in a source routed sytem as it effectively ages itself when each route is discovered.

If the source routing system uses a spanning tree, then an ageing process could help maintain the tree when topology changes occur. This assumes that an automatic from of STP is used in the source routing system through the broadcast of BPDUs. This would work as long as topology changes were relatively rare. This is the case with source routing as it was originally based on ring type MAC LAN structures.

Examiner's comment:

This question was well answered with most knowing bridging well as this was part of the coursework. The last section was more speculative and was well answered.

Q2 a) It is possible to configure and operate a network, which has no knowledge of its interconnection or the positions of stations upon it. All of the route configuration and set-up processing is performed by the stations at either end of the route or link(s). This is referred to as source routing and the bridge used in such catenets is known as a source routing bridge. Source routing inserts extra fields into the MAC frame, hence it is necessary to indicate presence of a source routed frame. This is done by using the first bit of the 48bit source address in the MAC frame, which normally defines unicast or multicast address format. This bit is normally wasted as multicast source address is meaningless, hence it is used to indicate (set to 1) a source routed frame.



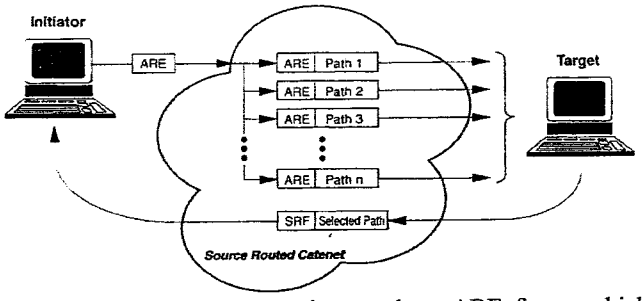
The routing field added to the frame consists of 2 parts: A 2 byte routing control segment (always present), A 0-28 byte variable length list of route descriptors. Big endian format, read MSB first.

Specifically routed frames (SRF) – Carry routing type 0b0XX. A SRF carries a list of route descriptors and is forwarded along this route. This is used for the bulk of data frames. Source frames which are not specifically routed are called *explorer frames* and have the first bit of the routing type set to 1

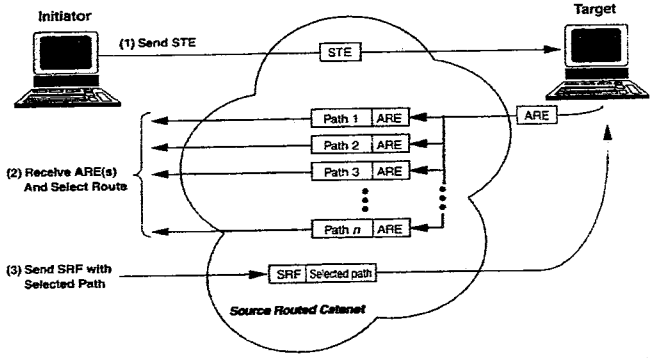
Spanning tree explorer (STE) frames – Routing type 0b11X. A SPE frame is only forwarded by bridges that are designed as part of the spanning tree. One copy of these frames will appear on each ring in the catenet (non-duplicated). They can be used for either multicast traffic or route discovery. STE frames are originally sent with no route descriptors. Forwarding bridges insert descriptors to identify the route taken.

All routes explorer (ARE) frames – Routing type 0b10X. An ARE frame is forwarded by all bridges along every possible path between source and destination. Thus the destination may receive multiple frames for different routes taken. ARE frames are sent with no route descriptors. These are added by each bridge as it forwards the frame.

b) Before an SRF can be sent, the route discovery process must be done to find a suitable route. There are several possible ways of doing this with STE and ARE frames, but here are two most common.



ARE request, specifically routed response – The sending station send an ARE frame which will arrive at the destination by multiple recorded routes. The receiving station selects a route and sends a SRF back to the sender to indicate the chosen route. This places a processing burden on the receiver, so a variation of this is for the receiver to send an SRF for every ARE it receives and let the sender select the route.



Spanning tree request – The sending frame sends an STE frame which is forwarded by the bridges on the spanning tree on all rings in the catenet and eventually to the target destination. The receiving station responds by sending an ARE frame back to the sender which will indicate to the sender all of the available routes. Hence the processing burden is returned to the sender who must select a route.

The route selection process can be a complex one, as it is unknown how many ARE responses will occur. There are two basic mechanisms.

- ✓ *Take the first route* – Just select the first ARE which arrives and discard the rest. The first ARE indicates the route of least delay. This is what the majority of stations will do.
- ✓ *Take the first route which meets a specific requirement* – select the route with a suitable MTU of with a minimum number of bridge hops.

c) There are some specific issues which must be considered when executing the process of route discovery. ***Session disruption*** – If the route is disrupted during a connection, then an error will occur and the discovery process re-initiated. If the route has been recorded by the end stations, it could be re-used, however it is likely to be disrupted, so the best option is to repeat the discover process to find an alternative route. ***Frame explosion*** – The ARE can lead to a large number of frames being generated across the catenet. Which will effect traffic loadings and could lead to a large number of frames arriving on the final ring of the destination. These frames could be very close together, generating congestion. This can be avoided by filtering frames and looking for repeated routes in the frame, thereby discarding the frame. Also, if the number of series interconnected rings is set carefully, then the explosion of frames can be controlled. ***Uneven traffic distribution*** – The first ARE indicates the fastest route at that instant. This may not be the case over the longer connection time. Source routing is often limited by heavily bursty traffic. This is best remedied by higher layer protocols monitoring delays or traffic flow and initiating a new route discovery if delays become too long or congestion occurs. One possible way to avoid this is to store a reserve route for use with congestion.

d) The first consideration is the different frame formats, including bit ordering between the two LAN protocols. The token frame must be converted or encapsulated in an ethernet frame and vice-versa when transmitting in reverse onto the token ring. This includes setting up length or type encapsulation. The first bit of the source address must also be set correctly either side of the bridge, otherwise the frames will be discarded. A key point on the source routing side is that the bridge must set the MTU during the route discovery procedure so that the maximum frame length of ethernet is not exceeded (1500 bytes).

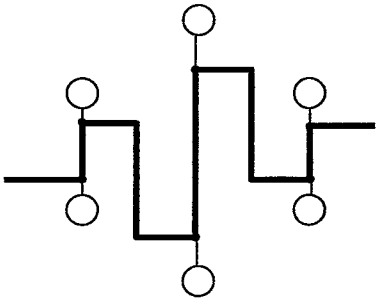
During the route discovery procedure, the bridge must mimic the response of the end station if it is known to be connected to the ethernet side of the LAN. Hence it will respond to the senders request appropriately and once the route has been selected, it will process the data frames sent and discard the routing information. Any unknown destination source routed frames will be discarded, hence the address tables must be kept up to date. The bridge must respond with the appropriate medium access at each port. A robust approach might be to let the bridge have higher layer functionality so that messages for each station can be converted between each LAN type. This would require buffering and higher layer (at least layer 4) knowledge of the protocol stack (such as TCP) to manage the interconnection. It would require a much more sophisticated bridge, however there would be a much greater reduction in the unnecessary transmission of frames across the bridge.

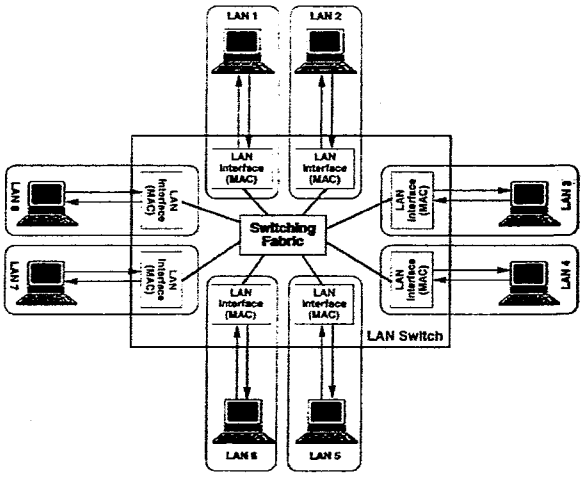
Examiner's comment:

The question was well answered, in the first two sections. There were some interesting answers to part c) with few finding 3 separate problems.

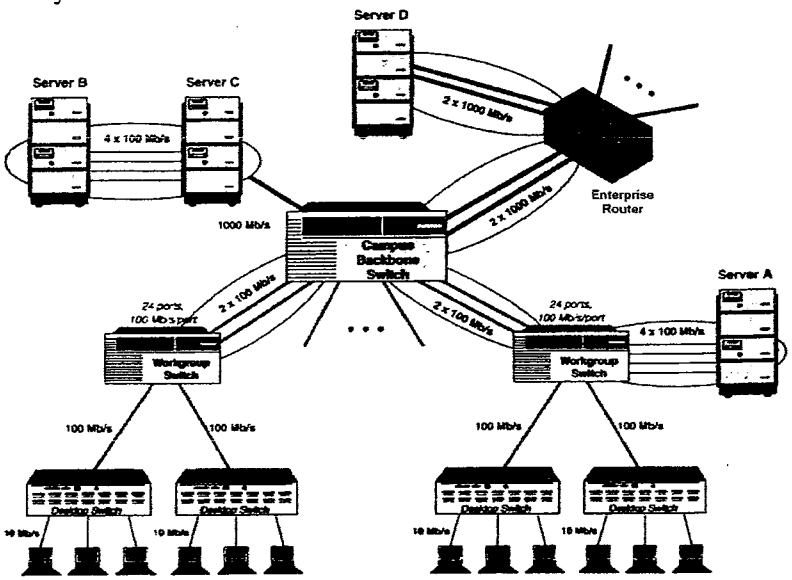
Q3 a) The term *wirespeed* refers to when a bridge or switch is capable of operating all of its ports at the fully specified data rates. One of the main advantages of LAN MAN and even WAN connections is the ability to operate at high data rates. This is especially the case with LANs, however their operation is often hampered by shared media or *half duplex* operation. A more desirable means of operation would be to operate in *full duplex*, where each direction of traffic to or from a station occupies its own physical channel. This has the effect of negating MAC protocols such as CSMA/CD and token rings. There are two main factors which allow full duplex operation.

- The use of dedicated media such as structured cabling
- The use of microsegmentation about a switch.





b) Wirespeed operation and full duplex means that multiple physical links can be combined or aggregated to increase the nett bandwidth between high traffic systems such a routers and servers.



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 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.

So how does an NIC or switch interface decide on individual conversations? This will depend on the applications used to send the data across the LAN. 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.

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.

c) The basic mechanism that can be used to control congestion is to send some sort of signal to the sending station to reduce its rate of transmission or throttle back its frames. There are mechanisms for both half and full duplex links which can avoid higher

layer interactions in order to reduce congestion. For half duplex links, the access control mechanism can be used to force a sending station to reduce its output in order to conserve buffer overflow in the switch.

- **Backpressure** – A switch uses the access protocol to slow the data arriving at its input ports.
- **Aggressive transmission** – A switch tries to remove data quickly at its exit ports by shortening the transmission procedure.

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*. 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.

d) The case for link aggregation at layer 3 is not so clear cut as the majority of the processes on this layer are implemented in software and do not always lead to efficient link operation ate wirespeed rates. There is normally a tradeoff for using layer 3 functions between data rate and flexibility and therefore low latency. Although layer 3 aggregation works in principle as there are no invariant limitations and there is the possibility to sequence, fragment and offset packets. Most of these features are in fact non-fastpath based processes which negates the advantage of link aggregation. Is there any real point in running two links together at half the data rate due to the complexity of the processing overheads.

Examiner's comment:

This question was very well answered by a select few. Not many got the final section right on layer 3 aggregation and the role of the addressing space.