Name……..…ID……..……..Section…….……Seat No……..….

Sirindhorn International Institute of Technology
Thammasat University

Final Examination: Semester 2/2006

Course Title : ITS 413 Internet Technologies and Applications

Instructor : Dr Steven Gordon

Date/Time : Friday 16 March 2007, 13:30 – 16:30

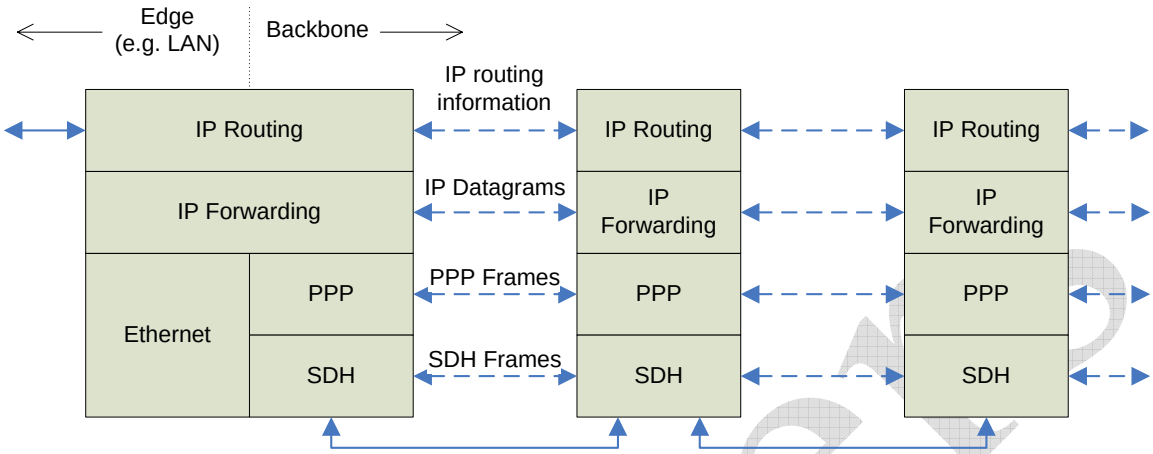
Instructions:

- This examination paper has __ pages (including this page).
- Condition of Examination
  Closed book (No dictionary, No calculator allowed)
- Students are not allowed to be out of the exam room during examination. Going to the restroom may result in score deduction.
- Turn off all communication devices (mobile phone etc.) and leave them under your seat.
- Write your name, student ID, section, and seat number clearly on the answer sheet.
- The space on the back of each page can be used if necessary.

Part A – Multiple Choice Questions [14 marks]

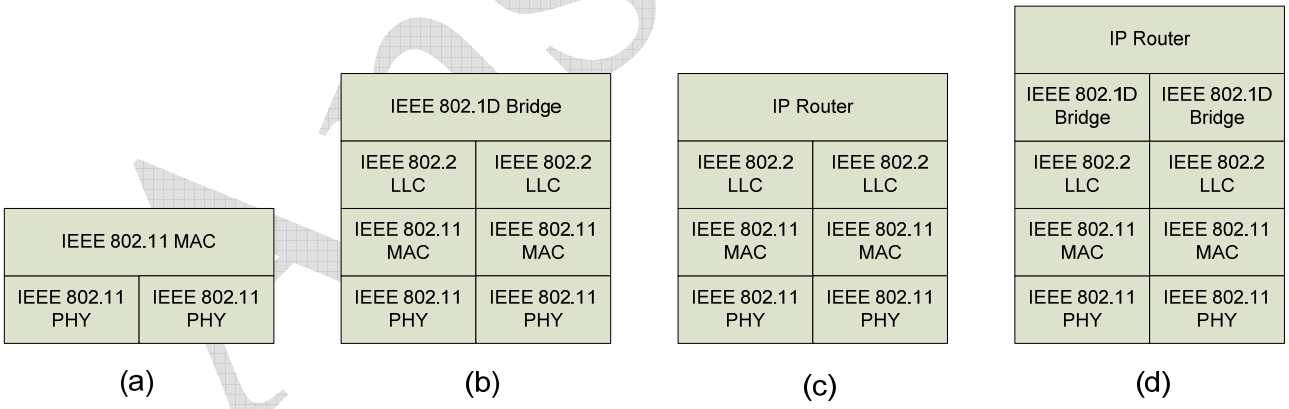
Select the most accurate answer (only select one answer). Each correct answer is worth 2 marks. You receive 0 marks for an incorrect answer or no answer.

1. The protocol architecture diagram below shows a configuration for:



   a. Using Ethernet as a backbone network
   b. Carrying Internet traffic over optical networks
   c. Using the advanced QoS features of ATM
   d. Using the MPLS standard

2. Which of the following protocol stacks best represent an IEEE 802.11 access point?



MAC = Medium Access Control; LLC = Logical Link Control; PHY = Physical; IP = Internet Protocol

b

3. If a small office network consists of two IEEE 802.11g 54Mb/s access points, then which of the following statements is true:
   a. Both access points should use the same channel to maximize the network throughput
   b. Clients using IEEE 802.11b cannot access the network
   c. The total network throughput will be 108Mb/s if the access points use different channels
   d. Users will experience a delay when a handover between access points occur

4. The auto-configuration features of IPv6:
    a. Require a DHCP server to be present
    b. Allows a computer to choose a IPv6 address based on its MAC address
    c. Is used as a transition method from IPv4 to IPv6
    d. Uses duplicate address detection to select the initial link-local address

5. IEEE 802.11 MAC defines a maximum value from which the back-off period is chosen from (e.g. 31). If this value was decreased in the standard then:
    a. Individual frame transmissions would be more efficient, but more collisions will occur
    b. Individual frame transmissions would be less efficient, but less collisions will occur
    c. Individual frame transmissions would be more efficient, but the nodes will not get fair (equal) access to the channel
    d. Less collisions will occur, but the nodes will not get fair (equal) access to the channel
    e. None of the above

6. Tunneling IPv6 over an IPv4 network
    a. Improves the efficiency of the end-to-end connection
    b. Provides a method for upgrading IPv4 to IPv6 networks
    c. Allows all nodes in the network to use auto-configuration
    d. Improves the security of the end-to-end connection
    e. None of the above

7. Which instant messaging protocols use asymmetric servers:
    a. Jabber
    b. AIM and MSN
    c. Yahoo Messenger and Jabber
    d. AIM
    e. Yahoo Messenger
    f. XMPP and Jabber
    g. XMPP and MSN

Part B – Short-Answer Questions [10 marks]

In the following questions write the appropriate word, phrase or protocol to complete the sentence. It is acceptable to use the protocol acronym instead of the full protocol name. Each question is worth 1 mark.

1. _____ is used by TCP as an indicator of congestion in the network.

A packet loss

2. Source and destination _____ and _____ are used to identify a connection in TCP.

IP addresses; ports

3. In TCP Reno, a retransmission may occur after _____ or after _____ are received.

a timeout; three duplicate acknowledgments

4. _____ is an example of a *service user* of TCP.

HTTP (or FTP or …)

5. If describing the protocol rules of TCP, then two events that may occur are _____ and _____ .

Receive DATA; Receive ACK; Timeout

6. _____ is a transport network technology that includes built in quality of service mechanisms.

ATM

7. The concept of _____ is used by both IPsec and methods for transitioning IPv4 to IPv6 networks.

Tunneling

8. Advantages of Peer-to-peer systems over client/server systems include

   _____ and _____ .

Load-sharing, fault-tolerance, scalability

9. _____ uses a client/server architecture for storing the resource index, but a peer-to-peer architecture when peers access resources.

Napster

10. _____ can be used to provide state information for web applications.

Cookies

Part C – General Questions [86 marks]

Question 1 [8 marks]

a) Explain the purpose of a robot exclusion file and how it works (including where the file is and what information the file contains, and how it controls robots – but *you do not have to give the format of the file*). [5 marks]

A robot exclusion file is intended to let web robots (or crawlers) know which parts of a website they should access. The file is "robots.txt" and is located inside a directory of the web server (usually the root directory, e.g. www.example.com/robots.txt). The file contains a set of directives as to what files/directories a robot can access (allow) or not (disallow). The directives can be targeted to specific robots (e.g. Googlebot) or all robots.

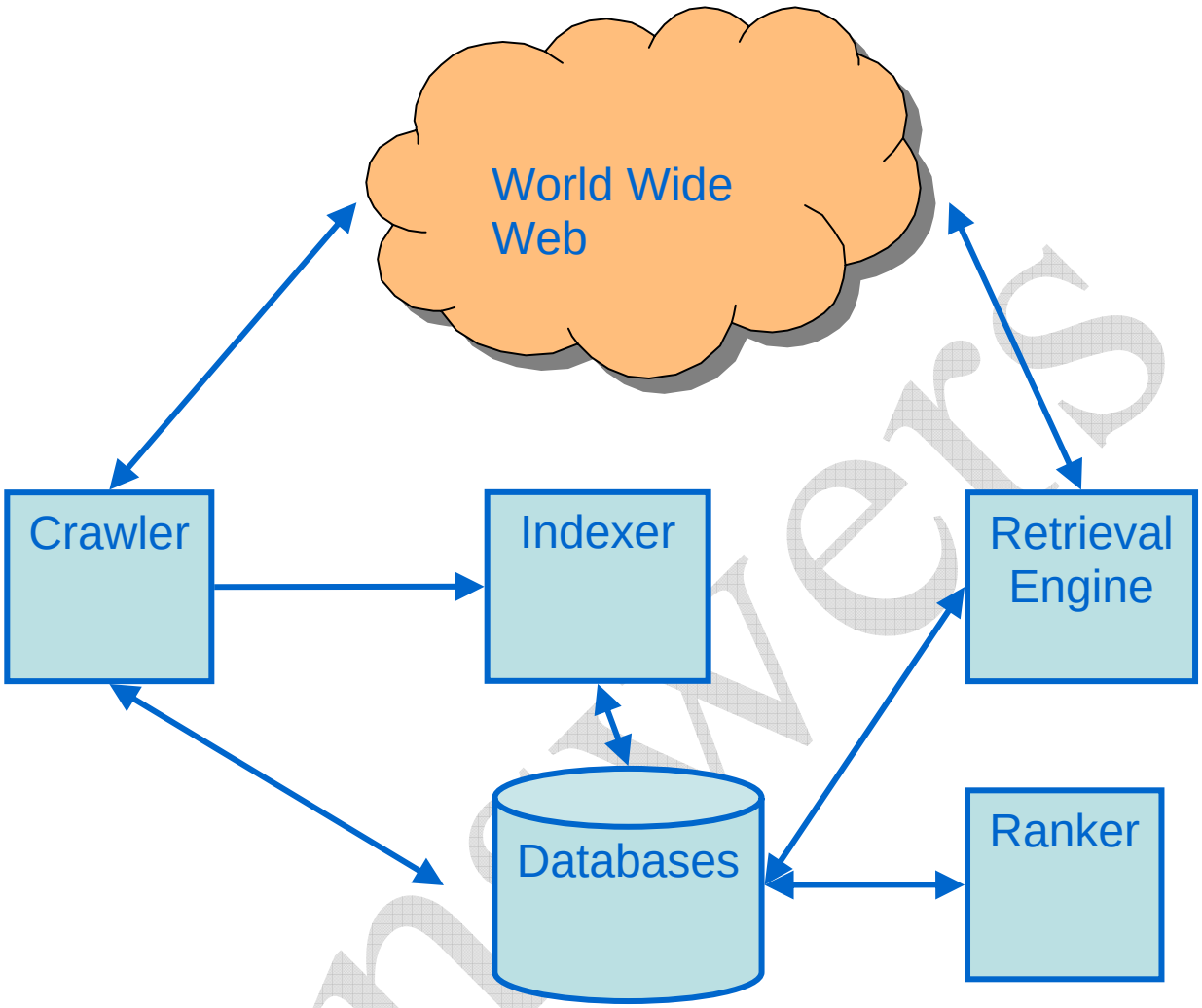
When a compliant robot accesses the web server, it first checks for the existence of the robots.txt file. If exists, the robot follows the directives, that is, only access the allowed files/directories.

b) Is a robot exclusion file suitable for protecting (that is, restricting access to) content on a web site? Explain your answer [3 marks]

No. A non-compliant robot (that is, a robot that does not adhere to the robots exclusion protocol) will not follow the directives and can then access the files. Also, web browsers usually do not adhere to the robots standard, and hence can still access the files.

Question 2 [25 marks]

a) Draw the architecture of a typical search engine. Make sure you label each component and show the connectivity between components. [8 marks]



b) Explain how a search engine crawler works. [4 marks]

A crawler is initialized with a set of URLs to visit. The crawler visits a URL, one-at-a-time, and retrieves the page and extracts the links from within the page. The links that have not yet be visited are added to a "To be crawled" list. Those that have been visited are stored in a "Already crawled" list. The crawler then selects a URL from the "to be crawled" list for the next page to visit, and repeats the process.

c) A search engine crawler picks a URL from a "To be crawled" list. Why is the picking algorithm important? [2 marks]

Because a crawler cannot usually traverse all pages on the web (because too many to index within a reasonable amount of time, and the pages change over time), the crawler must select the page to visit that provides must useful content.

d) Explain the difference between query dependant and query independent ranking algorithms. Give an example set of criteria for each and also give an advantage of each approach (compared to the other approach). [6 marks]

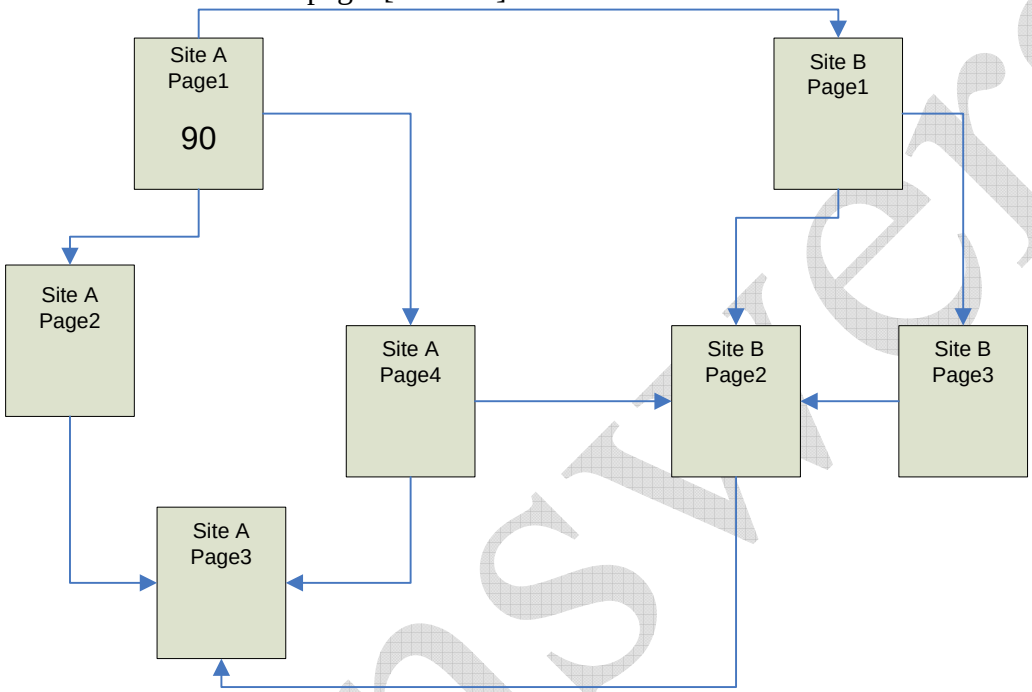
Query dependent ranks pages based in the search criteria/query supplied by the user, whereas query independent ranks pages independent of the user query.

Query dependant example: count of query terms in page; closeness of terms in page

Query independent example: number of links to page; from page; number of accesses

Advantage of independent is that the rank can be pre-calculated (before the user submits a query). Advantage of dependant is that is gives results matching the users query.

e) Calculate the Google PageRank for each page in the diagram below. Write the PageRank value inside the page. [3 marks]



Site A Page 2: 30
Site A Page 3: 90
Site A Page 4: 30
Site B Page 1: 30
Site B Page 2: 45
Site B Page 3: 15

f) Explain a limitation of the Google PageRank algorithm. [2 marks]

New sites, even those with very useful information, will initially have a low rank because few (if any) existing sites will link to them. As a result of a low rank they will appear lower in search results, meaning few people will visit them. Hence few people will find out about the new site, meaning very hard to get others to link to it.

Question 3 [20 marks]

a) Explain how the Gnutella Peer-to-Peer protocol works, including:
   - How do nodes join the network
   - How does a node search the network for a resource

   You *do not* have to give details of the protocol (e.g. assumptions, message formats, state diagrams), a brief textual description is sufficient. [6 marks]

Initially a node must be pre-configured with at least one other peer when it wants to join the network. It then sends a Ping message to this peer who may respond with a Pong, indicating it is willing to act as a permanent peer. The Ping may also be forward to other peers (in a broadcast fashion). When enough Pongs are received, the new node selects C peers as permanent peers.

When a node searches for a resource it sends a query message to its permanent peers. Each peer forwards the query to its permanent peer in a broadcast fashion until the resource is found. When the resource is found, the nodes returns the result to the initiating node, along the same path as the request/query came.

b) Using the diagram below (which shows a set of nodes and their C=3 permanent peers), answer the following questions (assume the nodes have already joined the network and a node forwarding a message counts as one transmission):
   i. How many times are messages sent in the network using the normal Gnutella protocol if node 1 searches for a resource that is located on nodes 11 and 18 (assume TTL=7)? [2 marks]
   ii. How long does it take for node 1 to receive a reply from part (i) (assume the hop time is 100 milliseconds)? [1 mark]
   iii. How many replies are received by node 1? Explain your answer. [2 marks]
   iv. Explain how the expanding ring search works, including how the number of messages, time for response and number of replies differs from parts (i) to (iii) (you do not have to give exact values, just explain *how* and *why* they would differ in the general case). [5 marks]

   In all parts you must explain any additional assumptions you make.

Part i. 41 messages. This assumes when two nodes have a message to transmit to each other, that one will transmit first and hence the other will refrain from sending. Also transmissions occur in rounds. There are 32 messages of sending queries and node 11 sends a response (3 hops) and node 18 sends a response (6 hops).

If we assumed that if two nodes were to transmit at the same time, then there would be an extra 5 messages: total 46.

Part ii. 6 x 100 = 600ms

Part iii. 2. A response from node 11 and a response from node 18.

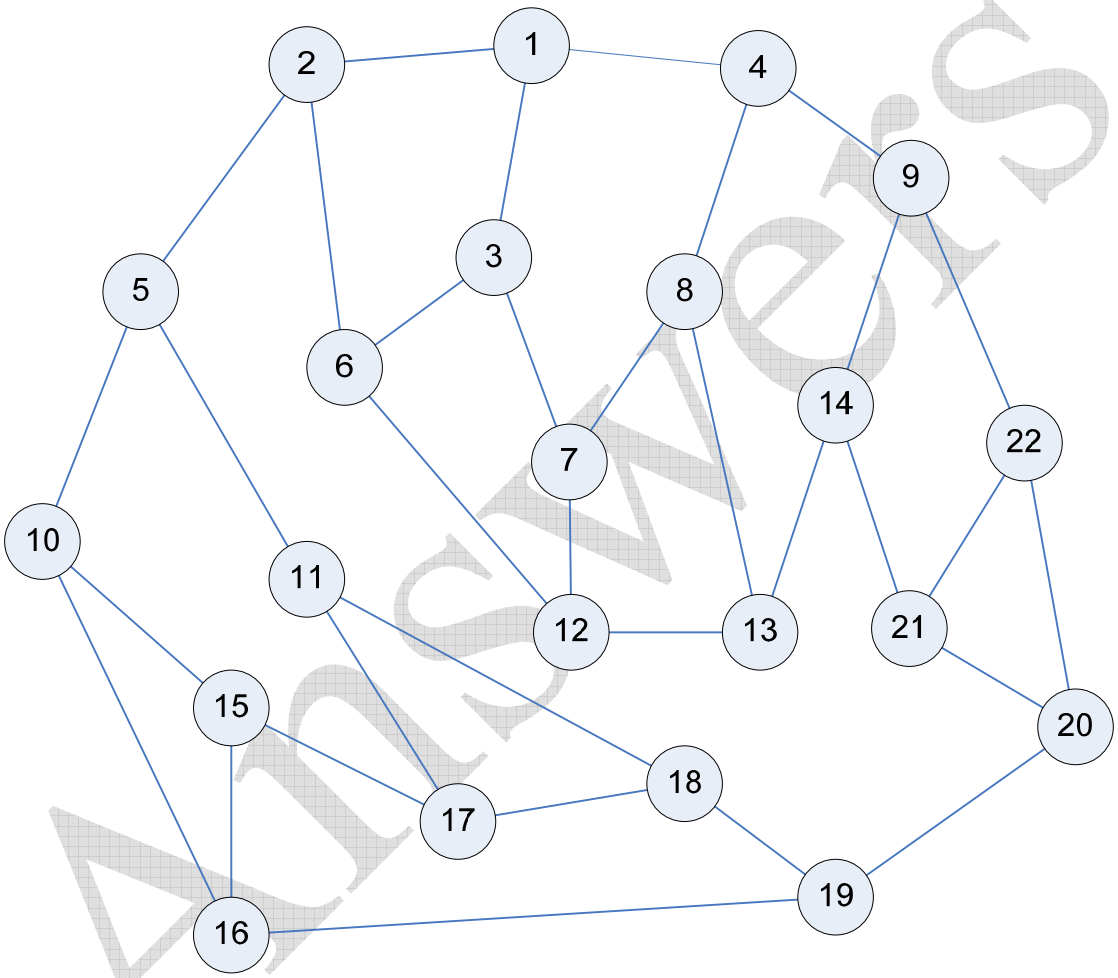
Part iv. Expanding ring first sets TTL=1 so that the broadcast only goes one hop. If no response is received within a timeout, then the query is sent again, but this time with TTL=2. This continues until a response is received or a maximum TTL is used.

Using an expanding ring the number of messages will be less than normal broadcast because when a response is received the messages will not be broadcast any further. E.g. if a response is received from a node 3 hops away, than effectively the TTL=3 in expanding ring. But in normal mode, if

TTL=7 and response is received from 3 hops away the message will still be broadcast up until 7 hops.

The time for response will be more than normal mode because the original node must wait for a timeout for each "ring". E.g. send a packet with TTL=1; wait for no response; send with TTL=2; wait for no response etc.

The Number of replies will be dependant in the expanding ring criteria, but general less than normal mode. If the initiator wants 2 replies than it will stop expanding the ring once it receives the 2 replies. In normal mode, although the initiator only wants 2 replies, because the messages are sent to many more nodes, possibly more replies will be received.



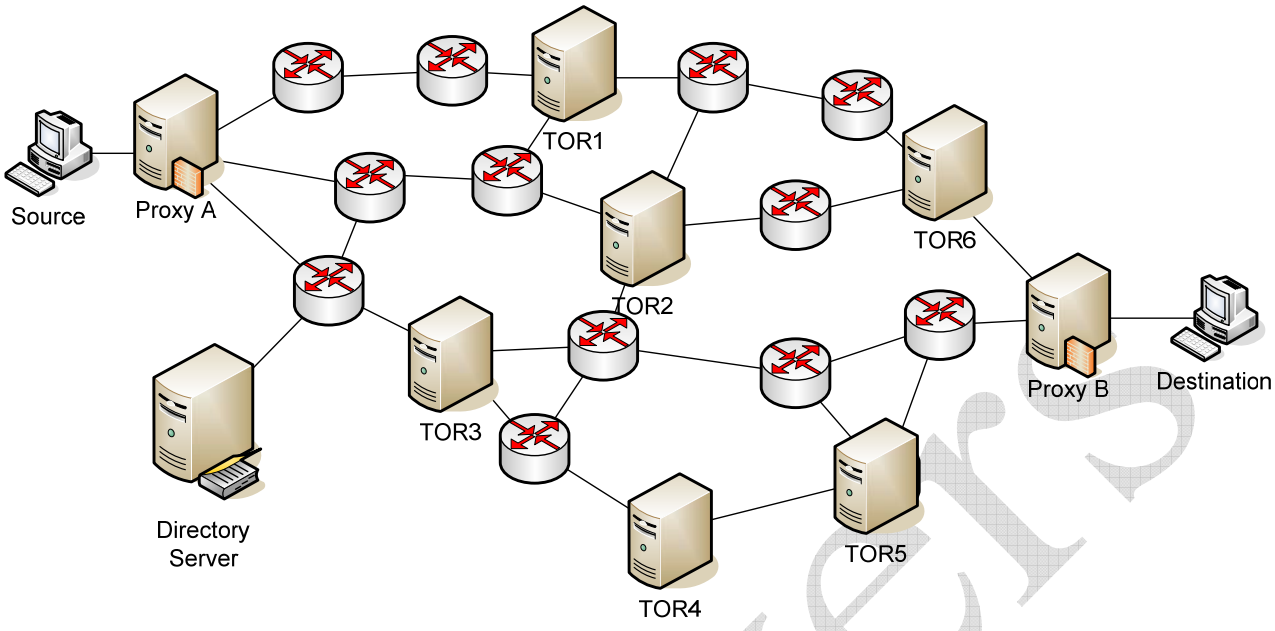
c) How do the following parameters assist in the broadcast (Hint: consider what happens if the parameters were *not* used):
   i. Time to live (TTL) [2 marks]
   ii. Query (or search) identifier [2 marks]

TTL limits the number of times a packet is forwarded so that the packet isn't forward forever (especially if loops exist).

Query ID allows nodes that have already received (and forwarded) a query to node forward it again.

Question 4 [13 marks]

The following diagram shows a TOR network.



a) What is the purpose of the Directory Server? [2 marks]

The Directory Server maintains information about all TOR routers in the network. When establishing a path, the Proxy obtains information about TOR routers in the network to select a path to the destination proxy.

Assume Proxy A selects a path to Proxy B via TOR3, TOR4 and then TOR5.

b) After connection setup is complete, what keys do the following nodes have [3 marks]:
   i. Proxy A
   ii. TOR3
   iii. TOR4
   iv. TOR5
   v. Proxy B

Proxy A: Secret key shared with: TOR3, TOR4, TOR5, and ProxyB
TOR3: Secret key shared with Proxy A
TOR4: Secret key shared with Proxy A
TOR5: Secret key shared with Proxy A
Proxy B: Secret key shared with Proxy A

c) In what order is a packet from Source to Destination encrypted. [2 marks]

Packet is first encrypted with Proxy B's key, then TOR5, then TOR4 then TOR3's key.

d) Explain how the encryption order in part (c) provides anonymity in TOR. [3 marks]

In general, a TOR router only knows the immediate previous router (where it received the packet from) and next immediate router (where it will send the packet to). That is because the original

source and final destination addresses are encrypted with keys such that a router cannot view the information (if more than 1 hop away).

e) Explain why using IPsec in tunneling mode (e.g. between proxy A and proxy B) does not provide similar level of anonymity as TOR. [3 marks]

With IPsec tunneling mode the original source and final destination IP addresses are hidden from any intermediate node because they are encrypted. However the IP address of the tunnel end-points (e.g. Proxy A and Proxy B) are known, and hence an intermediate node can identify: a node on network with proxy A is communicating with a node on network with proxy B. Whereas in TOR, the source and destination networks cannot be identified.

Question 5 [20 marks]

Chord is a protocol that uses Distributed Hash Tables for peer-to-peer applications.

a) Draw a diagram of an example Chord network that has 8 nodes (and can support no more than 8 nodes) [1 mark]

b) Using your example network where necessary, explain:

   i. How are nodes given identifiers that represent their position in the Chord network? [2 marks]

   ii. What is the relationship between keys and resources in Chord? [1 mark]

   iii. What is the relationship between keys and nodes? [2 marks]

   iv. What other nodes does a node maintain routes to? [2 marks]

   v. In addition to the addresses of other nodes, the routing information maintained by a node should also contain what? [2 mark]

Part i. An address of a node (e.g. IP address) is hashed to determine an ID of a node.

Part ii. A resource identifier is hashed to determine a resource key.

Part iii. A key is stored on the node with the same value of ID. If the node does not exist, then the key is stored on the next node in the ring that does exist.

Part iv. A node maintains routes to nodes that are 2n positions away, e.g. 1, 2, 4, 8, …. In the example, node 1 maintains routes to nodes 2, 3, 5. Node 2 maintains routes to nodes 3, 4, 7.

Part v. The keys stored by that node (and as a consequence, the keys that node is responsible to routing to).

Assume nodes 3, 4 and 7 have left your example network (that is, there are only 5 nodes remaining).

c) For each node, list the keys the node stores. [2.5 marks]

Node 1: 1
Node 2: 2
Node 5: 3, 4, 5
Node 6: 6
Node 0: 0, 7

d) For each node, list the other nodes the node maintains routes to. [2.5 marks]

Node 1: 2, 5
Node 2: 5, 6
Node 5: 6, 0, 1
Node 6: 0, 2
Node 0: 1, 2, 5

e) Describe the path taken if node 2 sends a query for resource with key 1. [2 marks]

Node 2 sends the query to node 6. Node 6 sends the query to node 0. Node 0 sends the query to node 1.

f) If Chord was modified such that each node maintained routes to every second node, then explain an advantage and disadvantage compared to the actual Chord protocol. [3 marks]

Advantage: Search would be faster, because a node knows about more nodes (and hence more keys) in the network

Disadvantage: Route maintenance would be much higher, because a node must maintain routes with n/2 nodes