EXAM
TCP/IP NETWORKING
Duration: 3 hours
With Solutions

January 2025

INSTRUCTIONS

1. Verify that you have a quiz, 4 problems + one figure sheet.
2. You can remove the figure sheet from the booklet and use it at your convenience when solving the exercises.
3. Write your solution into this document and return it to us (you do not need to return the figure sheet). You may use additional sheets if needed.
4. Do not forget to write your name on **the quiz and each of the four problem sheets** and **all** additional sheets of your solution.
5. For grading, the justification is as important as the solution itself.
6. If you find that you need to make additional assumptions in order to solve some of the questions, please describe such assumptions explicitly.
7. You can bring and use 4x A4 sheets = 8x AA4 pages of hand-written or type-written notes or the exam booklet that we offer in Moodle (in printed form). You can also use your pocket calculator (i.e. a simple calculator without extra storage or graph plotter).

Good luck!

QUIZ (16 PTS)

For each question, please circle a single best answer.

1. **(1 pt)** The Internet, in general, relies on packet switching (PS) as opposed to "relaying entire messages". This is because PS:
   (a) results in lower bit error rates.
   (b) reduces the buffer size required in routers and most of the time the end-to-end overall delay. *Correct*
   (c) can provide better security and privacy guarantees.
   (d) This is a trick question, the Internet relies on "relaying entire messages."

2. **(1 pt)** You type in your web client `http://www.epfl.ch` and you get an "Unable to connect to the Internet" error. Then you type `http://128.178.50.12` (which is the IP address of `www.epfl.ch`) and you get the right response. The reason for the error may be:
   (a) The link that connects your computer to the Internet is down.
   (b) The web server `www.epfl.ch` is down.
   (c) The local DNS server that you are using is down. *Correct*
   (d) You need to use IPv6 instead.

3. **(1 pt)** CSMA/CA (collision avoidance) is a randomized medium-access control protocol that:
   (a) is expected to improve the overall throughput compared to deterministic protocols, whenever few users share the same cable and have lots of data to transmit, always at the same time.
   (b) is used to avoid collisions by forcing the sources to back-off for a fixed amount of time in case a collision is detected.
   (c) is used by WiFi, to avoid collisions in the presence of a hidden terminal. *Correct*
   (d) is used by WiFi, only when users are close to each other and there is no hidden terminal.

4. **(1 pt)** In Switched Ethernet:
   (a) CSMA/CD is used to detect and manage collisions.
   (b) the hosts run a more elaborate Ethernet protocol than historical Ethernet.
   (c) half-duplex cables are used to increase the overall throughput.
   (d) switches forward frames to a single interface if an exact match exists in their forwarding table. *Correct*

5. **(1 pt)** The destination MAC address of a frame that encapsulates a Neighbor Advertisement (NA) packet is *always* equal to
   (a) the broadcast MAC address `ff:ff:ff:ff:ff:ff`.
   (b) a multicast MAC address, where the first 16 bits are `33:33` (in hex format) and the last 32 bits are copied from the Solicited Node Multicast Address.
   (c) the source MAC address of the host which asked for the neighbor solicitation and to which the NA is destined to. *Correct*

(d) the MAC address of the internal interface of the gateway router.

6. **(1 pt)** How many neighbor solicitation (NS) will be initiated by a link-layer switch whenever it has to forward a packet?

   (a) Exactly 0. *Correct*

   (b) Exactly 1: NS for the destination MAC address.

   (c) Exactly 2: NS for the destination MAC address and source MAC address.

   (d) We cannot say exactly, it all depends on the contents of its neighbor cache (or ARP table).

7. **(1 pt)** A disadvantage of the Spanning Tree Protocol (STP) is that

   (a) the cost that it assumes for the links is a decreasing function of their bit rate.

   (b) it uses a variant of the Bellman-Ford algorithm to compute the shortest paths, as opposed Dijkstra, which always identifies shorter paths.

   (c) a direct link between two (non-root) switches may not be used, even if it is optimal. *Correct*

   (d) this is a trick question: STP has no disadvantages.

8. **(1 pt)** The TCP retransmission timeout:

   (a) should be smaller than the round-trip time (RTT).

   (b) is calculated using the current estimate of the smoothed RTT and the RTT variability. *Correct*

   (c) depends on the size of the buffer at the receiver.

   (d) is calculated at the receiver side and sent back to the sender.

9. **(1 pt)** A TCP sender sents a series of segments to a TCP receiver. Which mechanism does the receiver rely on, in order to identify whether one of these segments was lost?

   (a) Fast retransmit (i.e. a certain number of duplicate acknowledgements is received).

   (b) A timeout event (i.e. either RTO or PTO expires).

   (c) Both of the above.

   (d) None of the above. *Correct*

10. **(1 pt)** Suppose you access a small web page that fits in one packet, by using HTTP from your browser. The round-trip (RTT) time between your host and the web server is constant. How many RTTs will it take, *at least*, between the moment when you finish typing the URL in the browser and the moment when the last bit of the web page has been received by your HTTP client?

    (a) 1 RTT. *(Correct: TCP Fast Open is used and you have visited the web page in the past)*

    (b) 2 RTTs.

    (c) 3 RTTs.

    (d) 4 RTTs.

11. **(1 pt)** Alice and Bob communicate over an established TLS connection.
    When Alice sends a message that fits into one segment:

    (a) Nobody can read and understand the message other than Bob.

    (b) Bob can verify that the message is indeed coming from Alice.

    (c) Bob can verify that the message was not altered by an adversary.

(d) All of the above. *Correct*

12. **(1 pt)** Software Defined Networking (SDN)...

    (a) uses deep packet inspection and per-flow forwarding to control routing (i.e. the paths followed by the packets). *Correct*

    (b) uses software routers that typically forward packets based on destination IP addresses and the longest-prefix-match principle, at a higher rate than hardware routers.

    (c) is mainly used at the link layer to update the software of the switches.

    (d) should never be used with other routing algorithms; as it is more efficient.

13. **(1 pt)** In which multicast protocol, the destination hosts have to subscribe to the multicast group by sending JOIN messages to their gateway routers via the Internet Group Management Protocol (IGMP) or Multicast Listener Discovery (MLD)?

    (a) Only in Protocol Independent Multicast (PIM).

    (b) Only in Bit Index Explicit Replication (BIER).

    (c) Only in PIM, and whenever source-specific multicast (SSM) is used.

    (d) In both PIM and BIER. *Correct*

14. **(1 pt)** In a network with UDP and TCP flows, all TCP flows use TCP Reno and ECN is not supported. We decide to replace Reno with DCTCP (which always uses ECN) and update all routers to support ECN. With this change, we expect that...

    (a) no packet loss will be ever observed.

    (b) packet loss due to congestion in routers will be reduced. *Correct*

    (c) packet loss due to congestion in both switches and routers will be reduced.

    (d) loss due to packet corruption will be reduced.

15. **(1 pt)** A domain network is composed of border and backbone routers. Each border router connects to external domains and at least one backbone router. The backbone routers are not directly connected to any external networks and form a full mesh with each other. The network supports only IPv6 with addresses in the block 2001:620:618::/48. All routers run OSPF, and all border routers run BGP. BGP is redistributed into OSPF, but OSPF is not redistributed into BGP. In order to make sure that all internal prefixes of the domain are reachable from hosts in other domains, the network operator must:

    (a) run a Network Address Translation function (NAT) at each and every border router.

    (b) manually configure the border routers to originate all of the internal prefixes ("Static configuration"). *Correct*

    (c) configure redistribution of all directly attached subnets ("Redistribute-connected").

    (d) Both (a) and (b), taking only one of the two actions does not work.

16. **(1 pts)** A tier-1 ISP B does not want to carry transit traffic between two other tier-1 ISPs, say A and C, with which B has peering agreements. To implement this policy, ISP B:

    (a) should *not* advertise to A routes that pass through C;
    and should *not* advertise to C routes that pass through A. *Correct*

    (b) should *not* advertise to A routes that pass through C;
    but should advertise to C routes that pass through A.

    (c) should advertise to A routes that pass through C;
    but should *not* advertise to C routes that pass through A.

    (d) should advertize to A routes that pass through C;
    and should advertise to C routes that pass through A.

PROBLEM 1 (23 PTS)

Consider the network for Problem 1 in the figure sheet. H1, H2, H3, H4, and H5 are hosts and DNS is a local DNS server serving hosts in its subnet. R1 and R2 are routers (Layer 3). Each router runs a seperate DHCP relay function to allocate IPv4 addresses within each subnet it is connected to. IPv6 addresses are allocated using stateless SLAAC. S1, S2, S3, S4, and S5 are switches (Layer 2). N is an IPv4 NAT and an IPv4/IPv6 router. O1 and O2 are observation points. All machines are dual-stack. All links are full-duplex Ethernet. We assume that all machines are correctly configured (unless otherwise specified), proxy ARP is not used and there is no VLAN. **The hosts, switches and routers have been running for some time, their different protocols have converged and the forwarding tables of all routers and switches are in their final state.** There is no other system or interface than those shown on the figure.

In the following questions, use interface letters to represent MAC addresses: for example, "S1w", "R2e", etc, or the names of hosts themselves, e.g., "H1". Do the same to represent IP addresses that are not explicitly given.

Question 1 (9 pts):

1. **(2 pts)** Write the two IPv6 addresses of H1 in uncompressed format.

   Solution.

   (a) fe80:0000:0000:0000:0000:0000:0000:0100

   (b) 3300:000a:000b:0000:0000:0000:0000:0011

2. **(3 pts)** Given the IPv4 addresses in the figure, what are the minimum and maximum possible IPv4 subnet mask lengths for hosts H3, H4, and H5?

Host	Minimum	Maximum
H3	17	28
H4	17	25
H5	17	25

   Solution.

   The maximums are determined by the range of allocated IPs per subnet. For example, in the subnet of H3, we see 2 allocated IPs: 192.168.16.1 and 192.168.16.13. Therefore, the host portion must be at least 4 bits, which allows 16 IP addresses to allocated.

   The minimums are determined by the constraint that different subnets should have different subnet prefixes.

   Focusing on non-common last 16 bits of each relevant subnet, we have:
   `x.x.00010000 00000000` (H3)
   `x.x.10000000 00000000` (H4, H5)

   Therefore, the minimum subnet mask such that the subnets are distinct is 17 bits.

3. **(4 pts)** Assuming that all IPv6 subnets shown in the figure have the same IPv6 prefix length, which of the following prefix lengths are valid?

Proposed length of common IPv6 subnet mask	valid	invalid
48	☐	☒
56	☒	☐
64	☒	☐
80	☒	☐

Solution.

The expanded public IPv6 addresses shown in the figure are:

```
1.  3300:000a:000b:0000:... (H1, H2)
2.  3300:000a:000b:ffff:... (H3)
3.  3300:000a:000c:0000:... (H4, H5, DNS)
--------16---32---48---64
```

Using a common 48 bit mask means that subnets 1. and 2. have a common prefix (invalid).
Using a common 56 bit mask makes each prefix distinct.
The same applies for 64 and 80.

Question 2 (3 pts): For this question only, assume that there was a power outage and **all nodes** in the topology (hosts, switches, routers, etc) lost all their configuration information. Further, assume that DHCP relay functions at router R2 did not restart properly and are not responsive. Which of the following actions will be successful after all other devices have been restored?
Hint: Recall a DHCP relay function is similar to a DHCP server within a LAN.

Action	Success	Failure
H4 sends an IPv6 packet to H5	☒	☐
H3 sends an IPv4 packet to H4	☐	☒
H3 sends an IPv6 packet to H5	☒	☐

Solution.

H4 can send an IPv6 packet to H5 using link-local addresses.
H3 cannot send an IPv4 packet to H3 because none of the two have obtained routable (non link local) IPv4 addresses.
H3 can send an IPv6 packet to H5 because IPv6 addresses are configured using stateless SLAAC which is unrelated to the non-responsive DHCP relays.

Question 3 (7 pts):

1. **(4 pts)** H4 is interacting with a website hosted at H1 using HTTP and IPv4. We observe the packet headers at the MAC, IP and transport layer at points O1 and O2.

   Give possible header field values in the tables below for the given direction. You can use given variable names when the specific value is not known (e.g., "Ns", "H4"). Assume that the client application is using any port number between 50000 and 50100.

Direction from H4 to H1						
	MAC addresses		IPv4 addresses		Port Numbers	
At	src	dst	src	dst	src	dst
O1	H4	R2e	192.168.128.14	200.0.0.11	50000	80
O2	Nn	R1s	Nn	200.0.0.11	a number $>$ 1024 (e.g. 20000)	80

Direction from H1 to H4						
	MAC addresses		IPv4 addresses		Port Numbers	
At	src	dst	src	dst	src	dst
O1	R2e	H4	200.0.0.11	192.168.128.14	80	50000
O2	R1s	Nn	200.0.0.11	Nn	80	The port $>$ 1024 the NAT used

2. **(3 pts)** Fill in the forwarding table of switch S3 below, assuming that both H4 and H5 have recently accessed `www.epfl.ch`. You can use given variable names when the specific value is not known

(e.g., “Ns”, “H4”).
How does S3 know this information? Does the table-filling mechanism work well without the Spanning Tree Protocol? Explain.

MAC Address	Port
R2e	S3w
H4	S3n
H5	S3ne
DNS	S3ne

Note: the table above assumes that the S3-S5 link was disabled by the spanning tree protocol. S3-S4 or S4-S5 could have been disabled alternatively.

S3 maps MAC addresses to output interfaces via MAC Learning: the switch dynamically creates the mappings by observing the source MAC address of each incoming packet and mapping it to the interface on which it is received. Without STA, the presence of a loop in the topology can cause packets to loop indefinitely which can cause timeouts and congestion.

Question 4 (4 pts): Suppose that suddenly H5 reboots and its caches become empty. At time t_1, H4 is compromised by an adversary. At time $t_2 > t_1$, H5 tries to configure its IPv4 interface and access `www.epfl.ch`, using IPv4. Briefly describe at least **two** ways in which the adversary can prevent H5 from getting any content from `www.epfl.ch`.

Solution.

Here are many possible attacks:

Option 1: H4 sends a fake DCHP reply responding to the DHCP discovery broadcast of H5 and giving itself as the default gateway.

Option 2: H4 responds to the ARP request H5 makes to learn the MAC address of the local DNS server and responds as if it is a DNS server. It then proceeds to send a fake IP address for www.epfl.ch.

Option 3: H4 responds to the ARP request H5 makes to learn the MAC address of R2e and responds as if it is the gateway to the subnet. It then drops all packets H5 sends it.

PROBLEM 2 (18 PTS)

In the following questions, we assume that the BGP decision process uses the following criteria in decreasing order of priority.

1. Highest LOCAL-PREF.
2. Shortest AS-PATH.
3. E-BGP is preferred over I-BGP.
4. Shortest path to NEXT-HOP, according to IGP.
5. Lowest BGP identifier of sender of route is preferred; the comparison is lexicographic, with A<B<C<D and $1 < 2$; for example A1 is preferred over A2, A2 is preferred over B1, etc.

Furthermore, **unless otherwise specified**:

- When receiving an E-BGP announcement, every BGP routers tags it with LOCAL-PREF = 0. No other optional BGP attribute (such as MED, etc.) is used in BGP messages.
- No aggregation of route prefixes is performed by BGP.
- The policy in all ASs is that all available routes are accepted and propagated to neighbouring ASs, as long as the rules of BGP allow it.
- Every router redistributes internal OSPF destinations into BGP.
- Every router performs recursive forwarding-table lookup.
- No confederation or route reflector is used.

In the following questions, you will need to fill in the local RIBs (route information bases) of various routers. Do it as follows:

- If you think a prefix is not in the local database, cross the line.
- If you think a prefix is missing, you can add it at the end of the table.
- If you think a prefix has several entries, write them one on top of the other as shown in the example below.
- If you run out of space, you can add extra lines at the end of the table.

At Example :				
Destination Network	From BGP Peer	Next-Hop	AS-PATH	Best route ?
X.X.X.X/Y	Z1	z1w	ASN ASM	yes
	Z2	z2w	ASM	no
Justification: Briefly explain.				

Question 1 (10 pts): Consider the network for Problem 2 in the figure sheet. There are six ASs (AS1 to AS6), with border routers (A1, A2, B1, B2, B3, C1, etc). In each domain, there is an **I-BGP mesh** that is not shown in the figure. $AS2$ has set its LOCAL-PREF as shown in the figure. **Consider the situation after BGP has converged**. Justify each answer.

1. **(4 pts)** List the route(s) in the local RIBs of Router $A1$.

At $A1$:				
Destination Network	From BGP Peer	Next-Hop	AS-PATH	Best route ?
110.68.0.0/13	D1	d1s	AS4	yes
110.68.0.0/13	A2	b1w	AS2 AS3 AS4	no
110.72.0.0/14	A2	b1w	AS2	yes
110.72.0.0/14	D1	d1s	AS4 AS3 AS2	no
110.74.0.0/15	D1	d1s	AS4 AS3 AS2 AS5	no
110.74.0.0/15	A2	b1w	AS2 AS5	yes
110.76.0.0/14	D1	d1s	AS4 AS3 AS2 AS6	no
110.76.0.0/14	A2	b1w	AS2 AS6	yes
110.84.0.0/15	-	-	-	-
110.88.0.0/13	D1	d1s	AS4 AS3	yes
110.88.0.0/13	A2	b1w	AS2 AS3	no
Justification: 110.88.0.0/13 has two routes with the same AS-path length, the route learned from $D1$ is preferred because it comes from eBGP.				

2. **(4 pts)** List all the routes in the local RIBs of Router $B3$.

At $B3$:				
Destination Network	From BGP Peer	Next-Hop	AS-PATH	Best route ?
110.68.0.0/13	B2	c2s	AS3 AS4	no
110.68.0.0/13	B1	a2e	AS1 AS4	yes
110.72.0.0/14	-	-	-	-
110.74.0.0/15	E1	e1w	AS5	yes
110.76.0.0/14	F1	f1w	AS6	yes
110.84.0.0/15	B1	a2e	AS1	yes
110.84.0.0/15	B2	c2s	AS3 AS4 AS1	no
110.88.0.0/13	B2	c2s	AS3	no
110.88.0.0/13	B1	a2e	AS1 AS4 AS3	yes
Justification: 110.68.0.0/13 has two routes, but AS2 has set LOCAL-PREF: routes from $AS1$ have higher value, so the route through $B1$ is chosen. 110.88.0.0/13 also has two routes: routes from $AS1$ have higher LOCAL-PREF value, so they are chosen even if the AS-path is longer.				

3. **(2 pts)** Suppose that there exists a malicious AS, which is a customer of AS3 (i.e. a stub domain served by AS3), starts sending bogus announcements for prefix 110.75.0.0/16.

   (a) How do these announcements change the local RIB of Router $A1$? List only the changes.
   (b) What is the AS-path followed by packets going from 110.84.1.1 to 110.75.1.1?

(a)

At $A1$:				
Destination Network	From BGP Peer	Next-Hop	AS-PATH	Best route ?
110.75.0.0/16	D1	d1s	AS4 AS3	yes
110.75.0.0/16	A2	b1w	AS2 AS3	no
Justification: The prefix is advertised by $C1$ and $C2$, so it will have two routes: as both have the same AS-path length, $A1$ will select the one learned by E-BGP.				

(b) 110.75.1.1 is contained **both** in the bogus prefix 110.75.0.0/16 and in the prefix 110.74.0.0/15 advertised by $AS5$. As the prefix match is longer for the bogus announcement, the path is AS1 - AS4 - AS3.

Question 2 (6 pts): Consider again the same network as before (Problem 2 in the figure sheet without the malicious AS). $AS2$ has set its LOCAL-PREF as shown in the figure. **For the purposes of this question, aggregation is used by BGP and whenever possible ASs also aggregate internally-originated prefixes with prefixes that are learnt from E-BGP.** Consider the situation after BGP has converged. **Justify each answer**.

1. **(4 pts)** List all the routes in the local RIBs of Router $A1$.

At $A1$:				
Destination Network	From BGP Peer	Next-Hop	AS-PATH	Best route ?
110.68.0.0/13	D1	d1s	AS4	yes
110.68.0.0/13	A2	b1w	AS2 AS3 AS4	no
110.72.0.0/14	-	-	-	-
110.74.0.0/15	-	-	-	-
110.76.0.0/14	-	-	-	-
110.84.0.0/15	-	-	-	-
110.88.0.0/13	D1	d1s	AS4 AS3	yes
110.88.0.0/13	A2	b1w	AS2 AS3	no
110.72.0.0/13	A2	b1w	AS2	yes
110.72.0.0/13	D1	d1s	AS4 AS3 AS2	no
Justification: Aggregated prefixes are in bold. The route to 110.88.0.0/13 is chosen because of E-BGP preference.				

2. **(2 pts)** As previously, suppose that there exists a malicious AS, which is a customer of AS3 (i.e. a stub domain served by AS3), starts sending bogus announcements for prefix 110.75.0.0/16.
   (a) How do these announcements change the local RIB of Router $A1$? List only the changes.
   (b) What is the AS-path followed by packets going from 110.84.1.1 to 110.75.1.1?

   (a)

At $A1$:				
Destination Network	From BGP Peer	Next-Hop	AS-PATH	Best route ?
Justification: No new entry because 110.75.0.0/16 is aggregated with the prefix 110.72.0.0/13 already advertised by $AS3$.				

   (b) 110.75.1.1 is contained in the aggregated prefix 110.72.0.0/13, so the packet is sent to $AS2$ through $B1$ (shortest AS-path). At $AS2$, there are two routes for the packet: 110.74.0.0/15 through $E1$ and 110.75.0.0/16 through $C2$; the route to 110.75.0.0/16 is selected due to longest prefix match. So, the AS-path is AS1 - AS2 - AS3.

Question 3 (2 pts): Consider the same network as before (Problem 2 in the figure sheet). In each domain, there is an **I-BGP mesh** that is not shown in the figure. **For the purposes of this question, aggregation is used by BGP.** Suppose that some of the ASs on the figure besides $AS2$ set their LOCAL-PREF as follows:

- For $AS1$: from $AS2$: 40; from $AS4$: 80.
- For $AS3$: from $AS2$: 80; from $AS4$: 40.
- For $AS4$: from $AS1$: 40; from $AS3$: 80.

Assuming that no link goes down and all devices always work properly, will all routers converge to a stable RIB? Why? If yes, how would the RIBs from Question2 change? If no, modify the LOCAL-PREF values of a single AS to make the RIBs stable.

No, the routes will oscillate because there are circular dependencies in the preferences. To fix this issue, we can choose any of the ASs and change/remove the preferences (e.g. for $AS1$, we can change the preferences to $AS2$: 100, $AS4$: 100).

PROBLEM 3 (22 PTS)

Consider the network for Problem 3 in the figure sheet.

- R1, R2, R3, and R4 are routers, connected via point-to-point links. S1, S2, S3, and S4 are servers connected to the routers.
- The following table describes the flows in the network:

Flow	Source	Destination	Path
F1	S1	S3	R1 → R2 → R3
F2	S1	S4	R1 → R4
F3	S2	S4	R2 → R1 → R4
F4	S2	S4	R2 → R3 → R4
F5	S3	S4	R3 → R4

- All flows are unidirectional flows (as indicated by the arrows). There is no other system and no other flow than those shown in the figure.
- The capacity of the links are as follows: all links between routers are 5 Gb/s, all links between routers and servers are 12 Gb/s **except for** the link between R3 and S3 which is 1 Gb/s. The links are full duplex with the same rate in both directions.
- We neglect the impact of the acknowledgement flows in the reverse direction.
- We also neglect all overheads and assume that the link capacities can be fully utilized at bottlenecks.
- For i = 1, 2, 3, 4, and 5 we call f_i the rate of the flow F_i.

Question 1 (6 pts): Assume the rates are allocated by some central bandwidth manager according to max-min fairness. Which are the rates achieved by each flow? Justify.

Using water-filling algorithm:
Step 1: We maximize t such that $f_1 = f_2 = f_3 = f_4 = f_5 = t$ and all constraints are satisfied; We find $t = 1$ Gb/s, hence $f_1 = f_2 = f_3 = f_4 = f_5 = 1$ Gb/s; The link between R_3 and S_3 is saturated and thus F_1, and F_5 freeze at 1 Gb/s.
Step 2: We maximize t such that $f_2 = f_3 = f_4 = t$ with $f_1 = 1$ and $f_5 = 1$ Gb/s and all constraints are satisfied; We find $t = 2.5$ Gb/s, hence $f_2 = f_3 = f_4 = 2.5$ Gb/s; The link between R_1 and R_4 is saturated and thus F_2 and F_3 freezes at 2.5 Gb/s.
Step 3: We maximize t such that $f_4 = t$ with $f_1 = 1$, $f_2 = 2.5$, $f_3 = 2.5$, and $f_5 = 1$ Gb/s and all constraints are satisfied; We find $t = 4$ Gb/s, hence $f_4 = 4$ Gb/s; The link between R_3 and R_4 is saturated and thus F_4 freezes at 4 Gb/s.

All flows are frozen thus the allocated rates are $f_1 = 1, f_2 = 2.5, f_3 = 2.5, f_4 = 4$, and $f_5 = 1$ Gb/s.

Question 2 (3 pts): Assume the rates are allocated by some central bandwidth manager, which of the following rate allocations are Pareto-efficient? Justify.

1. $f_1 = 0, f_2 = 5, f_3 = 0, f_4 = 5, f_5 = 0$
2. $f_1 = 1, f_2 = 4.75, f_3 = 0.25, f_4 = 4, f_5 = 0$
3. $f_1 = 1, f_2 = 1, f_3 = 5, f_4 = 4, f_5 = 1$

1. The rate allocation is Pareto-efficient since the link constraints are satisfied with equality so it is not possible to unilaterally increase any rate.
2. The rate allocation is not Pareto-efficient since we can increase the rate of F_5 without decreasing the rate of any other flow.
3. The rate allocation is not Pareto-efficient, becasue it is not a feasible allocation.

Question 3 (3 pts): Show why the following rate allocation is not proportionally fair: $f_1 = 0.75, f_2 = 0.75, f_3 = 4.25, f_4 = 4.25, f_5 = 0.75$ Gb/s.

If we increase f_1 by $0 < \delta < 0.25$, then f_4 will decrease by δ. Also the rate of f_5 will increase by δ. The total relative change is: $\sum_i \frac{x_i^{'} - x_i}{x_i} = \frac{\delta}{0.75} + \frac{\delta}{0.75} - \frac{\delta}{4.25} = 26.43 * \delta > 0$; Thus, the allocation is not proportionally fair.

Question 4 (6 pts): In this question, **flow F_1 is using UDP** and sends data at a constant rate of 1 Gb/s. Assume that all other flows use **TCP Reno** and are long-running TCP flows. Also assume that all routers use RED queuing with ECN. The round trip times (RTTs) for TCP flows is the following: $RTT_{F_2} = 300ms, RTT_{F_3} = 500ms, RTT_{F_4} = 500, RTT_{F_5} = 400ms$; these RTTs include all processing times. All flows use the same MSS and the offered window is very large.

What are the rates attained by each flow in the long run?

The UDP source does not adapt its rate. TCP flows adapt to use the leftover capacity at every link.

TCP Reno loss throughput is:

$$r_{reno} = \frac{MSS \times 1.22}{RTT \times \sqrt{q}}$$

where q is the percentage of the ECN-marked packets, which we can assume is the same for both F_2 and F_3 as the goes through the same bottleneck link. Therefore for we have $\frac{F_2}{F_3} = \frac{RTT_{F_3}}{RTT_{F_2}} = \frac{5}{3}$. Since TCP Reno allocations are Pareto-efficient $f_2 + f_3 = 5$ Gb/s, as a result $f_3 = 1.875$ Gb/s and $f_2 = 3.125$ Gb/s.

q is also the same for both F_4 and F_5 as the goes through the same bottleneck link. On the other hand we have $\frac{F_4}{F_5} = \frac{RTT_{F_5}}{RTT_{F_4}} = \frac{4}{5}$. Since TCP Reno allocations are Pareto-efficient $f_3 + f_4 = 5$ and with respect to the capacity constraint on the link between R3 and S3, as a result $f_4 = 1$ Gb/s and $f_5 = 4$ Gb/s.

Question 5 (4 pts): In this question we have the same setting as in Question 4 but **flows F2 to F5 use TCP Cubic instead of Reno** as their congestion control algorithm. Which of flows F_2 and F_3 may experience a change in their rates and why?

As we have large RTTs and large throughput, TCP Cubic behaves in the region where it is less sensitive to RTT than Reno (see the slide in the exam booklet). Therefore we expect the rates of F_2 and F_3 to be more similar than with TCP Reno, meaning that F_3 will get more throughput than with TCP Reno but F_2 will get less throughput than with TCP Reno due to the capacity constraints.

PROBLEM 4 (21 PTS)

Question 1 (7 pts):

Consider the multi-area OSPF enterprise network of the figure for Problem 4 in the figure sheet. All lines represent physical links and the associated numbers/weights are the OSPF costs. All boxes are routers; Area 0 is the backbone area. Router A_i, (with $i \in \{1, \ldots, 4\}$) is border router between Areas i and the backbone area. All depicted subnets (of the form 10.0.X.0/24) are stub networks. All network interfaces have been correctly configured and OSPF has converged.

Suppose that in each of 10.0.0.0/24 and 10.0.1.0/24 networks, there exist 250 power measurement units with limited resources that report a power measurement (i.e. send a packet of 1 byte) per msec to three different monitors in networks 10.0.15.0/24 (Area 2), 10.0.23.0/24 (Area 3) and 10.0.31.0/24 (Area 4), by using UDP. All other networks contain a large number of hosts and servers that communicate over TCP connections.

As the network administrator, you have to enable source-specific multicasting (SSM); you prefer BIER and use your SDN controller as the mutlicast flow overlay (because your network happens to support SDN).

1. **(3 pts)** What would be the BIER forwarding table entries for each multicast group at router $A1$? You can use bit or set notation. For each multicast group the BIER forwarding table will be the same. There are only

3 destination BREFs. So:

Destination BFER	Forwarding Bit Mask	Next Hop
15	{8,9,10,11,12,13,14,15,16,24,25,26,27,28,29,30,31}	A2
23	{16,17,18,19,20,21,22,23}	A0
31	{8,9,10,11,12,13,14,15,16,24,25,26,27,28,29,30,31}	A2

2. **(2 pts)** Reason about your choice of BIER in this setting. Is it the best option? Briefly explain why.

Yes, because you have to enable SSM, and the number of sources is large (500). So, we would like to avoid keeping state at the backbone routers for such a large number of multicast groups.

3. **(2 pts)** Area 0 is crucial for the availability of the entire network, but it seems rather unprotected from congestion collapse events, because a malfunctioning sensor could easily overwhelm router $A1$ by sending lots of traffic. Sketch a solution to avoid this problem and briefly explain why your solution works. You may add any functionality/device in the current topology to achieve your goal.

At least two solutions are possible here:

1. *Per-class (or class-based) queuing at* $A1$*: the router will classify packets (using an access list) into at least two different classes—one for the UDP traffic of the mulicast groups and another one for all other (TCP) traffic; each class will be guaranteed a dedicated queue and a weight — hence a limited rate. For example, in our setting, the guaranteed rate of the UDP traffic should be a bit larger than* $500*8/(10^{-3}) = 4$*Mbits, which is the aggregate flow rate of all UDP traffic coming from the sensors. Hence, even a malfunctioning sensor could not overwhelm* $A1$*.*
2. *SDN with deep packet inspection: we can dynamically inspect the arriving traffic at* $A1$ *from the two sensor networks and send it to a scrubber whenever the arrival rate largely exceeds 4Mbits. Redirecting traffic to a particular device can happen by correctly configuring the Flow Table of router* $A1$*.*

Question 2 (11 pts):

Consider the following scenario: A client is connected to a server via a single TCP connection, which uses a simplified AIMD congestion-control algorithm, where the congestion window size is measured in number of segments, not in bytes; in additive increase, the congestion window size increases by one segment in each RTT, and in multiplicative decrease, the congestion window size decreases by half (if the result is not an integer, round down to the nearest integer).

Assume that:

- The connection uses a 10 Mbps link which **does not buffer any data (i.e. there is no buffer on the link)**, and this link is the only bottleneck link between the two communicating hosts.
- The server sends a huge file to the client, and the client's receive buffer is much larger than the congestion window.
- Each TCP segment size is 1,500 bytes.
- The RTT of this connection is 150 msec.
- TCP connection is **always in congestion avoidance** phase; slow start and fast recovery are not used.

1. **(3 pts)** What is the maximum window size (in segments) that this TCP connection can achieve?
   Let W denote the max window size measured in segments. Then, $W \cdot \frac{MSS}{RTT} \leq 10$*Mbps, as packets will be dropped if the maximum sending rate exceeds the link capacity. Thus, we have* $W \cdot \frac{1500 \cdot 8}{0.15} \leq 10^7 \Rightarrow W = 125$ *segments.*
2. **(2 pts)** What is the average window size (in segments) and average throughput (in bps) of this TCP connection?
   As congestion window size varies from $W/2$ *to* W*, then the average window size is* $0.75W = 94$ *(ceiling of 93.75) segments. Hence, the average throughput is 94*1500*8/0.15 =7.52Mbps.*
3. **(2 pts)** How long would it take for this TCP connection to reach its maximum window again after recovering from a packet loss?
   When there is a packet loss, W *becomes* $W/2$*, i.e., 125/2=62. So, it will take (125 - 62) *0.15 = 9.45 seconds, because the number of RTTs (that this TCP connection needs in order to increase its window size from 62 to 125) is 63. Recall the window size increases by one in each RTT.*
4. **(4 pts)** Replace the 10 Mbps link with a 10 Gbps link.
   (a) **(1 pts)** How long it would take for this TCP connection to reach its maximum window again after recovering from a packet loss?
   (b) **(3 pts)** Note that in your answer to the last part, you will realize that it takes a long time for the congestion window size to reach its maximum window size after recovering from a packet loss. Sketch a solution to solve this problem.

   Let W denote the max window size. Then, $W \cdot MSS/RTT = 10$*Gbps, as packets will be dropped if maximum sending rate reaches link capacity. Thus, we have* $W \cdot 1500 \cdot 8/0.15 = 10 \cdot 10^9$*, then W= 125000 segments. As congestion window size varies from W/2 to W, the time to reach the max window value again is* $125000/2 \cdot 0.15/60 = 156.25$ *mins.*
   We are eventually in a case of a long-fat network. So, using TCP Cubic instead of naive AIMD would solve this problem. Essentially, to speed up the window increase process, we need to increase the window size by a much larger value (at least until we reach W again), instead of increasing window size only by one in each RTT. Conversely, we could use a multiplicative decrease with a smaller factor, say 0.3 *instead of* 0.5*.*

Question 3 (3 pts):

In our discussion of TCP congestion control in the class, we implicitly assumed that the TCP sender always had data to send. Consider now the case that the TCP sender sends a large amount of data and then goes idle

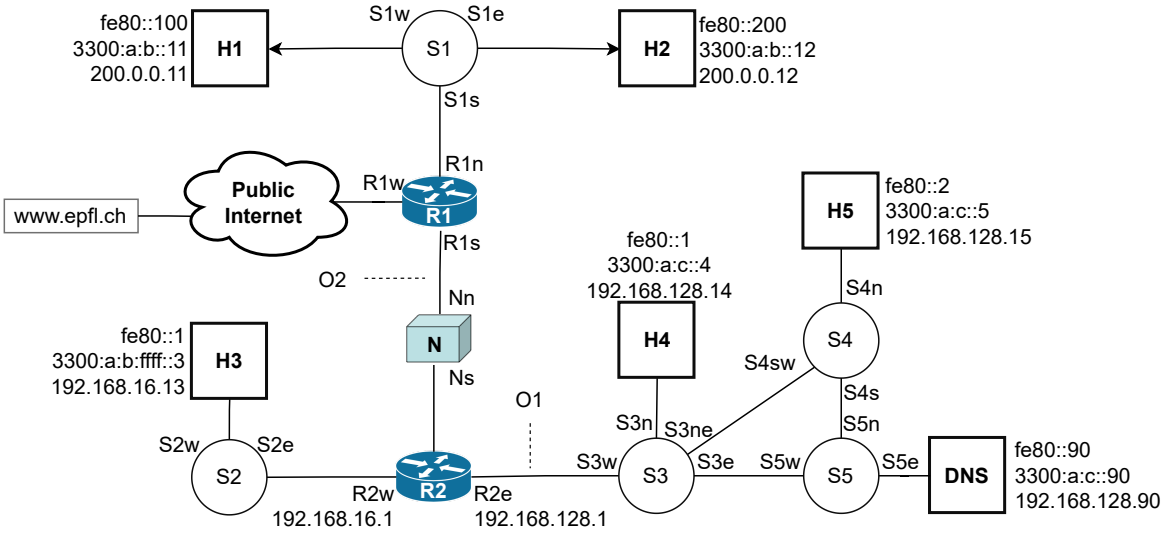
(since it has no more data to send) at time t_1. TCP remains idle for a relatively long period of time and then wants to send more data at time t_2. What is an advantage and a disadvantage of having TCP use the $cwnd$ and $ssthresh$ values from t_1 when starting to send data at t_2?

Advantage:
If network conditions are relatively stable, i.e., there is a similar amount of bandwidth available for this TCP connection as before, then starting with the appropriate window size is advantageous compared to spending time in slow start.

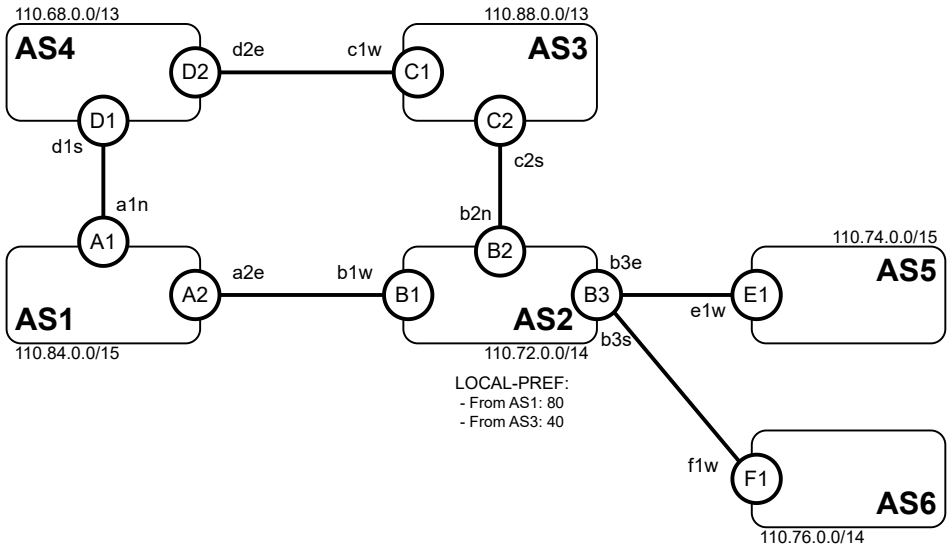
Disadvantage:
If there is less available bandwidth at t_2 (e.g., due to more contention), starting with the large window appropriate for t_1 may result in additional congestion and packet loss for this and other flows on the same path.

TCP IP EXAM - FIGURES

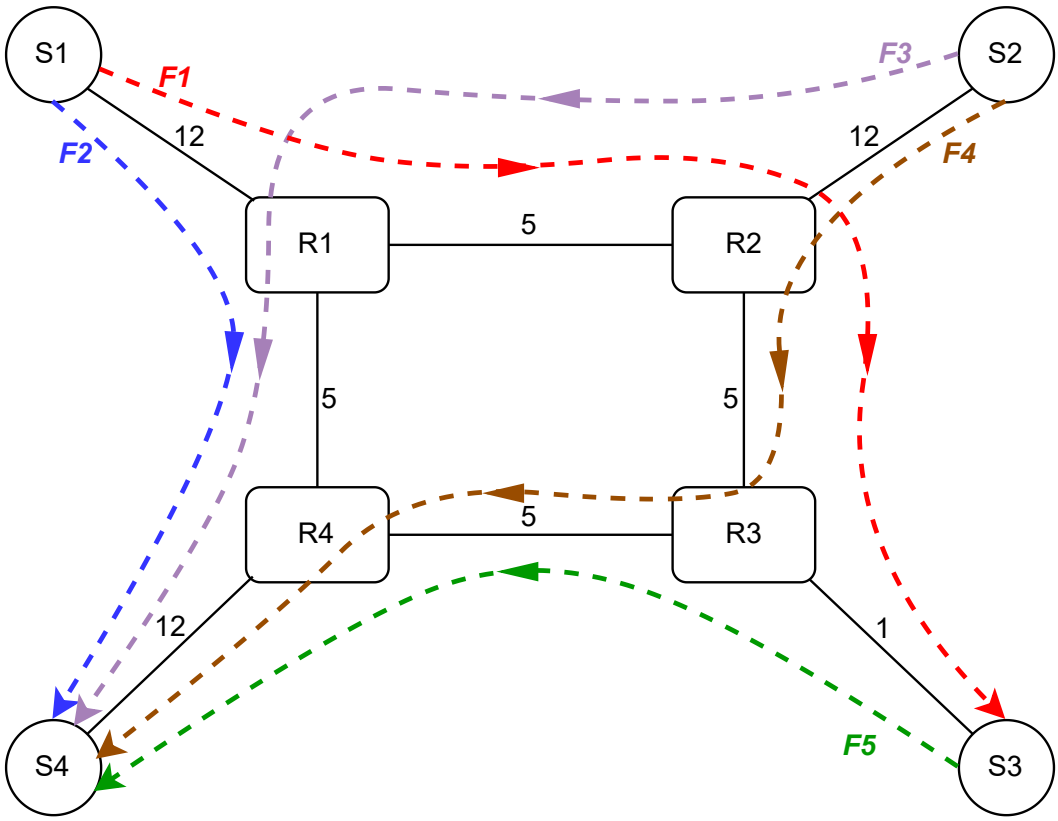
For your convenience, you can separate this sheet from the main document. Do not write your solution on this sheet, use only the main document. You do not need to return this sheet.



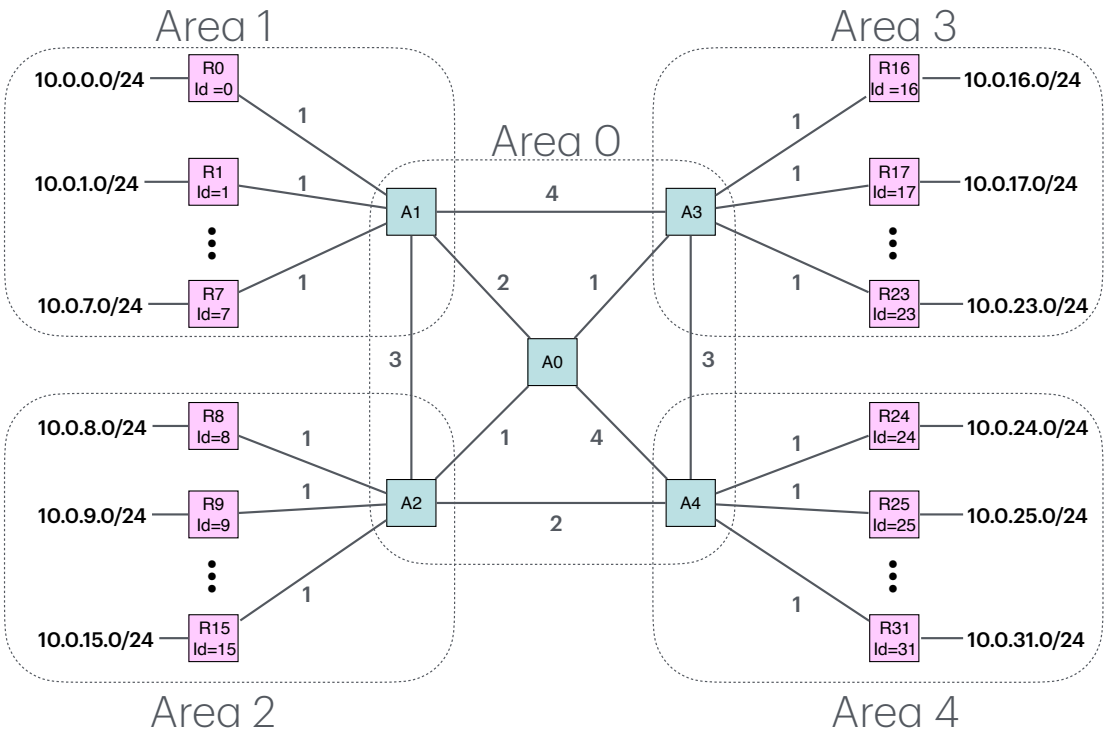
Problem 1.



Problem 2



Problem 3



Problem 4