

Exam: Advanced Topics in Communication Networks

19 February 2021, 09:30–12:00, Room HG F 7

- ▷ You **must wear a face mask at all times** during the exam, except for short eating and drinking breaks. Only medical (IIR) and FFP2 masks are allowed. Contact the assistants in case you need a spare mask.
- ▷ Write your **name** and your **ETH student number** below on this front page and **sign it**.
- ▷ Put your **legitimation card** on most accessible corner of your desk. Make sure that the side containing your name and **student number is visible**.
- ▷ Verify that you have received **all task sheets** (Pages **1 - 36**).
- ▷ **Do not separate** the task sheets. We will collect the exams **only after you have left** the room.
- ▷ Write your answers directly on the task sheets.
- ▷ All answers fit within the allocated space—often in much less.
- ▷ If you need more space, use the **extra sheets** at the end of the exam. Indicate the **task** in the corresponding field.
- ▷ Read each task completely before you start solving it.
- ▷ For the best mark, it is not required to score all points.
- ▷ Please answer in **English**.
- ▷ **Write clearly** in blue or black ink (not red) using a **pen**, not a pencil.
- ▷ **Cancel** invalid parts of your solutions **clearly** (e.g., by crossing them out).
- ▷ At the end of the exam, **place the exam face up on the top left corner** of your desk. Then collect all your belongings and **exit the room** according to the given instructions.
- ▷ No written material nor calculator are allowed.

Family name:

Student legi nr.:

First name:

Signature:

Do not write in the table below (used by correctors only):

Task	Points	Sig.
General Knowledge	/50	
P4 Program Analysis	/20	
ISP services	/20	
Design question	/60	
Total	/150	

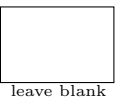
- (iii) Give one advantage and one disadvantage of using ordered label distribution control versus using independent label distribution control. (2 Points)

Advantage: _____

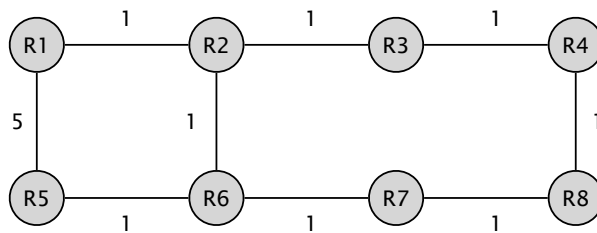
Disadvantage: _____

b) Traffic Engineering

(10 Points)



Consider the network topology below composed of 8 routers. Each link is annotated with its IGP weight which corresponds to its end-to-end delay (in μs). All the links have the same (bidirectional) capacity of 100 Gbps. In the following, we are interested in understanding how Label Switched Routers (LSRs) use the Resource Reservation Protocol (RSVP) to signal traffic-engineered Label Switched Paths (LSPs).



- (i) Assume the network just started and R1, R3, R4 *consecutively* try to establish an LSP towards R6 using RSVP. Each LSP reserves a capacity of 100 Gbps (primary objective), while minimizing delay (secondary objective). Indicate the path taken by each LSP. (3 Points)

Path of LSP R1 → R6: _____

Path of LSP R3 → R6: _____

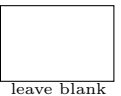
Path of LSP R4 → R6: _____

- (ii) Describe which RSVP message will R4 send to establish its LSP. How is the message routed and what type of state does it create in each intermediate router? (4 Points)

- (iii) Describe which RSVP message will R6 answer back to R4. How is the message routed and what type of state does it create in each intermediate router? (3 Points)

c) Quality of Service

(12 Points)



- (i) Consider a link shared by 5 clients. The link uses 5 token buckets (one per client) to rate limit their respective throughput. Each token bucket has a capacity (bucket size) of 10 packets and a filling rate of 1 packet per second. You can assume that the link capacity is infinite, meaning that the link itself is never a bottleneck, only the token buckets.

The table below characterizes the sending behavior of each of the 5 clients for 10 seconds. Each column corresponds to a 1 second slot and indicates how many packets each client is trying to send during that slot. Each token bucket is initialized with 10 tokens.

For each of the 5 client, indicate whether the respective token bucket will *allow* all packets to be sent or *limit* the client traffic. Circle the answer for each client. Furthermore, *if* the token bucket is limiting the client traffic, circle the *first* time slot at which the token bucket starts limiting the client. (5 Points)

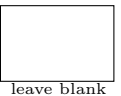
slot id \ client id	1	2	3	4	5	6	7	8	9	10	Bucket's status
client 1	1	1	1	1	1	1	1	1	1	1	allowing / limiting
client 2	2	2	2	2	2	2	2	2	2	2	allowing / limiting
client 3	10	1	10	1	10	1	10	1	10	1	allowing / limiting
client 4	10	0	0	0	0	1	1	1	2	2	allowing / limiting
client 5	5	0	5	0	0	0	0	0	0	5	allowing / limiting

- (ii) Consider a link with a total capacity of 23 units. This link is shared by 5 sources which, respectively, demand $R_1 = 1$, $R_2 = 2$, $R_3 = 5$, $R_4 = 10$, $R_5 = 10$ units. What is the max-min fair allocation for each source? Describe your computation. (5 Points)

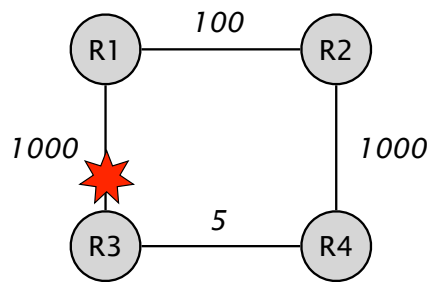
(iii) Why isn't fair queuing used in the entire Internet? It is such a great idea! (2 Points)

d) Fast Convergence

(10 Points)



Consider the network topology below composed of 4 routers (R1, R2, R3, R4). Each link is annotated with its IGP weight. At some point, the link between R1 and R3 fails. In the following, we are interested in understanding whether R3 can use Loop-Free Alternates (LFA) to protect the traffic it sends to destinations reachable behind R1.



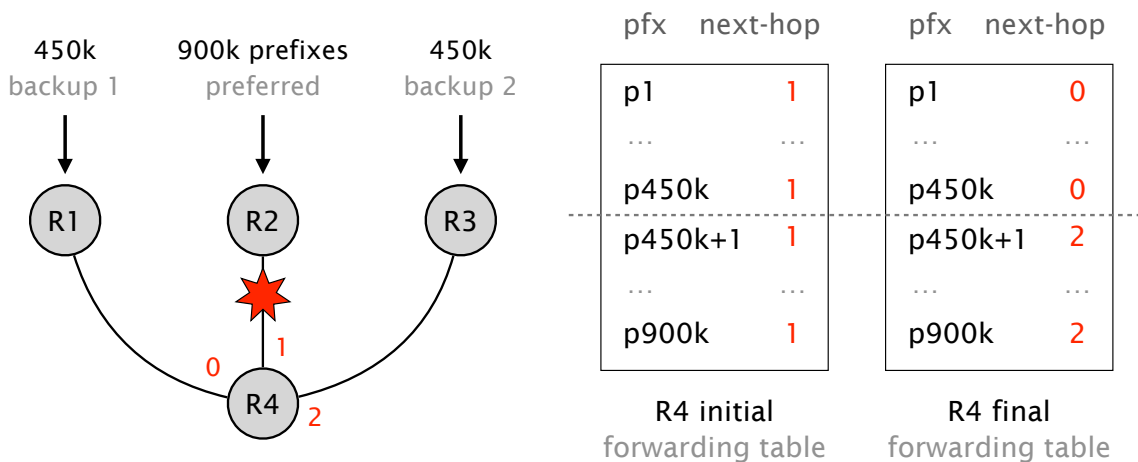
(i) Explain why R4 is *not* a LFA for R3. (2 Points)

(ii) Is it possible to adapt the link weight between R3 and R4 so that R4 can act as a LFA for R3? Give one possible link weight or explain why it is not possible. (2 Points)

- (iii) In the original topology: could R3 use remote LFA to protect against the failure of the link with R1? If so, (briefly) explain how. If not, explain why remote LFAs would not apply in this case. (2 Points)

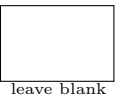
- (iv) Consider now the BGP network topology depicted below. It is composed of 3 border routers (R1, R2, R3) and one internal router R4. Each of the border router maintains one eBGP session with an external neighbor, and one iBGP session with R4. R2 learns 900k eBGP prefixes and is the preferred next-hop. In contrast, R1 and R3 learns 450k prefixes each, the union of which corresponds to the prefixes learned by R2.

R4 (flat) forwarding table initially maps each of the 900k forwarding entries to R2 (the preferred next-hop). At some point, the link between R2 and R4 fails, forcing R4 to update each of the 900k forwarding entry so that half of them map to R1 (resp. R2). Assuming an update time per entry of $\approx 100 \mu s$, the convergence time is 90 s.

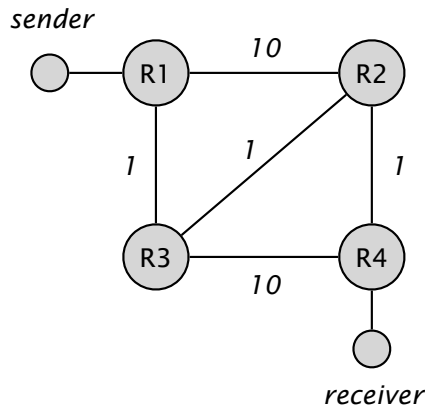


e) IP Multicast

(8 Points)



Consider the network topology below composed of 4 routers (R1, R2, R3, R4). Each link is annotated with its IGP weight. A multicast sender (resp. receiver) is connected to R1 (resp. R4). In the following, we are interested in understanding the process with which the routers reactively build a (multicast) distribution tree using the “Flood-and-Prune” strategy and Reverse Path Filtering (RPF).



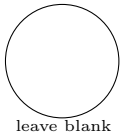
- (i) Assume the network just started. Write down (in the table below) the entire sequence of multicast transmissions which happens after the sender has sent its first packet. Write down the sender and the receiver of each transmission on a distinct line. When a router transmits more than one packet (i.e., to multiple neighbors), order the transmissions using lexicographic order on the destination (i.e., write down the transmission to R1 before the one to R2, etc.). (4 Points)

Hint: The correct answer might require less than 10 transmissions, but not more.

Transmission ID	Source	Destination
1	sender	R1
2		
3		
4		
5		
6		
7		
8		
9		
10		

- (ii) List the ID(s) of the unnecessary transmission(s). (2 Points)

- (iii) Briefly explain how to avoid these unnecessary transmissions using a sender-based approach. That is, explain how a sender can figure out which transmissions *not* to send. (2 Points)

**Task 2: P4 Program Analysis****20 Points**

You have just been hired as a P4 expert in a world-renowned Internet service provider. Your predecessor had to quit the company unexpectedly and could not finish polishing and documenting the latest P4 code she had been working on. You have been given the following piece of code (an ingress pipeline) that you must analyze to see whether you can extract any further information and continue her work.

```
1 control MyIngress(inout headers hdr,
2                   inout metadata meta,
3                   inout standard_metadata_t standard_metadata) {
4
5     register<bit<48>> 65536 table_timestamps; // index 16
6     register<bit<48>> 1024 table_rtt; // index 10
7     register<bit<10>> 1024 table_num_rtt; // index 10
8
9     action ipv4_forward(macAddr_t dstAddr,
10                       egressSpec_t port, <bit<10>> pref_index) {
11         hdr.ethernet.srcAddr = hdr.ethernet.dstAddr;
12         hdr.ethernet.dstAddr = dstAddr;
13         standard_metadata.egress_spec = port;
14         hdr.ipv4.ttl = hdr.ipv4.ttl - 1;
15         meta.monitor_metric = pref_index;
16     }
17
18     action drop() {
19         mark_to_drop();
20     }
21
22     table ipv4_lpm {
23         key = {
24             hdr.ipv4.dstAddr: lpm;
25         }
26         actions = {
27             ipv4_forward;
28             drop;
29         }
30         size = 1024;
31         default_action = drop();
32     }
33
34     action compute_index() {
35         bit<16> base = 0;
36         bit<16> cnt = 65536;
37         hash(meta.index, HashAlgorithm.crc16, base,
38             { hdr.ipv4.srcAddr,
39               hdr.ipv4.dstAddr,
40               hdr.ipv4.protocol,
41               hdr.tcp.srcPort,
42               hdr.tcp.dstPort },
43             cnt);
44     }
```

```
45
46     action store_timestamp(){
47         table_timestamps.write(meta.index,
48             standard_metadata.ingress_global_timestamp);
49     }
50
51     action compute_metric(){
52         bit<48> time_syn;
53         table_timestamps.read(time_syn, meta.index);
54         meta.metric = standard_metadata.ingress_global_timestamp - time_syn;
55     }
56
57     action aggregate_metric(){
58         bit<48> current;
59         table_rtt.read(current, meta.monitor_metric);
60         table_rtt.write(meta.monitor_metric, current + meta.metric);
61
62         bit<10> num;
63         table_num_rtt.read(num, meta.monitor_metric);
64         table_num_rtt.write(meta.monitor_metric, num + 1);
65
66         table_timestamps.write(meta.index, 0);
67     }
68
69
70     apply {
71         if (hdr.ipv4.isValid() && hdr.ipv4.ttl > 0) {
72
73             // Execute IPv4 forwarding
74             ipv4_lpm.apply();
75
76             if (hdr.tcp.isValid()) {
77                 compute_index();
78                 if (hdr.tcp.SYN == 1 && hdr.tcp.ACK != 1) {
79                     store_timestamp();
80                 } else if (hdr.ipv4.totalLen == 40 && hdr.tcp.ACK == 1) {
81                     compute_metric();
82                     aggregate_metric();
83                 }
84             }
85         }
86     }
87 }
```

- (i) Assume that a TCP SYN packet with a valid Time To Live (i.e., `ipv4.ttl > 0`) is received. Explain step-by-step how the ingress pipeline presented above would process such a packet. (4 Points)

- (ii) Consider the action `compute_metric()` on line 51. What is the purpose of this action? For which type of packets will it be executed? (2 Points)

- (iii) Consider now the action `aggregate_metric()` on line 57. What is the purpose of this action? What could be the reason behind using the `meta.monitor_metric` index instead of reusing `meta.index`? (2 Points)

Purpose of `aggregate_metric()`: _____

Reason behind `meta.monitor_metric`: _____

(iv) Explain the overall functionality of the ingress pipeline code presented above. (2 Points)

(v) There are cases where this code would not fulfill its intended functionality. Mention two potential problems with this code and explain how you would modify it to mitigate these problems. Discuss the trade-off with your proposal (if any). (6 Points)

Problem 1: _____

Mitigation 1: _____

Problem 2: _____

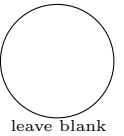
Mitigation 2: _____

Trade-offs: _____

- (vi) Imagine now that you are an attacker who knows that this code is running on a given switch, and you can send crafted packets to that switch. Describe two different attacks you could execute to degrade the performance of the P4 program. (4 Points)

Attack 1: _____

Attack 2: _____

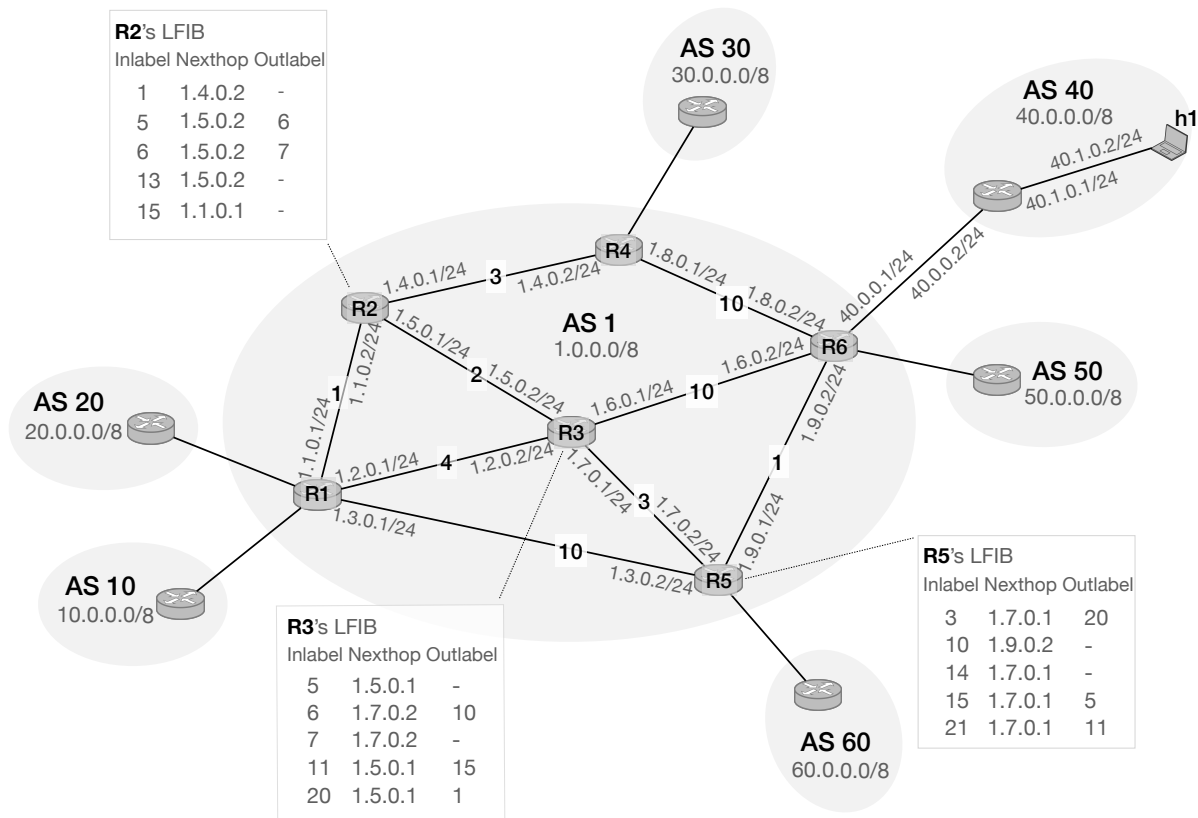


Task 3: Internet service provider services

20 Points

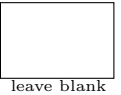
This question studies the Autonomous System (AS) 1 depicted below.

- ▷ AS 1 is an Internet Service Provider (ISP) interconnecting six customers ASes: AS 10, 20, 30, 40, 50 and 60.
- ▷ AS 1 uses OSPF for internal connectivity and BGP to receive and advertise prefixes to the neighboring ASes. The IP addresses configured on the router interfaces within AS 1 as well as the OSPF link weights are shown on the figure. Furthermore, each AS advertises its own public prefix; for instance AS X announces X.0.0.0/8.
- ▷ AS 1 has configured a BGP Free Core using MPLS, and the Label-Switched Paths (LSPs) are determined with LDP. It uses penultimate hop popping and has configured all relevant BGP properties on the edge routers.
- ▷ Finally, the figure shows the Label Forwarding Information Base (LFIB) of routers R2, R3 and R5. Routers R1, R4 and R6 also have an LFIB, but they are not shown on the figure. In the LFIB, the Inlabel column indicates the matched MPLS label of an incoming packet whereas the Outlabel column indicates the outgoing MPLS label. A dash (-) in the Outlabel column indicates that the incoming label is popped and no new label is added.



a) Investigating the MPLS tunnels

(10 Points)



- (i) Which customer prefixes could AS 1 group in the same FEC(s)? (1 Point)

- (ii) Which path within AS 1 does the traffic from AS 40 to AS 30 follow? How is this path determined? (2 Points)

- (iii) A host in AS 20 sends an IP packet to a host in AS 50. Indicate the path used by the packet within AS 1. For each hop, indicate the inbound and outbound MPLS labels, starting with the outbound label at R1. If one router does not know how to forward the packet, write “X” as outbound label and keep the following lines empty. Write “None” to indicate no label. (4 Points)

Router	Inbound label	Outbound label
R1	None	

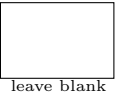
- (iv) The host `h1` in AS 40 launches a `traceroute` towards the router R2. In the table below, write down the IP addresses returned by the `traceroute`. In cases `h1` does not receive an ICMP Time to live Exceeded message for some initial TTL values, write “X” in the corresponding lines of the table. (3 Points)

Reminder. By default, a `traceroute` sends ICMP Echo Requests with an increasing Time To Live (TTL) value in the IP header (starting from 1). When the TTL value reaches 0, a router sends an ICMP Time to live Exceeded message back to the source, thus enabling the source to collect information about which routers forwarded the packet.

Initial TTL	Output
1	
2	
3	
4	
5	
6	

b) BGP VPN under the microscope

(10 Points)



In this question, we consider that all ASes are connected to AS 1 using virtual routing and forwarding (VRF) and announce both their own public prefix (X.0.0.0/8) **and** their internal—private—prefix (192.168.0.0/16) to AS 1. For example,

- AS 10 announces 10.0.0.0/8 and 192.168.0.0/16.
- AS 20 announces 20.0.0.0/8 and 192.168.0.0/16.

Note that all customers are using the same internal prefix.

The routers in AS 1 exchange VPN routes using Multiprotocol BGP (MP-BGP). Routes from VRFs are shared via MP-BGP tagged with both a *route distinguisher* (rd) and a list of *route targets* (rt). This is configured in FRR by using the `rd vpn export`, `rt vpn export`, and `rt vpn import` commands. The following snippet shows an extract of the VPN configuration of the border routers in AS1.

The route map RT_FILTER is used to differentiate between private and public prefixes. It is applied during export and **removes** the route target 1:1 for /16 prefixes. For example, consider the announcements from AS 10 received by router 1 in VRF_1. Router 1 tags

- 10.0.0.0/8 with route targets 1:1 and 10:1, and
- 192.168.0.0/16 with route target 10:1 only.

```
# Router 1
router bgp 1 vrf VRF_1
neighbor 10.0.0.0/8 remote-as 10
neighbor 20.0.0.0/8 remote-as 20
route-map vpn export RT_FILTER
rd vpn export 1:1
rt vpn export 1:1 10:1
rt vpn import 1:1 10:2
```

```
# Router 4
router bgp 1 vrf VRF_2
neighbor 30.0.0.0/8 remote-as 30
route-map vpn export RT_FILTER
rd vpn export 1:2
rt vpn export 20:1
rt vpn import 20:2
```

```
# Router 5
router bgp 1 vrf VRF_3
neighbor 60.0.0.0/8 remote-as 60
route-map vpn export RT_FILTER
rd vpn export 1:3
rt vpn export 1:1 10:2
rt vpn import 1:1 10:1
```

```
# Router 6
router bgp 1 vrf VRF_4
neighbor 40.0.0.0/8 remote-as 40
route-map vpn export RT_FILTER
rd vpn export 1:4
rt vpn export 1:1
rt vpn import 1:1
```

```
router bgp 1 vrf VRF_5
neighbor 50.0.0.0/8 remote-as 50
route-map vpn export RT_FILTER
rd vpn export 1:5
rt vpn export 20:2
rt vpn import 20:1
```


- (iii) Below are 6 packets that were captured on the link R2–R3. Each packet has two MPLS labels, as shown in the table below.

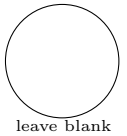
Packet Number	SrcIP	DstIP	OuterLabel	InnerLabel
1	60.0.0.1	10.0.0.1	15	81
2	192.168.0.1	192.168.0.2	1	85
3	10.0.0.1	60.0.0.1	7	81
4	30.0.0.1	50.0.0.1	6	85
5	192.168.1.1	192.168.1.2	7	81
6	192.168.1.1	192.168.1.2	?	85

Between which pair of ASes is the sixth packet sent? What is the OuterLabel of the sixth packet? If you cannot be certain about the pair of ASes or the label, indicate the possible options. Explain your reasoning. (4 Points)

Pair(s) of ASes: _____

Outer label(s): _____

Reasoning: _____

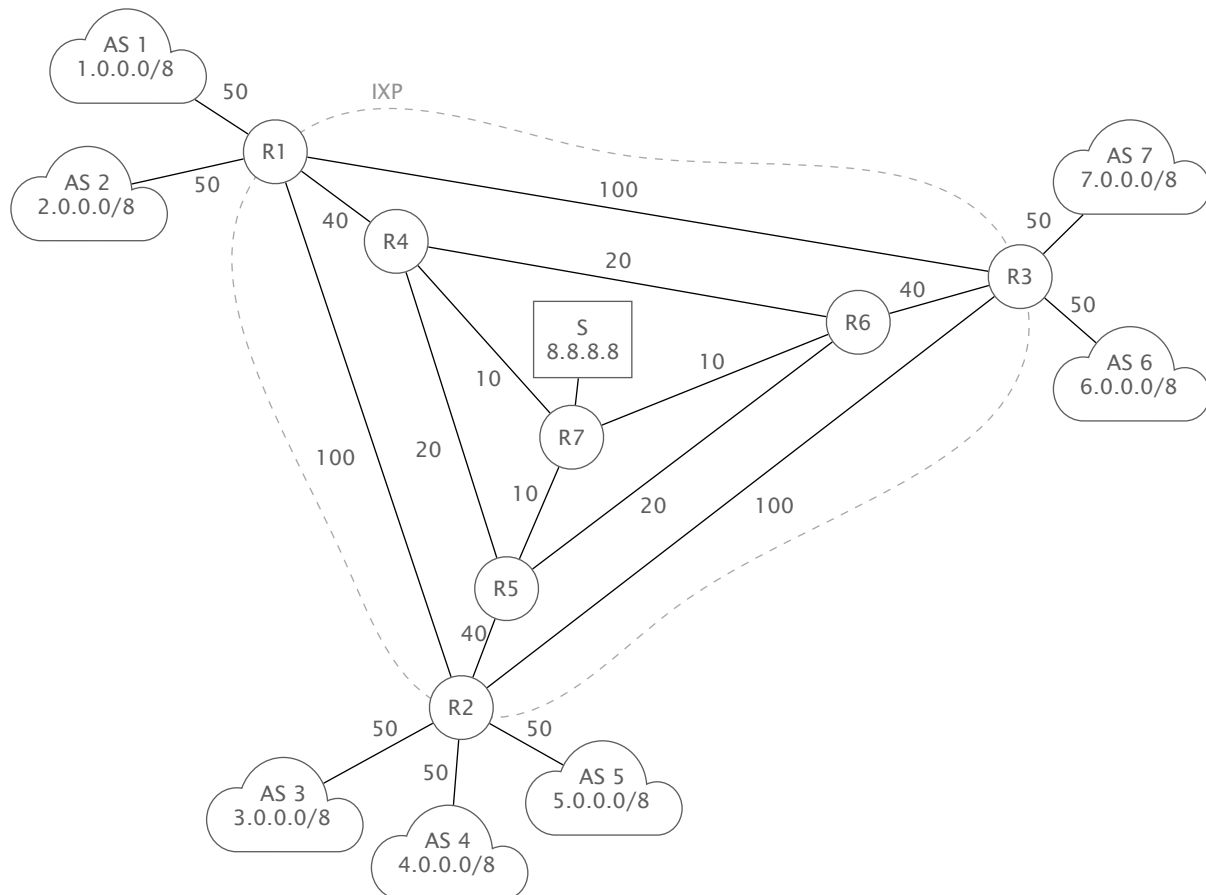
**Task 4: Design question****60 Points**

This question contains two parts which relate to configuring and testing networks.

- ▷ Part 1 is subdivided into four dependent questions. Part 2 is independent of Part 1.
- ▷ There is not always a unique right answer to the questions: different answers may give full points. In some questions, you are asked for two distinct solutions. You will not get more points for providing more solutions. On the contrary, if you do not clearly indicate which are your two proposed solutions, we will consider the “worst two.”
- ▷ In several questions you are asked to describe which solution you would use; you must also **explain how** you would use it and/or **why** this solution is appropriate. For example, simply writing “we can use MPLS” is not enough and will give zero point. Your answers should be detailed enough to enable a knowledgeable engineer to implement your solution.
- ▷ You **should not** write code snippets. While you may write pseudo-code to describe a part of your solution, this is **not expected**. Do so only if necessary.
- ▷ If your solution for one question (say, **b**) extends your solution to a previous question (say, **a**), you do not need to repeat your previous solution; only specify which parts you inherit. For example, you may write “I use the solution from **a** and extend it as follows. (...)” You can also do this if you did not solve one task, but you know how to extend it to solve a subsequent task.

Part1—Design and configuration of an IXP network

Consider the following network topology belonging to an Internet eXchange Point (IXP). As a reminder, an IXP interconnects border routers from different Autonomous Systems (ASes). In this question, the IXP interconnects 7 ASes.



We make the following hypotheses.

- ▷ All routers run FRR, are P4-enabled, and run a FRR and P4 control plane. There is **no central controller**.
- ▷ You have full access to the routers. For example, you can run commands such as `tc` on all the routers' interfaces.
- ▷ Edge labels indicate the links' bandwidth in Gbps. For example, the link between R1 and R2 has a bandwidth of 100 Gbps.
- ▷ A server S is connected to R7; its usage is described in the relevant question.

(ii) Discuss 2 advantages that solution 1 has over solution 2 and vice-versa. (4 Points)

1st advantage of solution 1: _____

2nd advantage of solution 1: _____

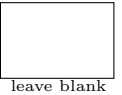
1st advantage of solution 2: _____

2nd advantage of solution 2: _____

(iii) Explain which solution you would favor and why. (1 Point)

Favored solution: _____

Reason: _____

c) Fast reroute**(15 Points)**

In practice, multiple failures may happen in the IXP network. You must propose a fast reroute system that fulfills the following requirements:

- Connectivity should be preserved between all ASes as long as the IXP network remains connected.
- The overall convergence time should be independent of the number of prefixes involved.
- For fairness reasons, each AS can utilize at most 20 Gbps of each of the 40 Gbps links.

- (i) Describe the four main sources of convergence delay and explain which one(s) has(ve) the largest impact on convergence delay, and why. (5 Points)

Source 1: _____

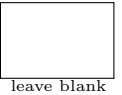
Source 2: _____

Source 3: _____

Source 4: _____

Most impactful delay source(s) and reason: _____

- (ii) Describe a fast reroute solution that satisfies the requirements. (10 Points)

d) VoIP traffic monitoring**(5 Points)**

You aim to monitor how much of VoIP traffic crosses the IXP network; for that, you must design a monitoring system. We make the following hypotheses.

- VoIP packets are uniquely identified by a value of 123 in the TOS field in the IP header. No other traffic has its IP header TOS field set to 123.
- In all routers, packets' ingress timestamps are stored in an intrinsic metadata variable called `meta.ts`.
- There is no failure nor network congestion.

The only requirement for your monitoring system is that the server `S` should receive a clone of each VoIP packet entering from one of the ASes, together with the timestamp of the packet's entry time into the IXP network.

- (i) Describe one solution that satisfies the requirement. Elaborate on your data-plane design. (3 Points)

- (ii) How would you validate that your solution monitors the VoIP traffic as expected? (2 Points)
