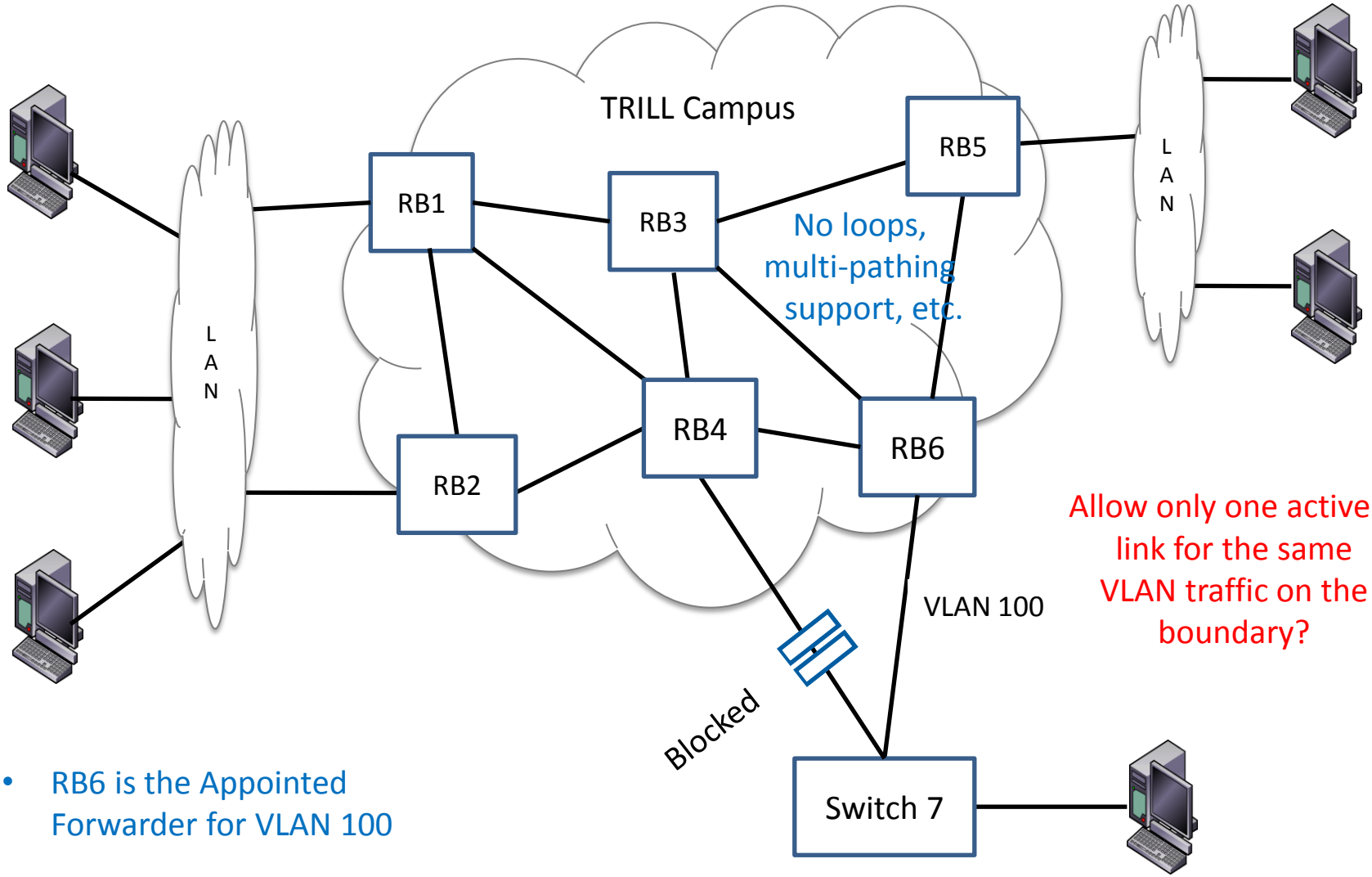
t-LAG:
A Distributed LAG Mechanism for TRILL Enabled Fabrics

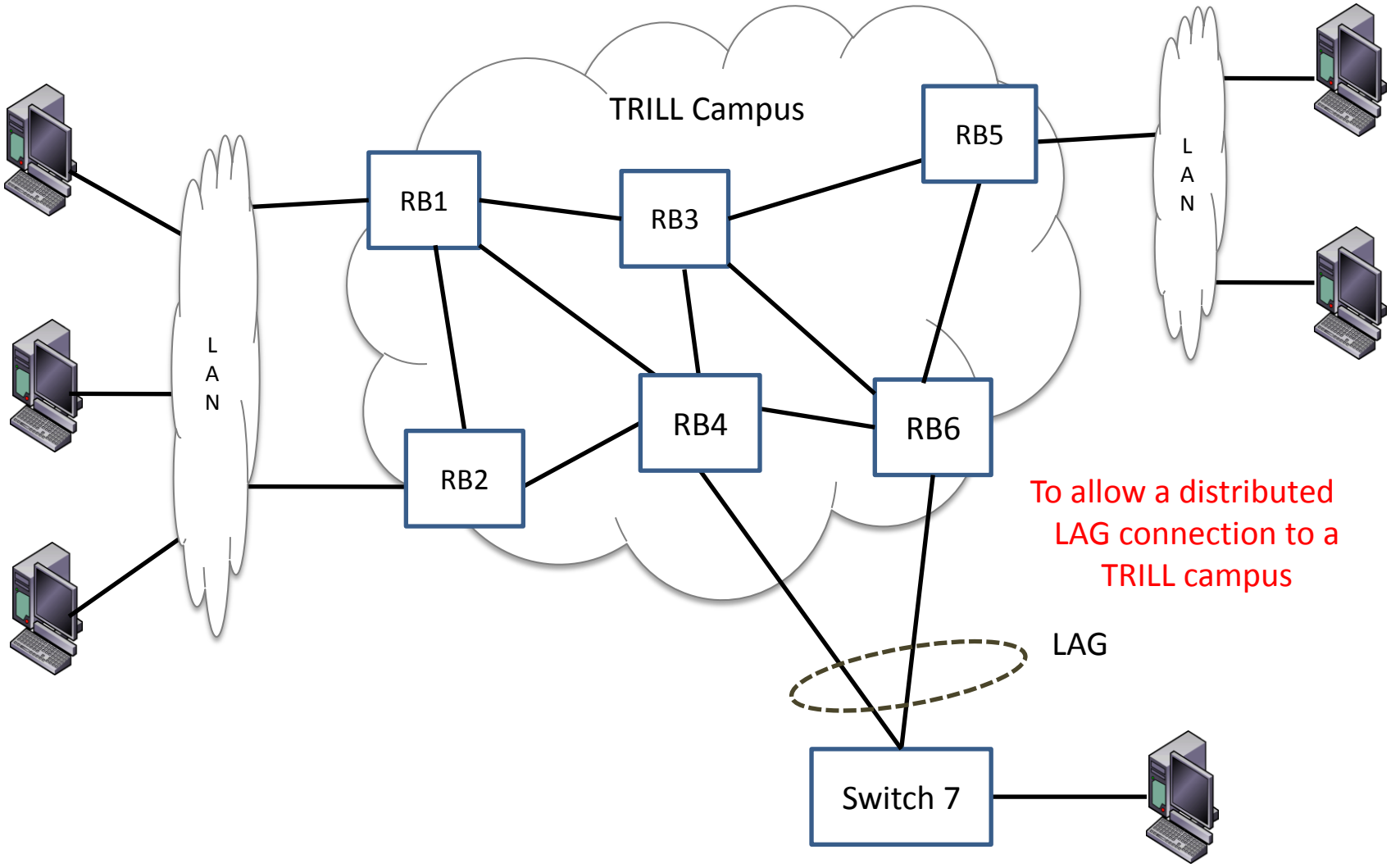
Dar-Ren Leu, Vijoy Pandey
dleu@us.ibm.com
9/9/2011

Problem Statement

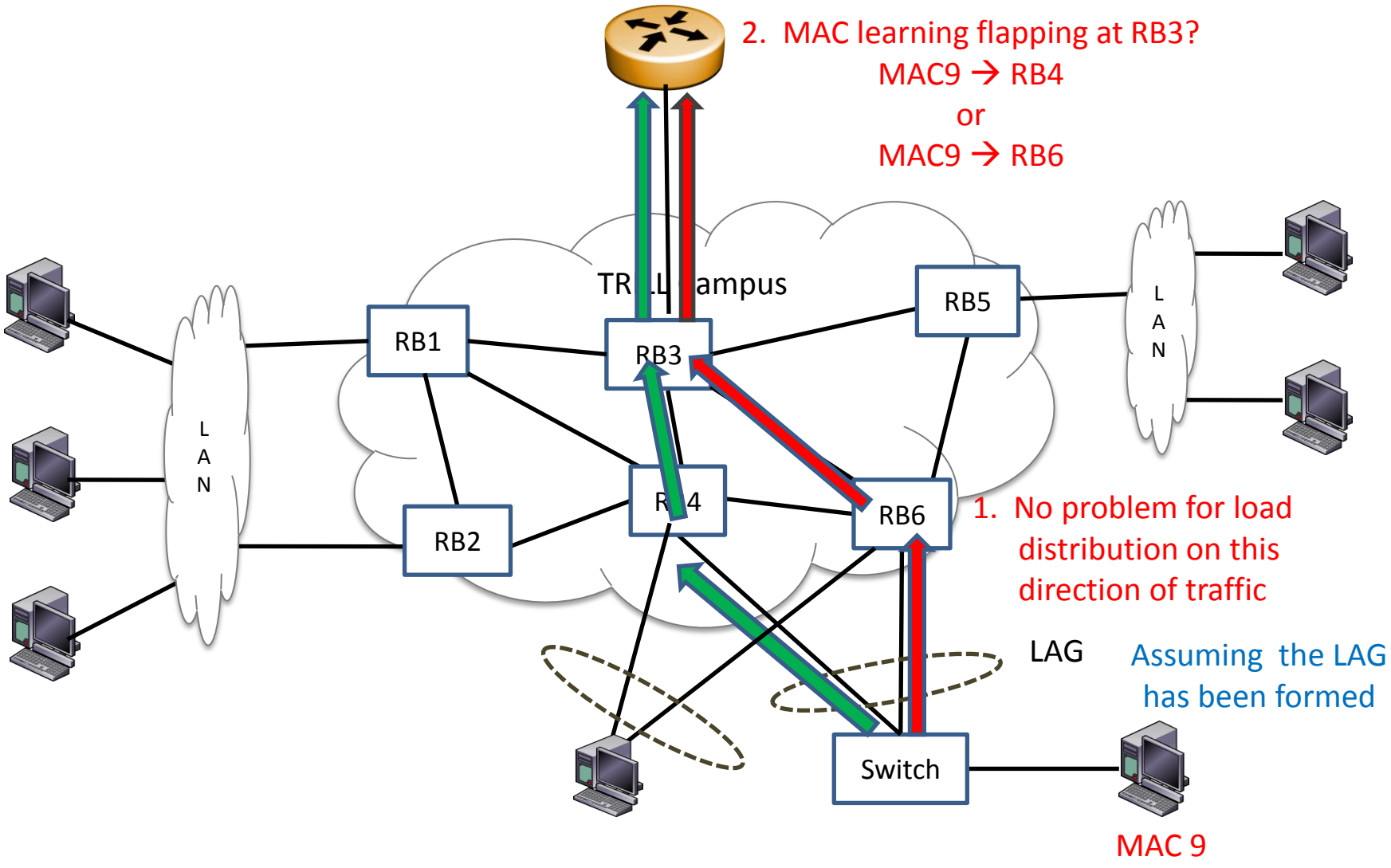


- RB6 is the Appointed Forwarder for VLAN 100

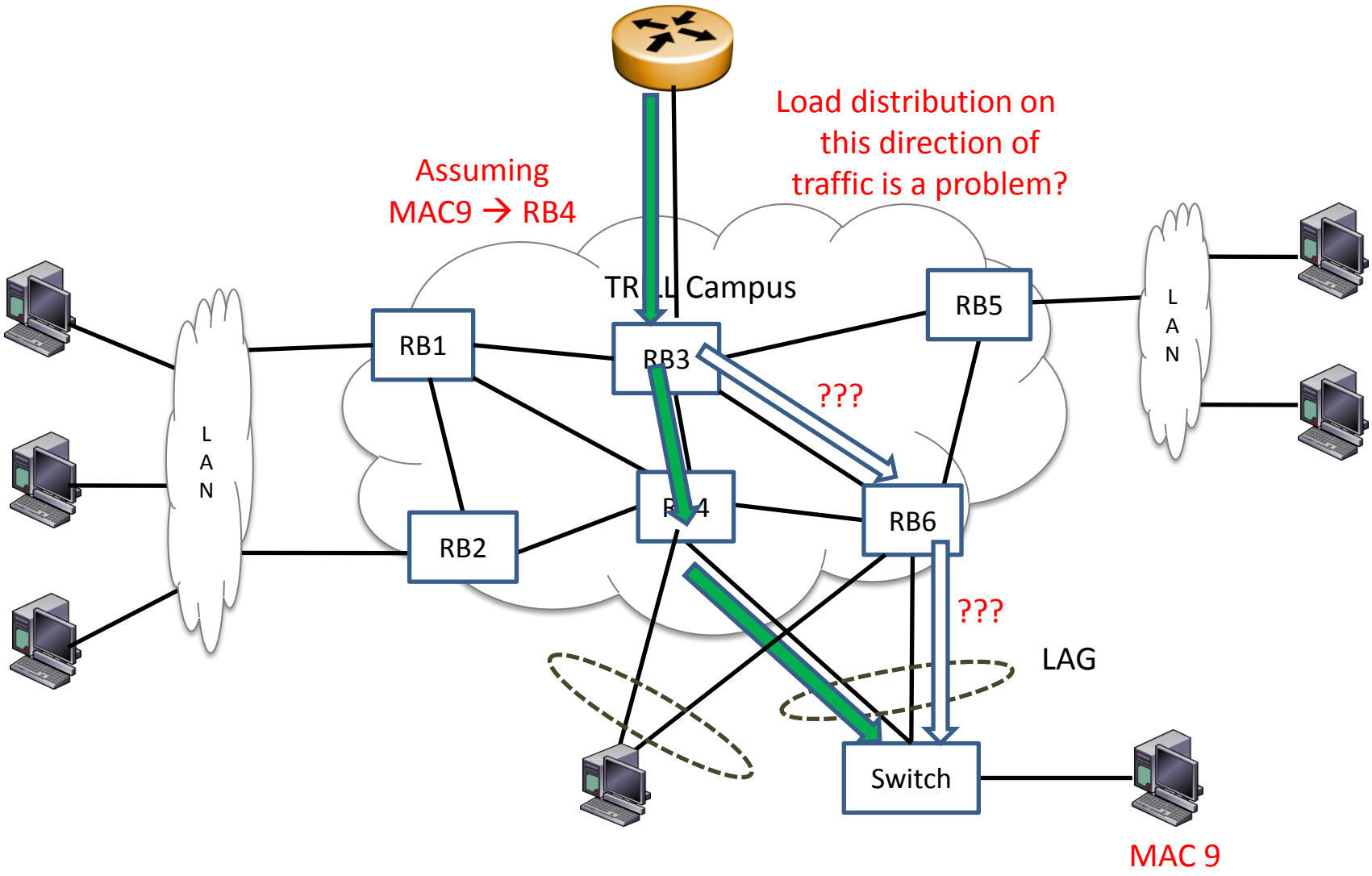
Desirable


TRILL Campus
RB1
RB2
RB3
RB4
RB5
RB6
LAN
LAN
Switch 7
LAG
To allow a distributed LAG connection to a TRILL campus

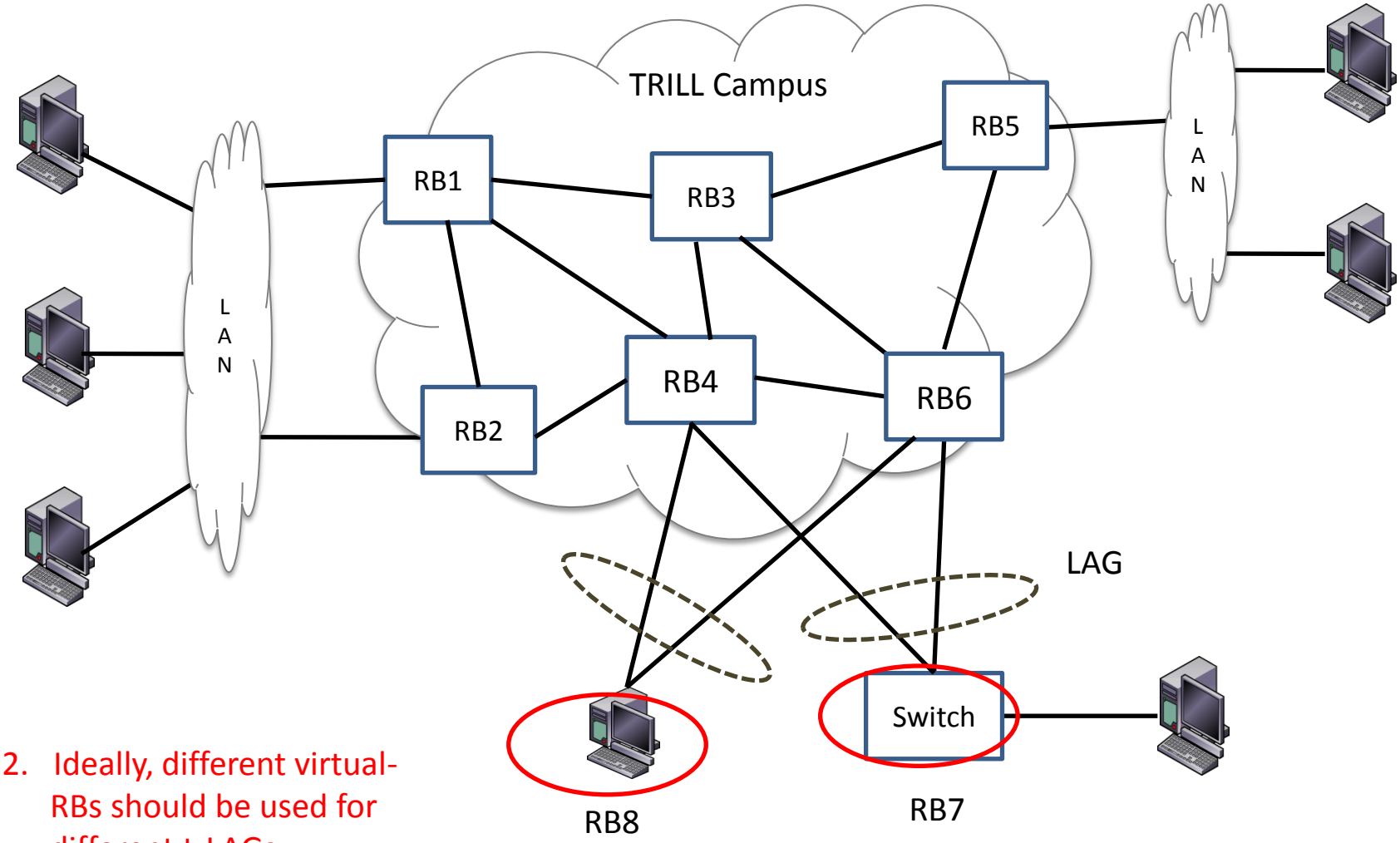
Traffic Ingress at a LAG



Traffic Destined to a LAG



The t-LAG



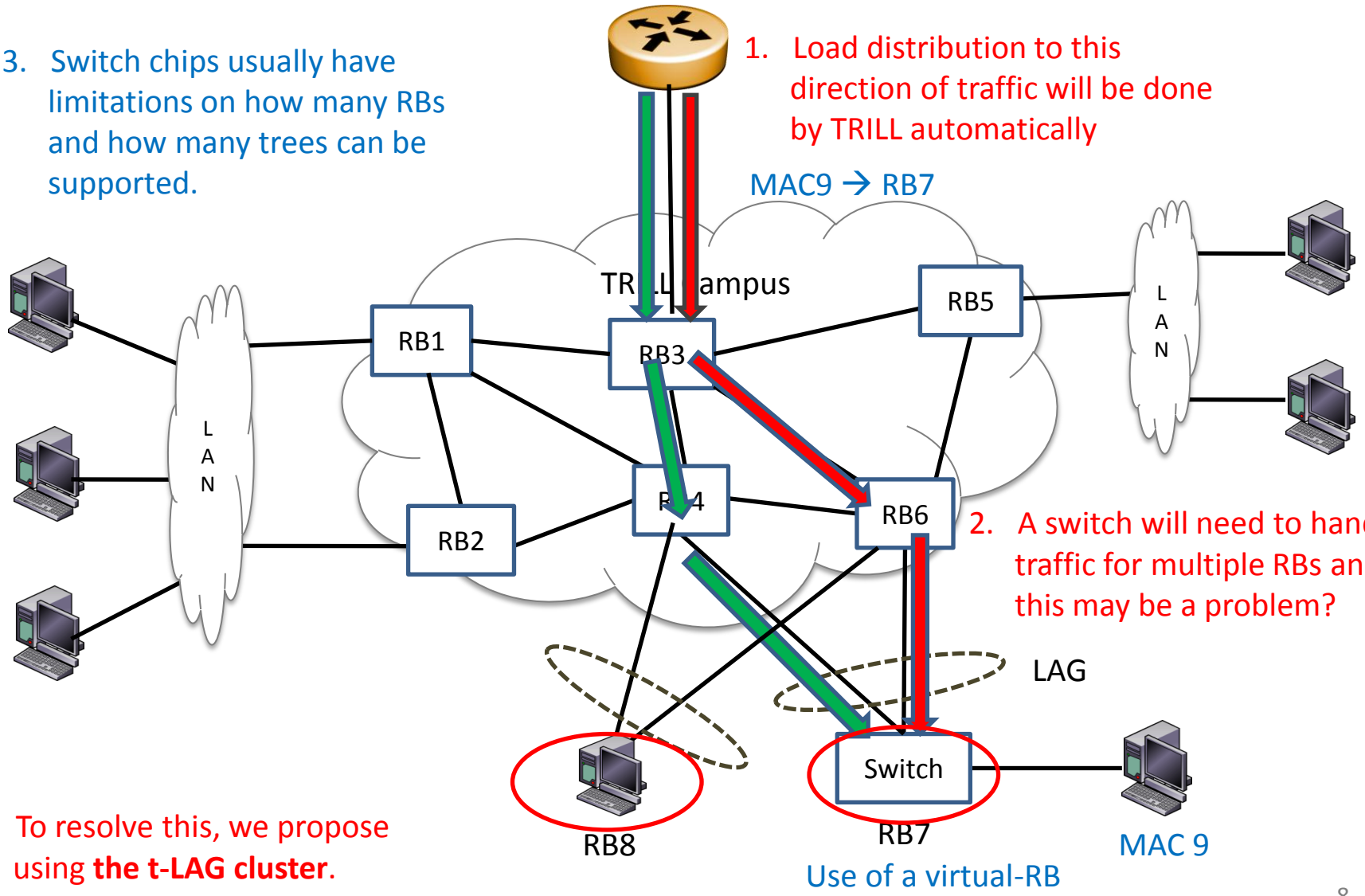
2. Ideally, different virtual-RBs should be used for different t-LAGs

1. Use of a virtual-RB

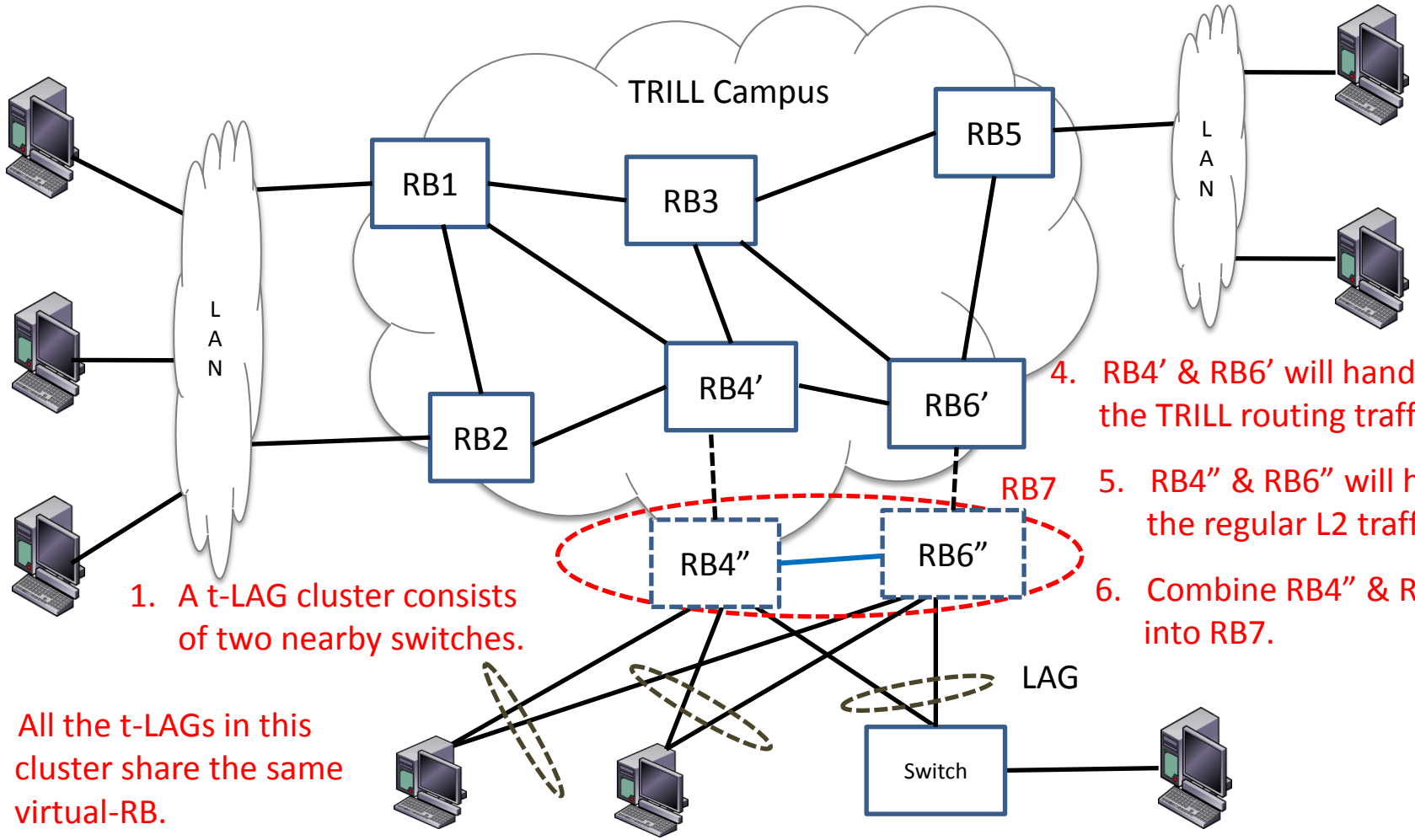
The t-LAG

- Through the use of a virtual-RB for each t-LAG; all the RBs connected to the same t-LAG should share the same nickname for that virtual-RB.
- Involve all these virtual-RBs in the TRILL IS-IS communication as well as the SPF calculation.
- The switch that has t-LAGs on it should take care of the needed IS-IS as well as the ESADI communication for all related virtual-RBs.
- The switch chips for those t-LAG enabled switches should handle traffic for both ingress/egress at t-LAGs in a proper way.

Traffic Destined to a LAG



The t-LAG Cluster



1. A t-LAG cluster consists of two nearby switches.
2. All the t-LAGs in this cluster share the same virtual-RB.
3. Separate RB4 to RB4' & RB4"; same for RB6.
4. RB4' & RB6' will handle the TRILL routing traffic.
5. RB4" & RB6" will handle the regular L2 traffic.
6. Combine RB4" & RB6" into RB7.
7. All the regular L2 traffic in RB4" & RB6" should be handled within RB7. An ISL is needed in between RB4" & RB6" for failure handling.

The t-LAG Cluster

- Group two nearby switches together to form a t-LAG cluster (e.g., RB4 and RB6 on previous slide); these two switches will be reserved mainly for t-LAG purpose.
- All the t-LAGs in a t-LAG cluster should share the same virtual-RB; however, more than one virtual-RB can be used if desired.
- For each switch (say, RB4) in the cluster, separate its functions into two domains: the TRILL routing domain (denoted by RB4') and the local access domain (denoted by RB4").
- Connect the local access domains (RB4" and RB6") in a t-LAG cluster by ISL, called the t-LAG ISL (denoted by the blue link on previous slide); the local access domains in a t-LAG cluster will then be combined into a virtual-RB and assigned a unique ID (say, RB7).
- All the regular L2 traffic in RB4" & RB6" should be handled within RB7. That is, the t-LAG ISL in between RB4" & RB6" will be used for failure handling if a local t-LAG link fails.

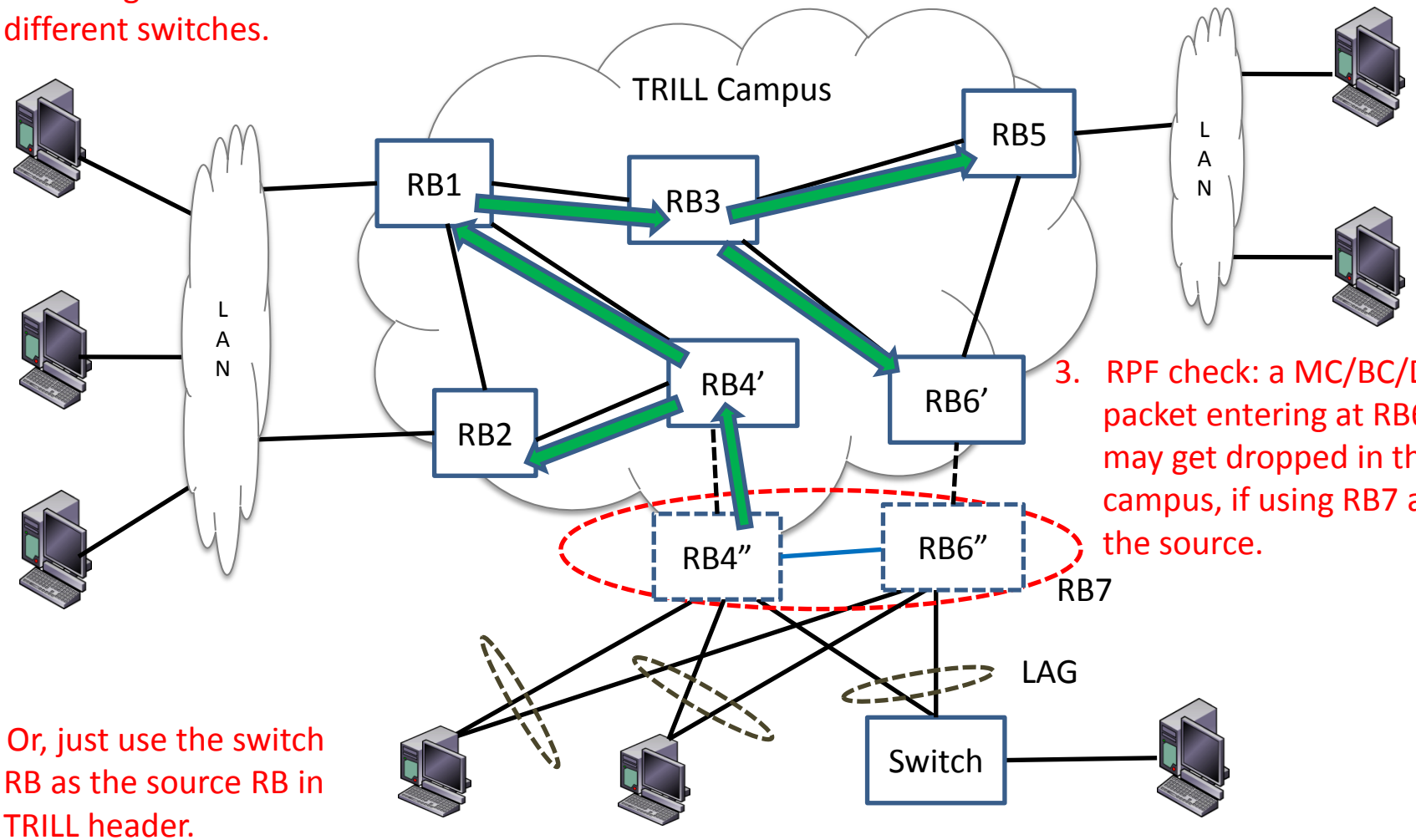
Which RB to Use as the Source?

- In the above, we have assumed using the ingress virtual-RB as the source RB in TRILL header.
  - This might cause problems to a MC/BC/DLF distribution inside the TRILL campus
- Alternative is to use the switch RB as the source RB in TRILL header.

A Problem for MC Traffic

4. Alternative is to use different trees for traffic ingress at different switches.

1. Assuming the ingress virtual-RB is used as the source.



3. RPF check: a MC/BC/DLF packet entering at RB6 may get dropped in the campus, if using RB7 as the source.

5. Or, just use the switch RB as the source RB in TRILL header.

2. Assuming the tree rooted at RB7 is used at both RB4 & RB6.

MAC Learning

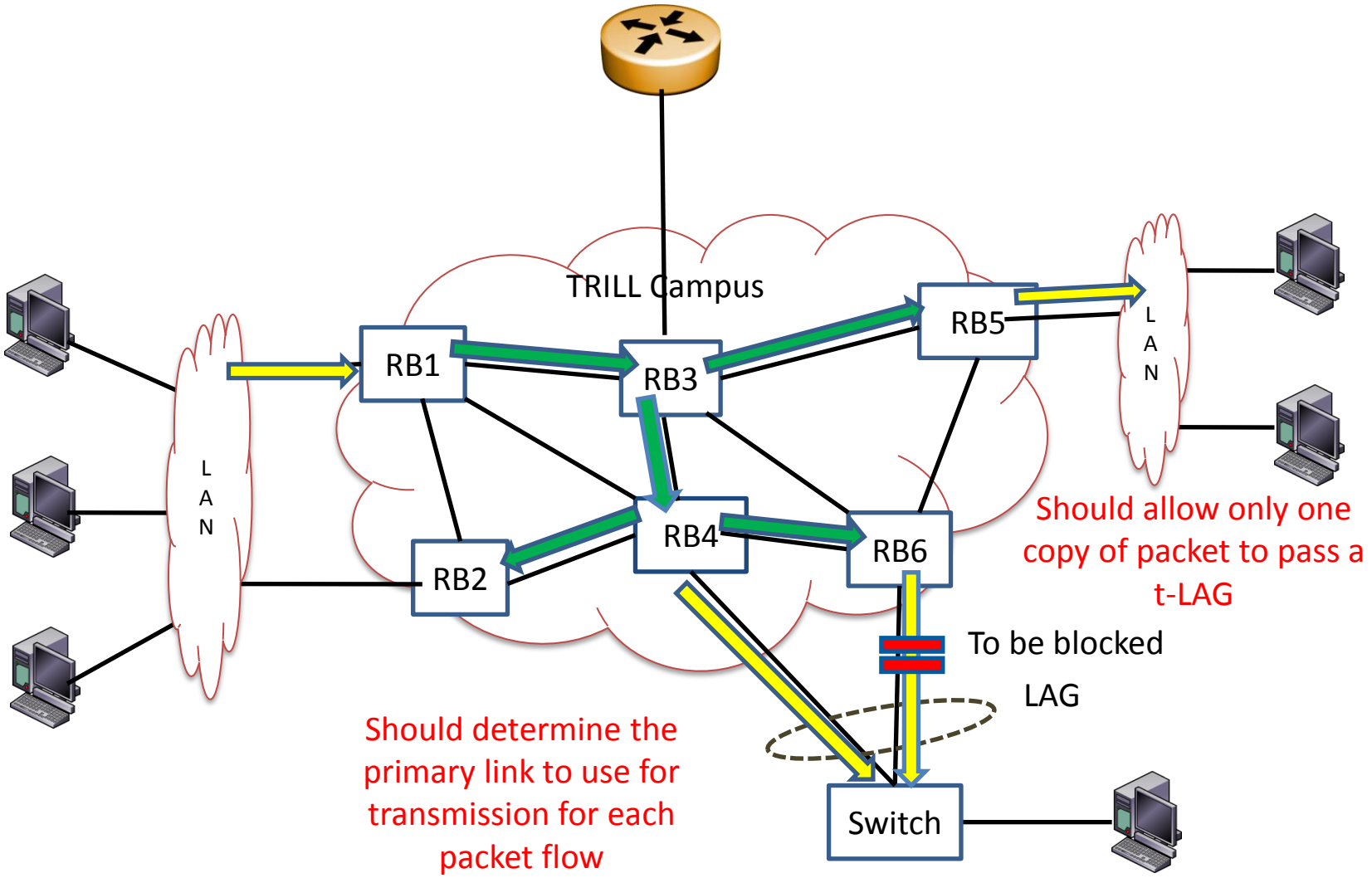
- If using the ingress virtual-RB as the source RB in TRILL header,
  - MAC learning via h/w is not an issue

- If using the switch RB as the source RB in TRILL header, two MAC learning methods are allowed: one s/w based, the other h/w based.
  - S/W based:
    - A new MAC entry learnt at a t-LAG will need to be passed to the CPU on ingress switch and the entry is then modified to map to the corresponding ingress virtual-RB by the S/W
    - After that, the MC entry is propagated to other RBs via ESADI
    - Allow RBridges of different vendors to talk to each other
  - H/W based:
    - The MAC entry on chips will be vport based
    - Chips allow to map multiple RBs (e.g., RB4, RB6, RB7) into the same vport
    - In this case, all the RBridges in the campus will need to know how to do this proper mapping and, thus, all the switches will need to come from the same vendor (a limitation)

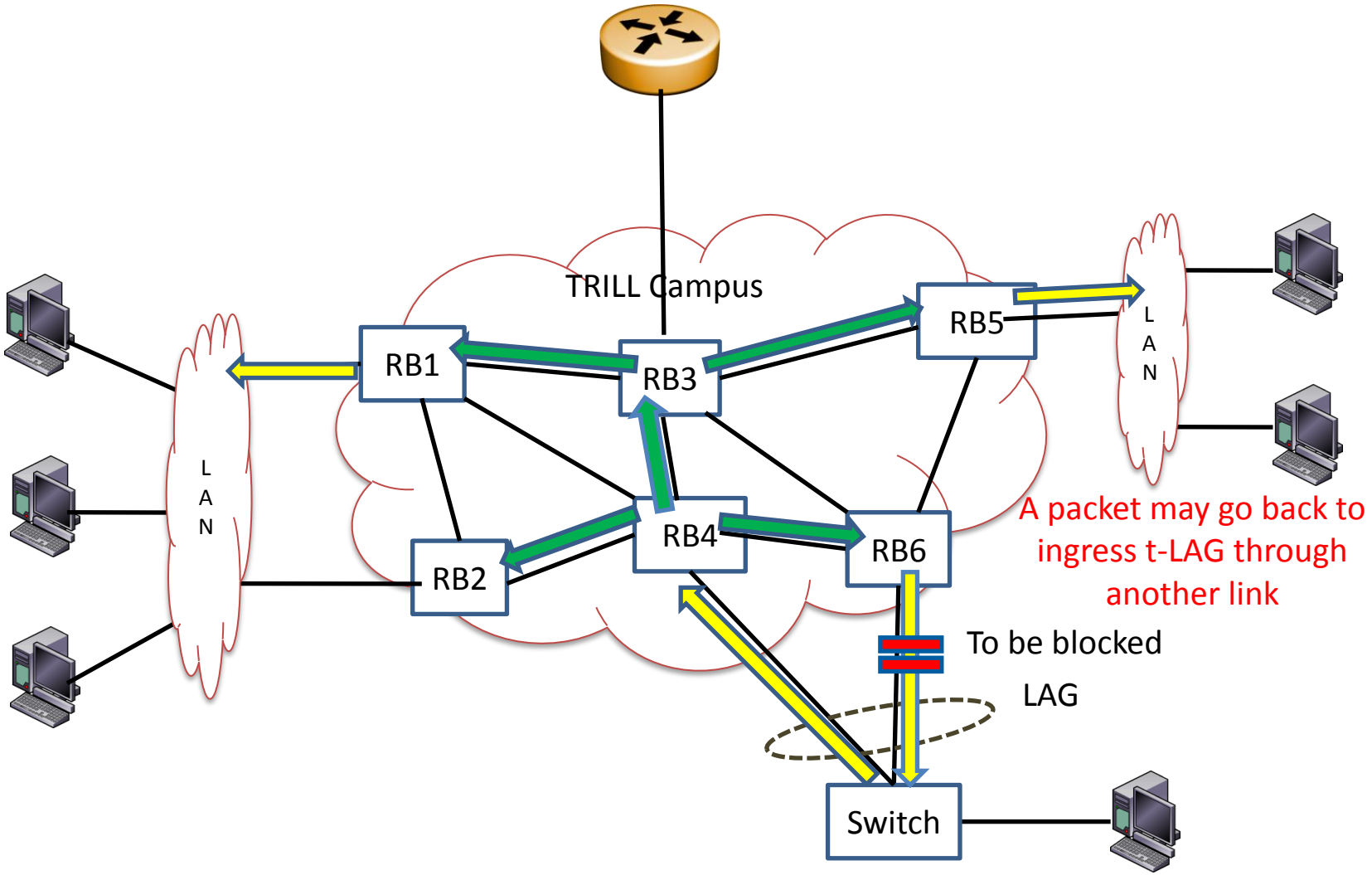
Other Problems for MC Traffic

- Multiple copy of the same packet to a t-LAG is possible and must be eliminated
  - Need to determine the primary link to use for transmission for each tree and each t-LAG
  - Different primary link for each t-LAG can also be chosen for different packet flows of MC/BC/DLF traffic
    - System based
    - (Tree, VLAN, DMAC) based
- Source port (t-LAG) pruning
  - ACLs or some other schemes will be required

Multi-Copy Issue



Source t-LAG Pruning



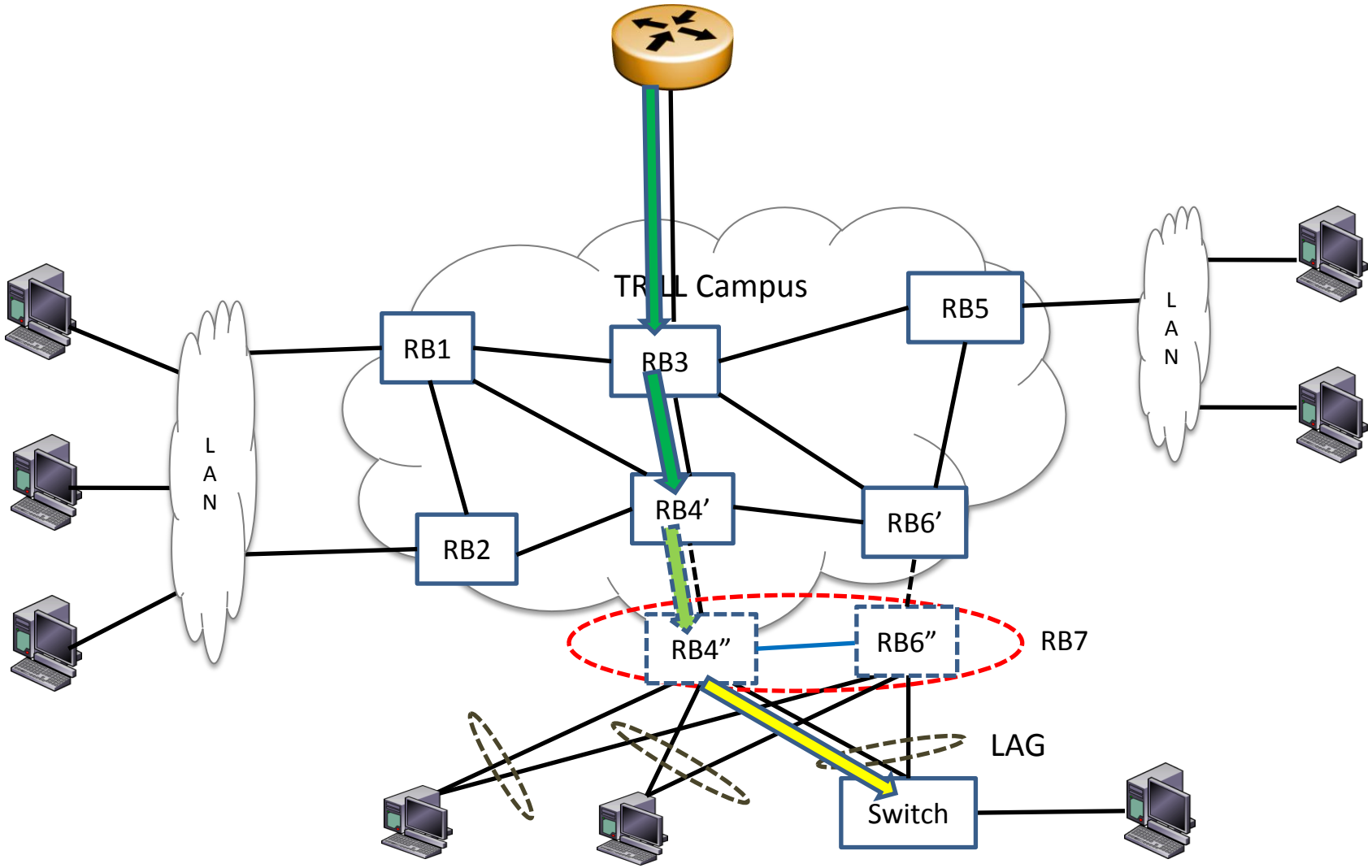
Failure Handling

- Problem: a t-LAG link may go down at run time; some traffic destined to that t-LAG may get dropped due to this
- Several possible solutions for this:
  1. The use of the t-LAG ISL for packet redirection
  2. Stop claiming the connectivity in between the switch RB and the virtual-RB at run time, if too many local t-LAG links go down (for UC)
  3. Change the primary link for a t-LAG (for MC/BC/DLF)
- Due to that multiple t-LAGs sharing the same virtual-RB, should implement all three above to better conquer the failures for various scenarios.

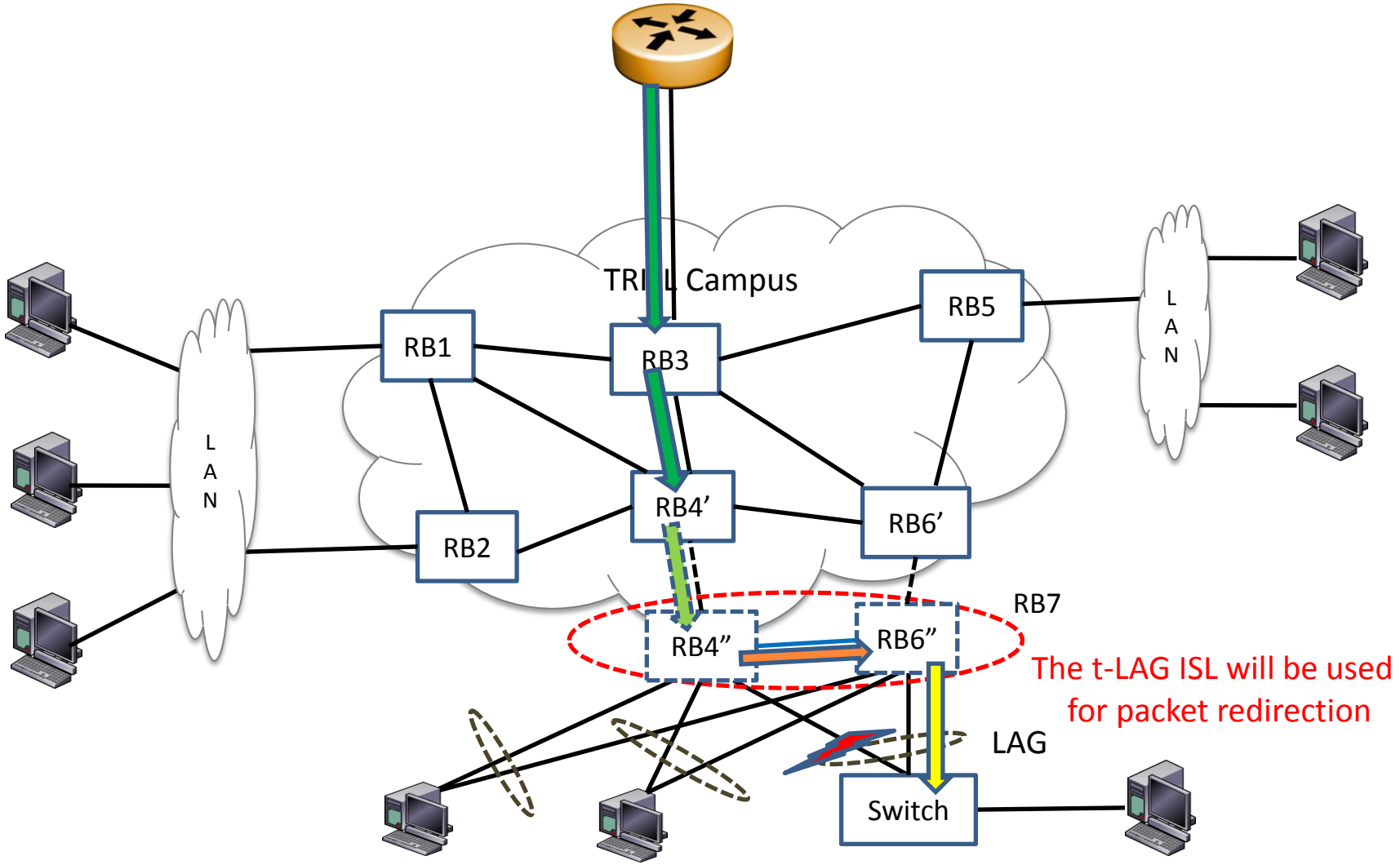
Scenarios

- Case 1: Packet flow of UC to a t-LAG via TRILL.
- Case 2: Packet flow of MC/BC/DLF traffic from a t-LAG.

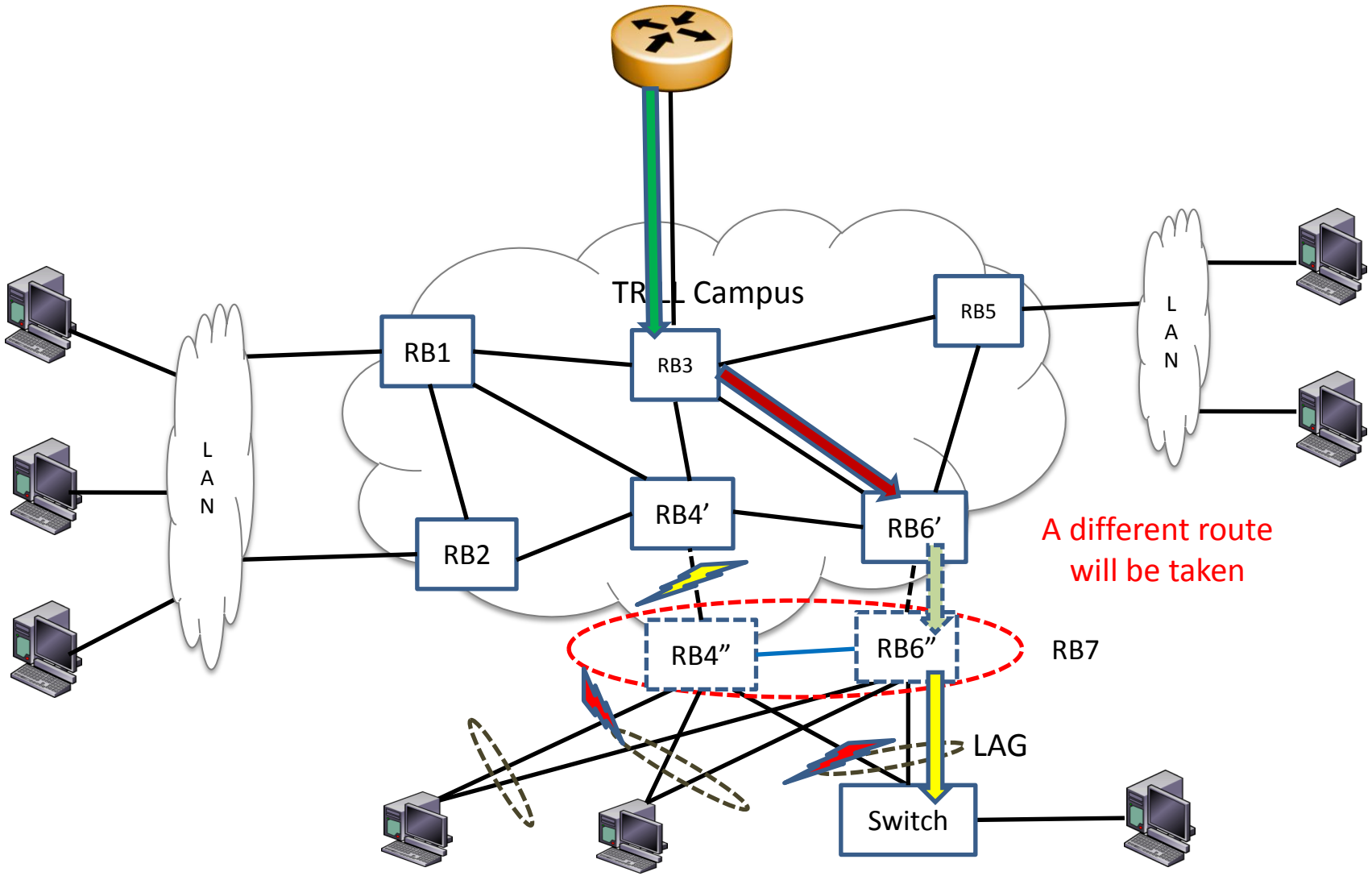
Scenario 1: The normal packet flow of UC traffic to a t-LAG



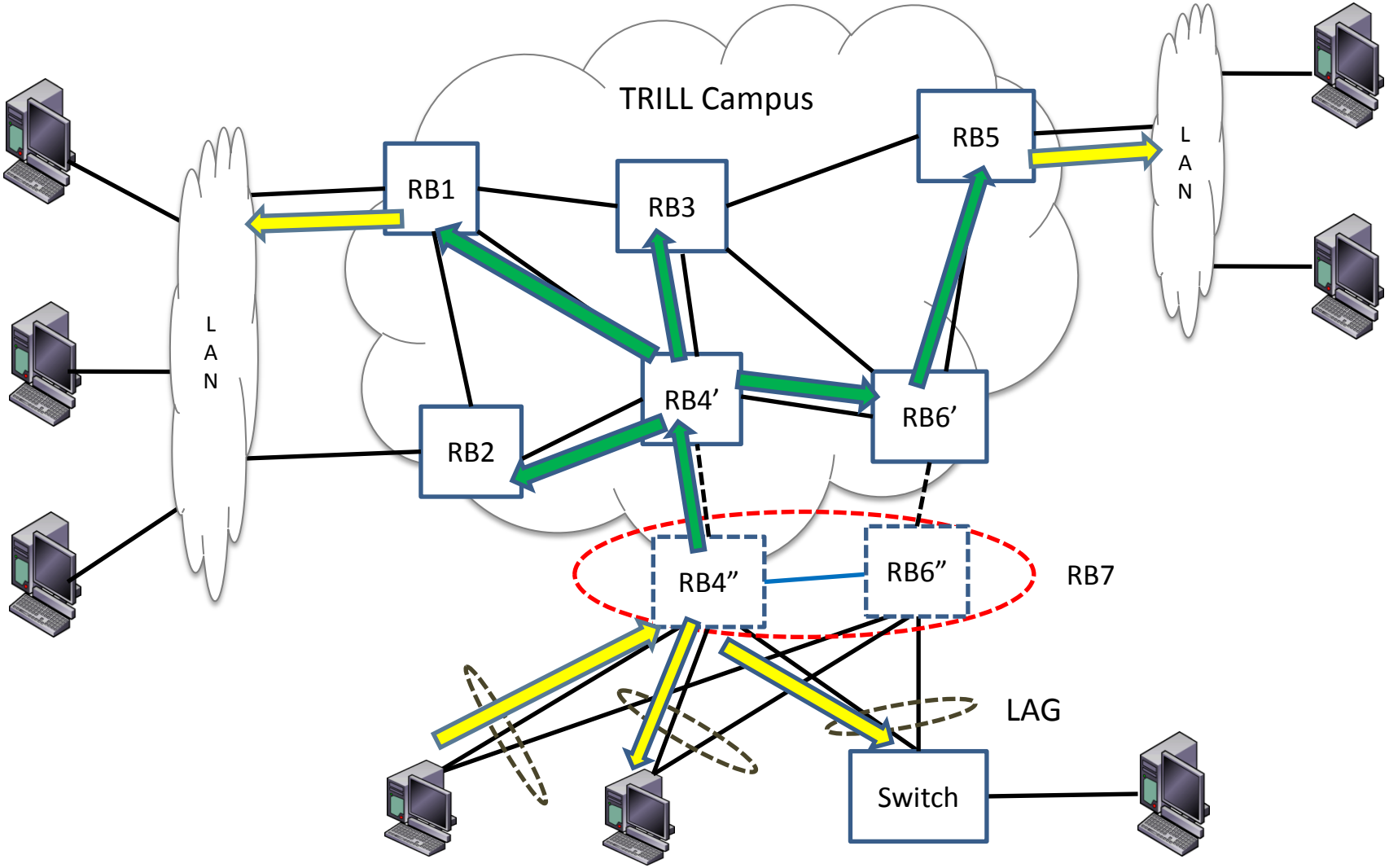
Scenario 1: If a local t-LAG link fails



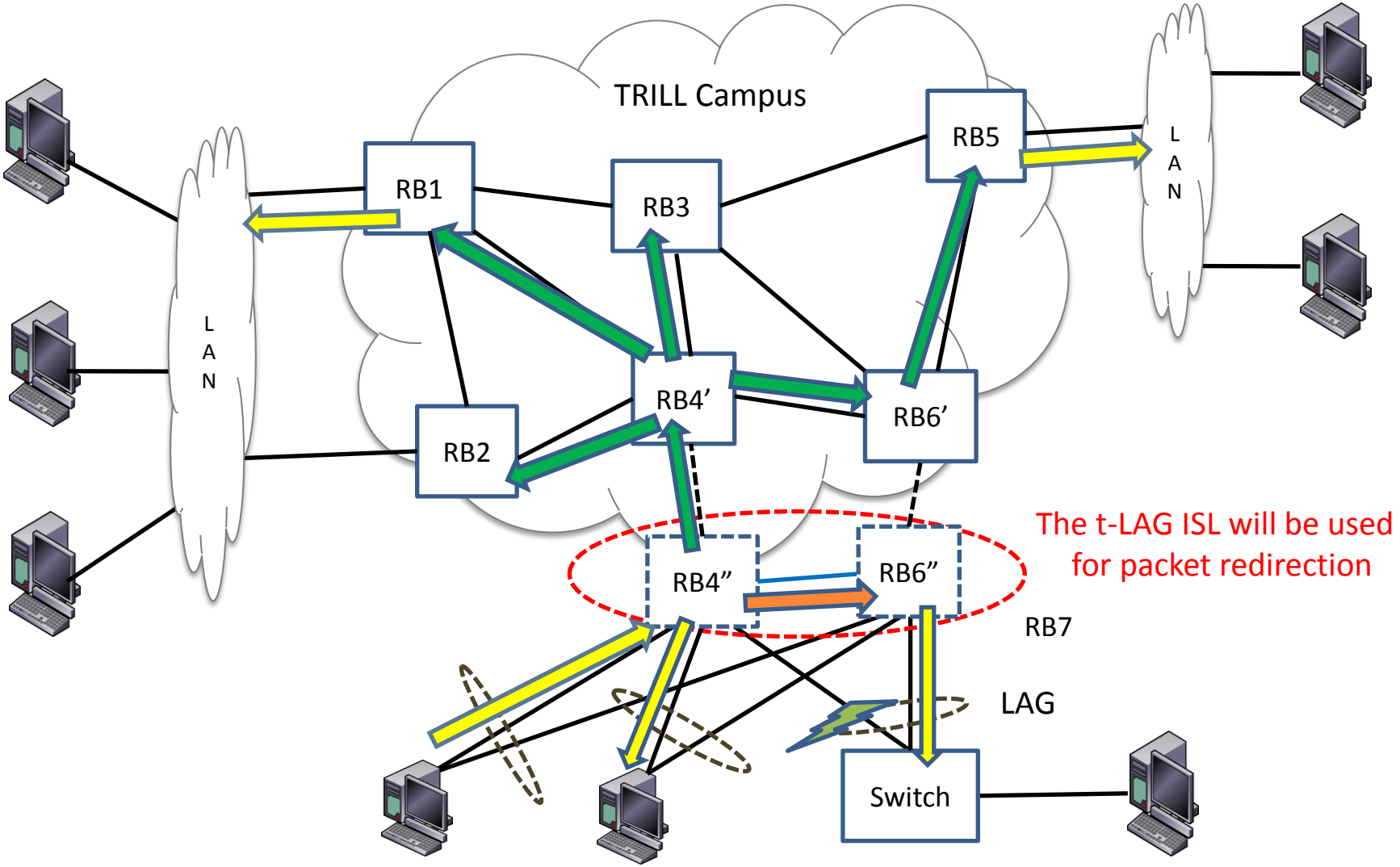
Scenario 1: If too many t-LAG link fails



Scenario 2: The normal packet flow for MC/BC/DLF traffic ingress at a t-LAG



Scenario 2: If a local t-LAG link fails



Thank You