Cloud and Datacenter Networking

Università degli Studi di Napoli Federico II
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI
Laurea Magistrale in Ingegneria Informatica

Prof. Roberto Canonico

Datacenter networking infrastructure

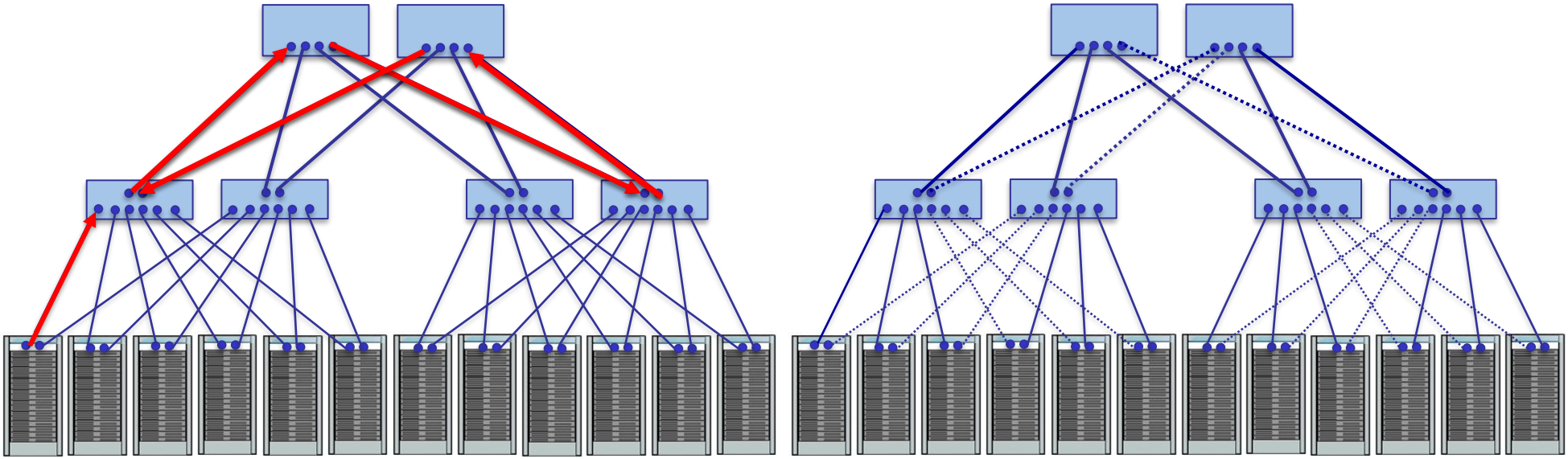
Part III

Lesson outline

- Datacenter networks and loops
  - STP and why STP is not the solution for DC networks
- Datacenter networks and alternate paths exploitation
  - TRILL protocol
    - Paper: Radia Perlman and Donald Eastlake. *Introduction to TRILL*. The Internet Protocol Journal, Volume 14, No. 3, pp. 2-20, September 2011
  - ECMP

Network topologies and loops

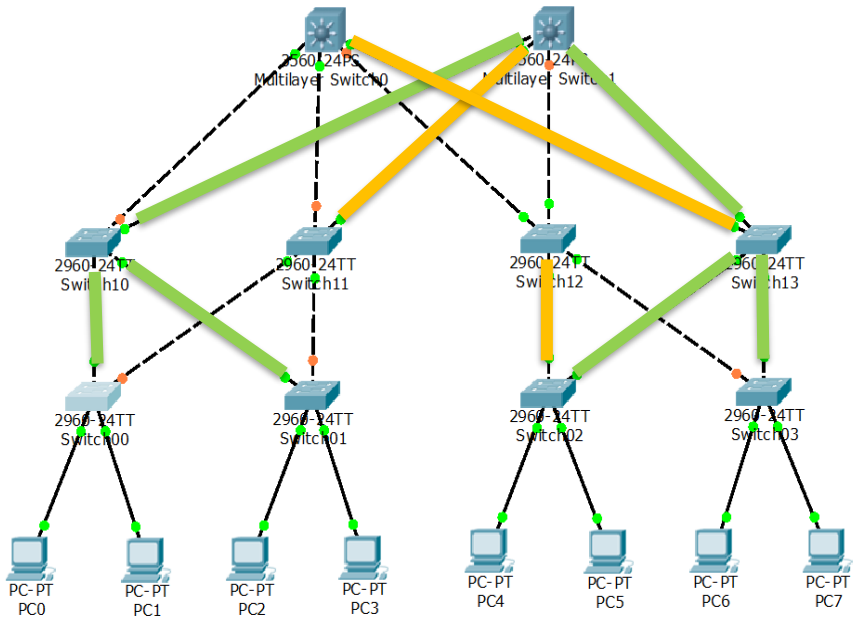
- This network topology has a problem: loops !

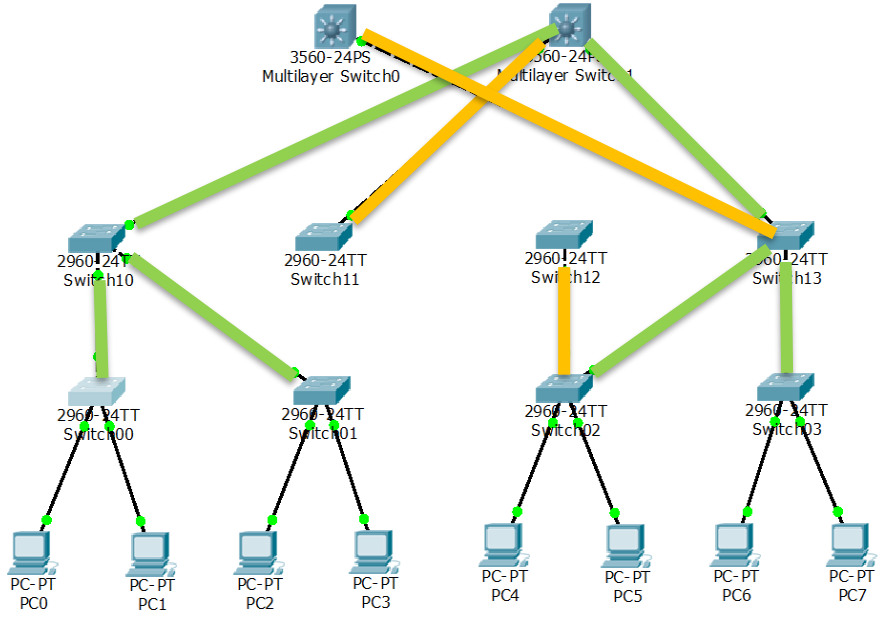


- Switches operate at layer 2 (no TTL)
- Mechanisms are needed to avoid that packets entering a loop are forwarded forever
- Traditional loop-management approach in Ethernet networks: STP
  - IEEE 802.1D standard protocol invented by Radia Perlman
  - Conceived for campus-like networks
  - To cut loops, network topology is transformed into a tree (*loop-free topology*) by disabling a subset of links
  - Inconvenience: only a fraction of network capacity may be utilized in this way and oversubscription is not reduced
- Alternative solutions: TRILL, FabricPath (Cisco), VCS (Brocade), M-LAG, QFabric, SPB, ...

Example of a DC network with STP in action

- Switches decide which interfaces should be switched off to prevent loops
- One end of a disabled link is turned off while the other is still on
- This results in a *spanning tree* connecting all the end systems as well as all the switches without any loop





STP: Spanning Tree Protocol

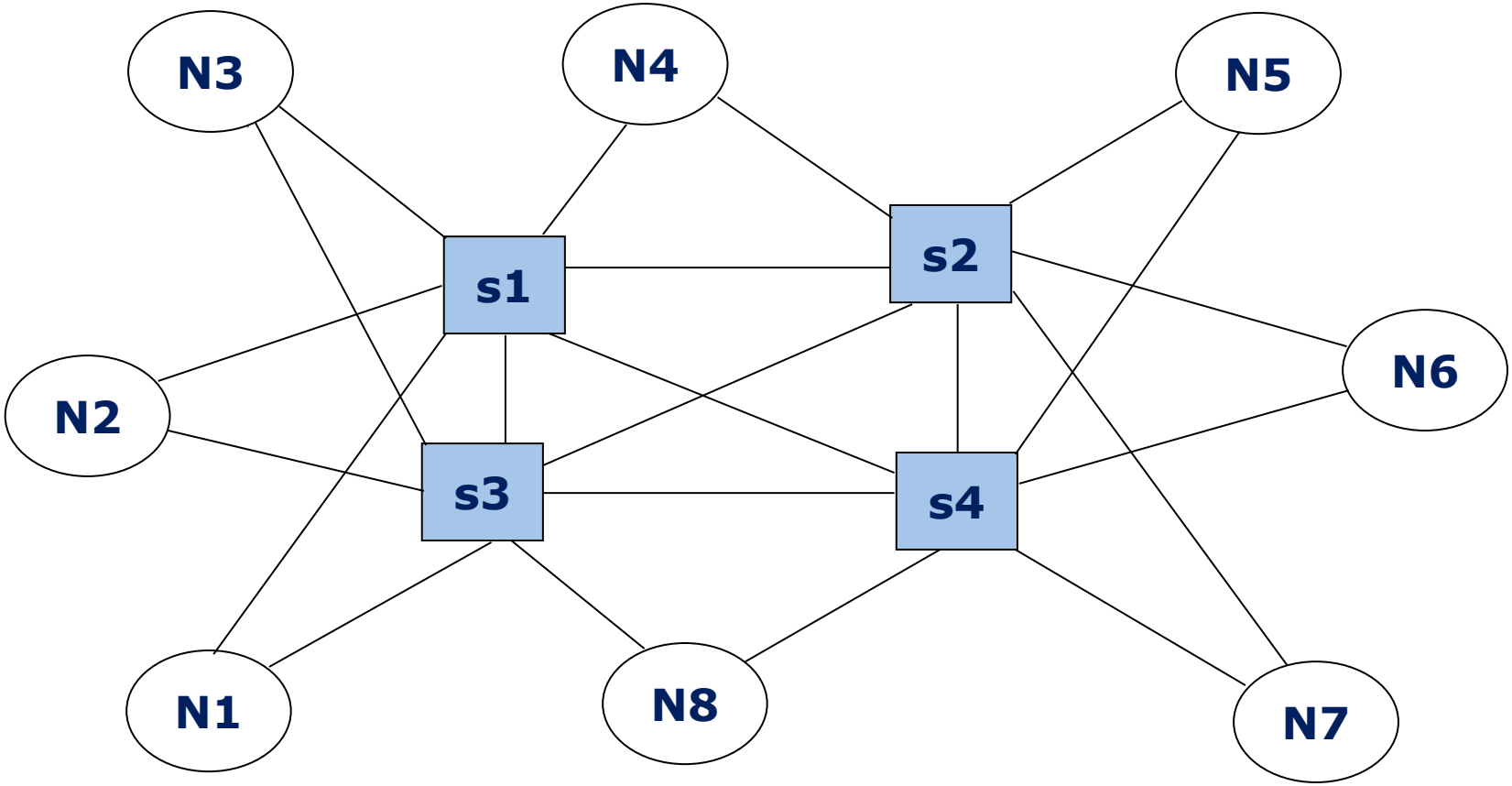
- Switches periodically exchange Configuration *Bridge Protocol Data Units* (BPDUs) to build the topology database
  - BPDU's are forwarded out all ports every 2s, to the dedicated MAC multicast address of 01:80:C2:00:00:00
  - Configuration BPDUs contain the switch *bridge ID*
- Each bridge starts out thinking it is the Root bridge
- Eventually, all switches agree that the Root bridge is the switch with smallest bridge ID
- Through BPDU exchanges, tree converges, which means all switches have same view of the spanning tree
- Each port of a switch may be in one of the following states:
  - Forwarding, Blocking, Listening, Learning

STP: how it works

STP creates a tree that provides a single unique path to each destination as follows:

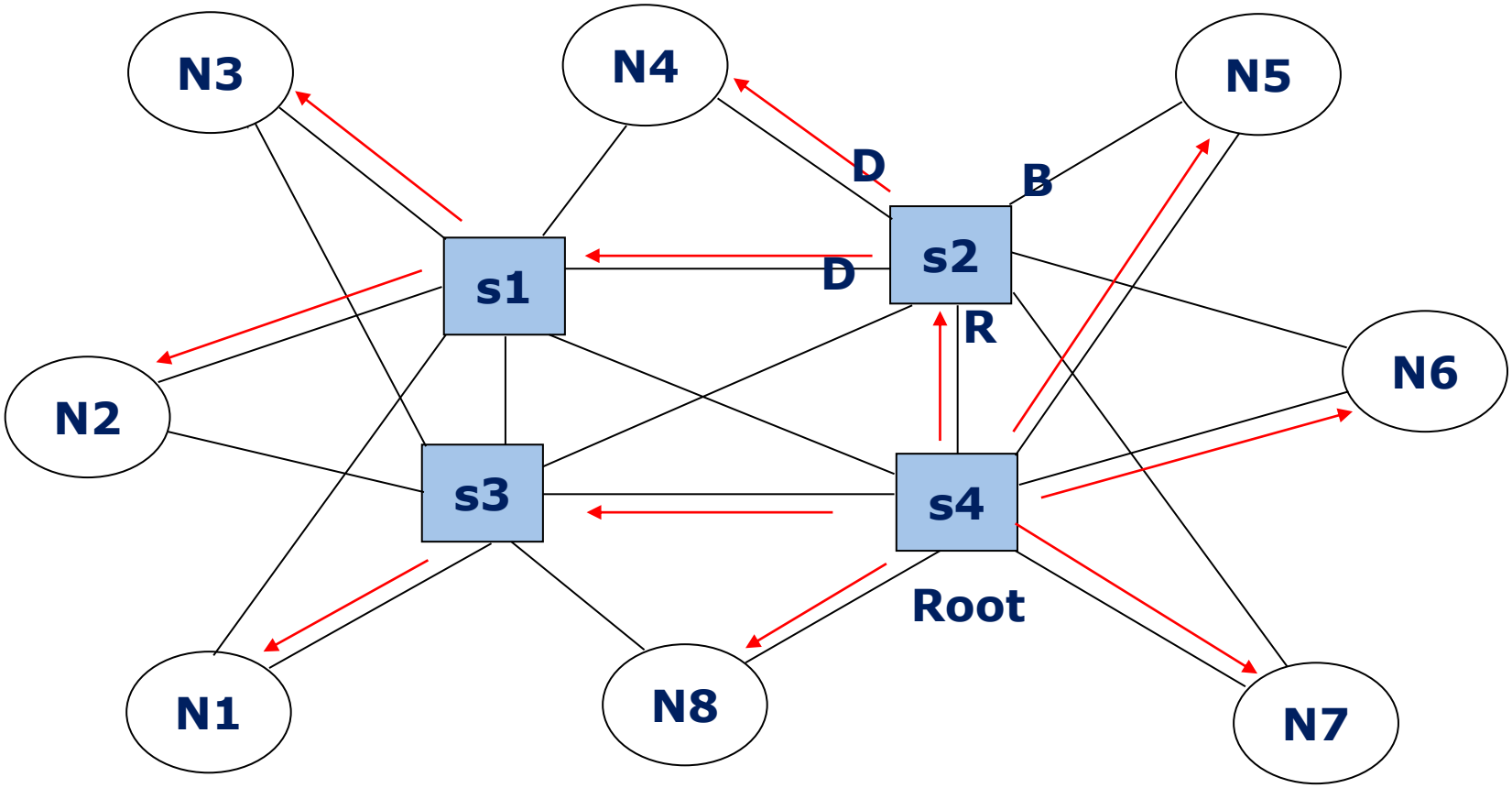
- switches elect a root bridge acting as the root of the spanning tree
- each bridge calculates the distance of the shortest path to the root bridge
- each bridge determines a *root port*, which will be used to send packets to root
- for each segment, a *designated port* is identified, i.e. the port closest to the root
- root ports and designated ports are set to *forwarding* state;
  all other ports are set to *blocking* state
  - Packets will not be received or forwarded on blocked ports

Spanning Tree Protocol in action (1)



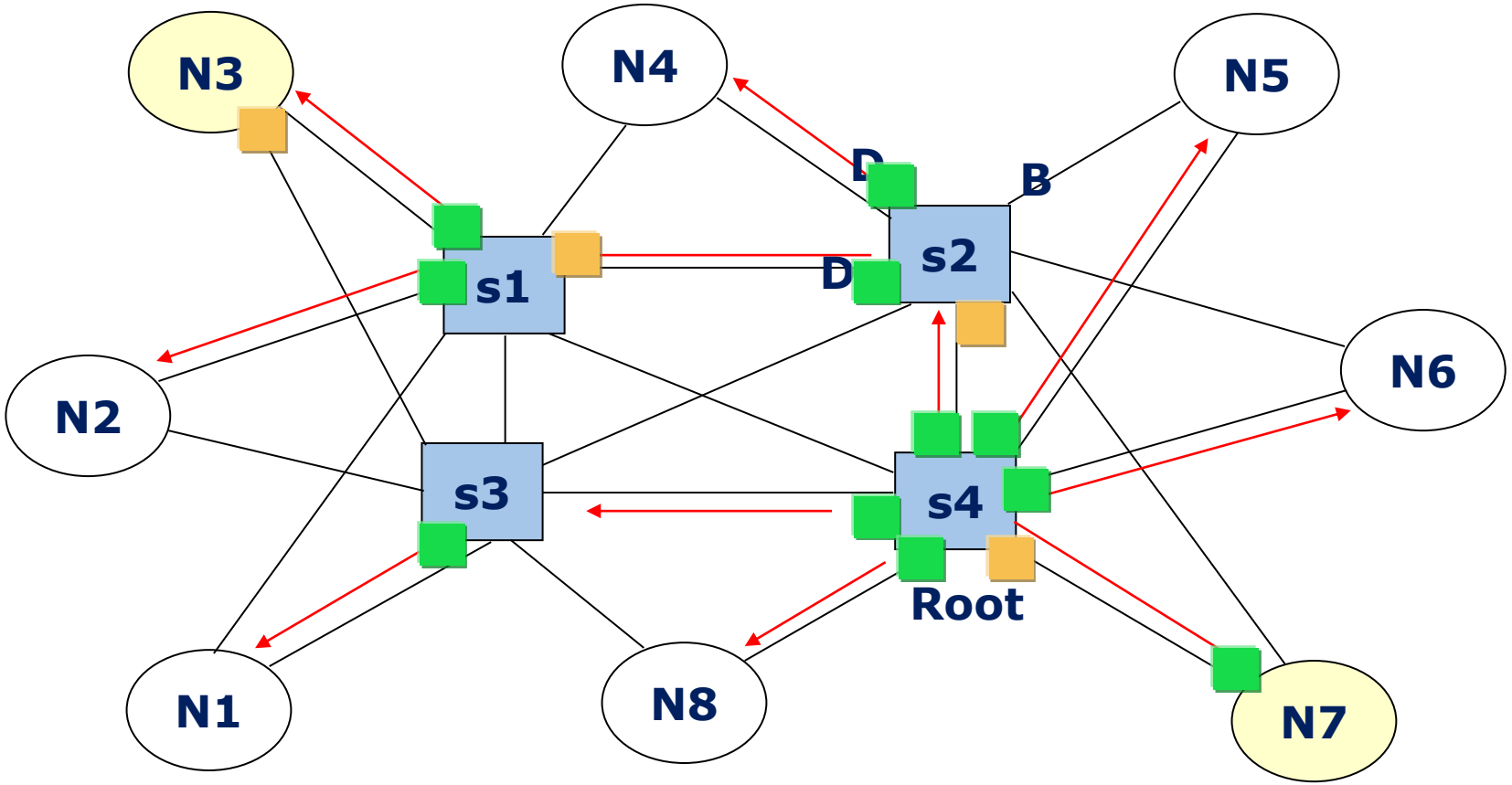
Spanning Tree Protocol in action (2)

- Having elected s4 as the Root Bridge, this is the resulting spanning tree
- Red arrows show the path of a broadcast packet from the root to any possible destinations
- Links not marked with an arrow are not traversed by any packets



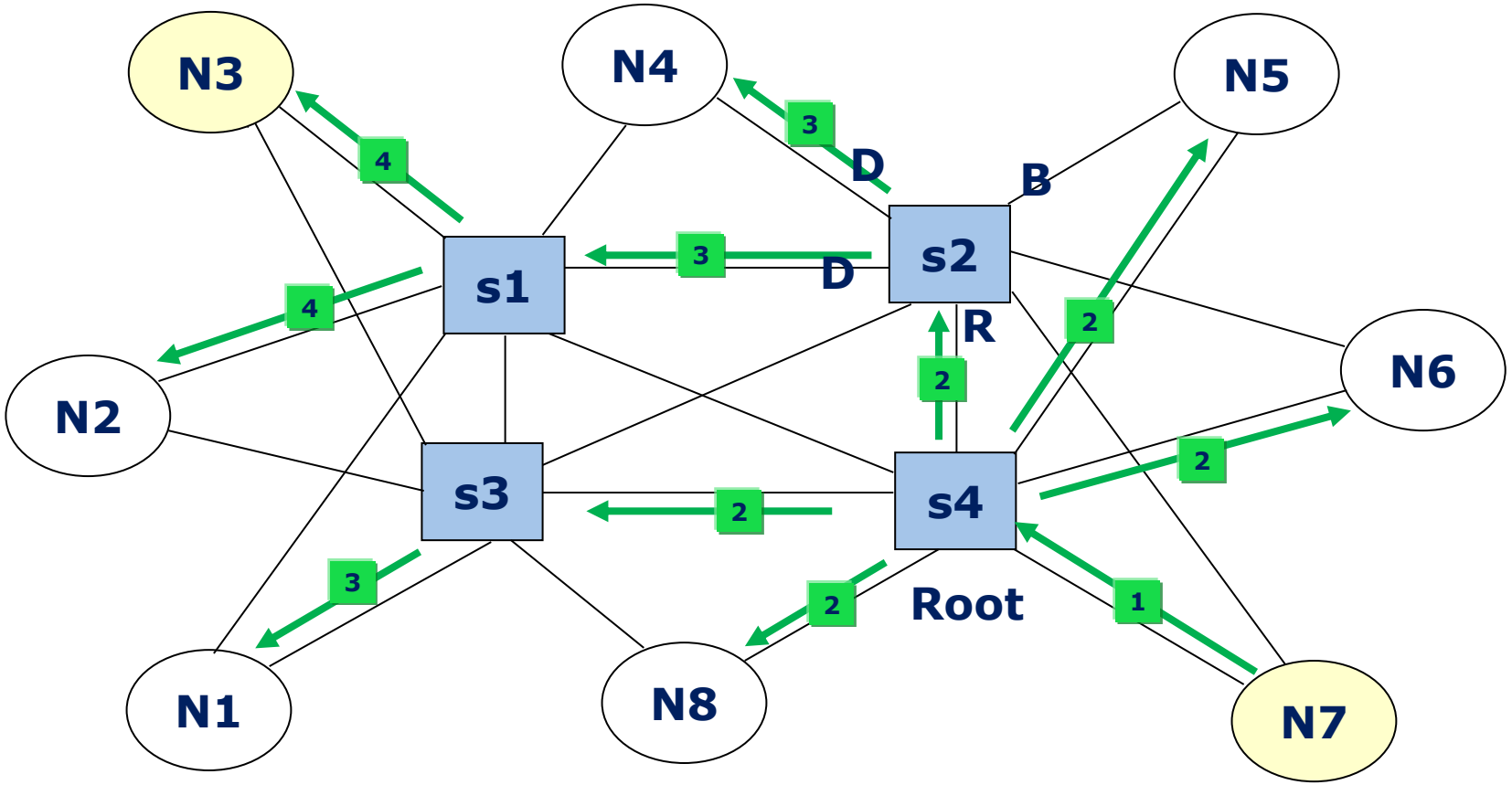
Spanning Tree Protocol in action (3)

- N7 sends in broadcast (i.e. to FF:FF:FF:FF:FF) an ARP request querying for N3's MAC
  - Packet is sent to the root s4 and from the root down the spanning tree towards all destinations
- N3 replies to N7 with its own MAC address
  - switches have learned where N7's MAC is located from the previous transmission



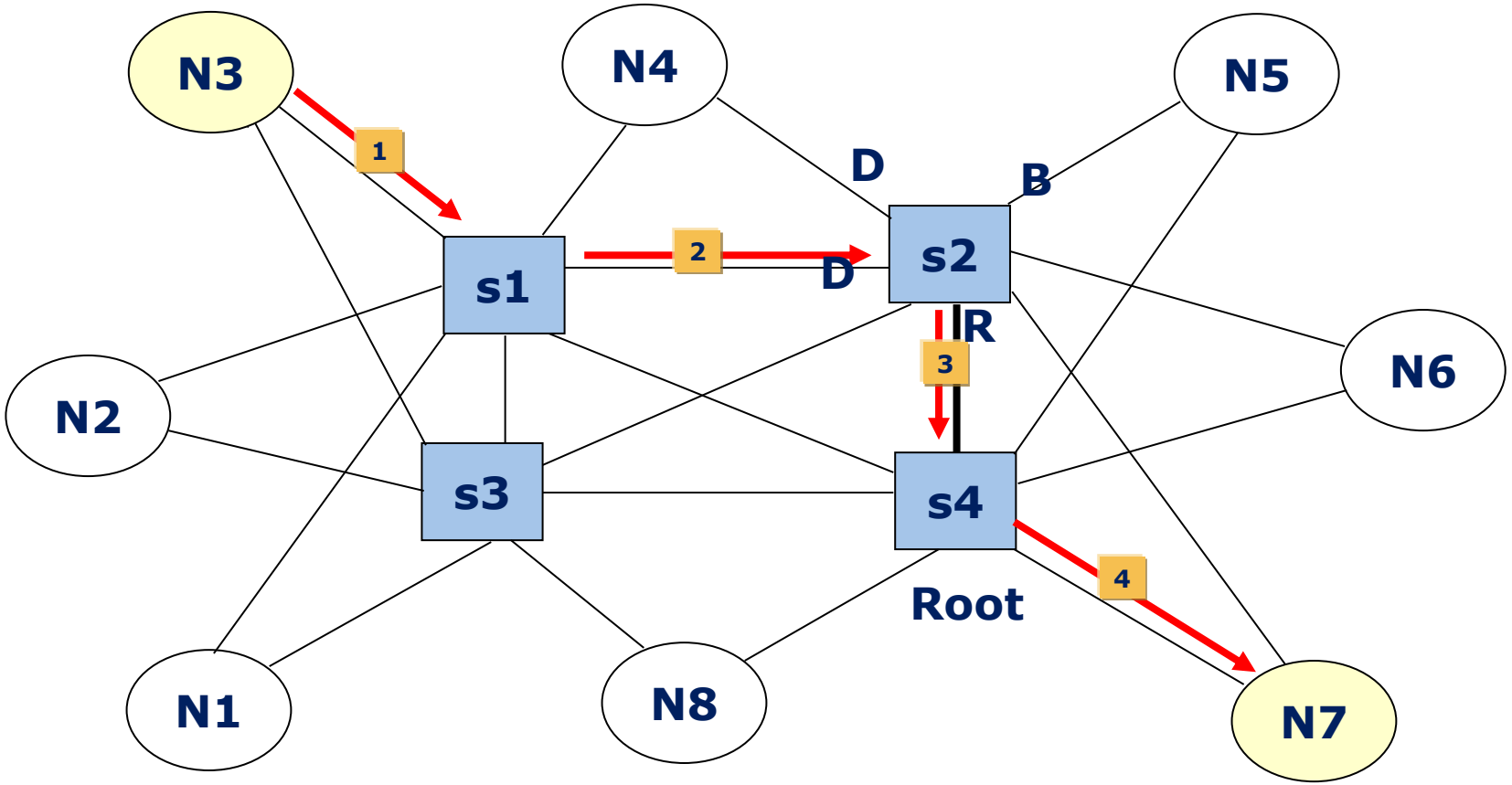
Spanning Tree Protocol in action (3)

- N7 sends in broadcast (i.e. to FF:FF:FF:FF:FF) an ARP request querying for N3's MAC
  - Packet is sent to the root s4 and from the root down the spanning tree towards all destinations
- N3 replies to N7 with its own MAC address
  - switches have learned where N7's MAC is located from the previous transmission



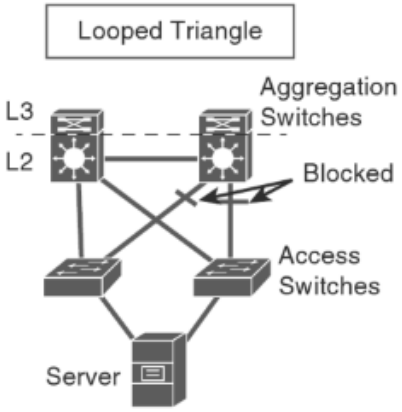
Spanning Tree Protocol in action (4)

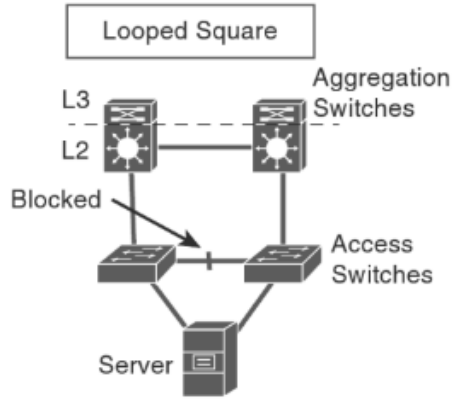
- N7 sends in broadcast (i.e. to FF:FF:FF:FF:FF) an ARP request querying for N3's MAC
  - Packet is sent to the root s4 and from the root down the spanning tree towards all destinations
- N3 replies to N7 with its own MAC address
  - switches have learned where N7's MAC is located from the previous transmission

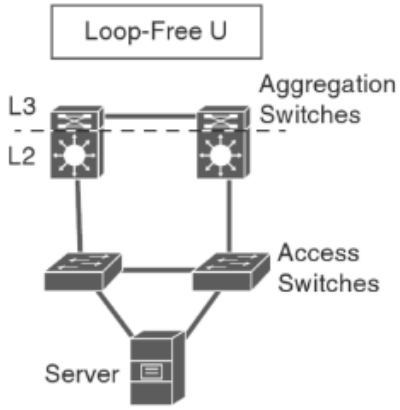


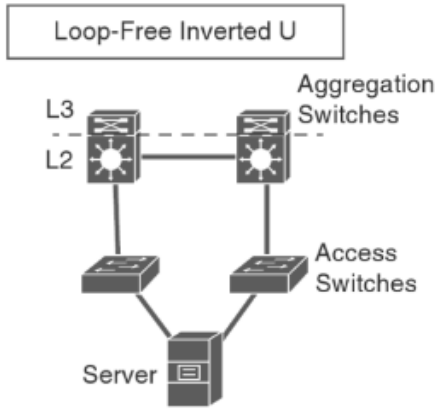
Access-Aggregation connection options

- Looped Triangle topology
  - STP blocks 2 uplinks out of 4
- Looped Square
  - STP blocks 1 horizontal link
  - In case of failure of an uplink, traffic is routed to the adjacent access switch → oversubscription doubles
- Loop-Free U
  - Communication between aggregation switches is L3
  - No loops → no links blocked by STP
- Loop-Free Inverted U









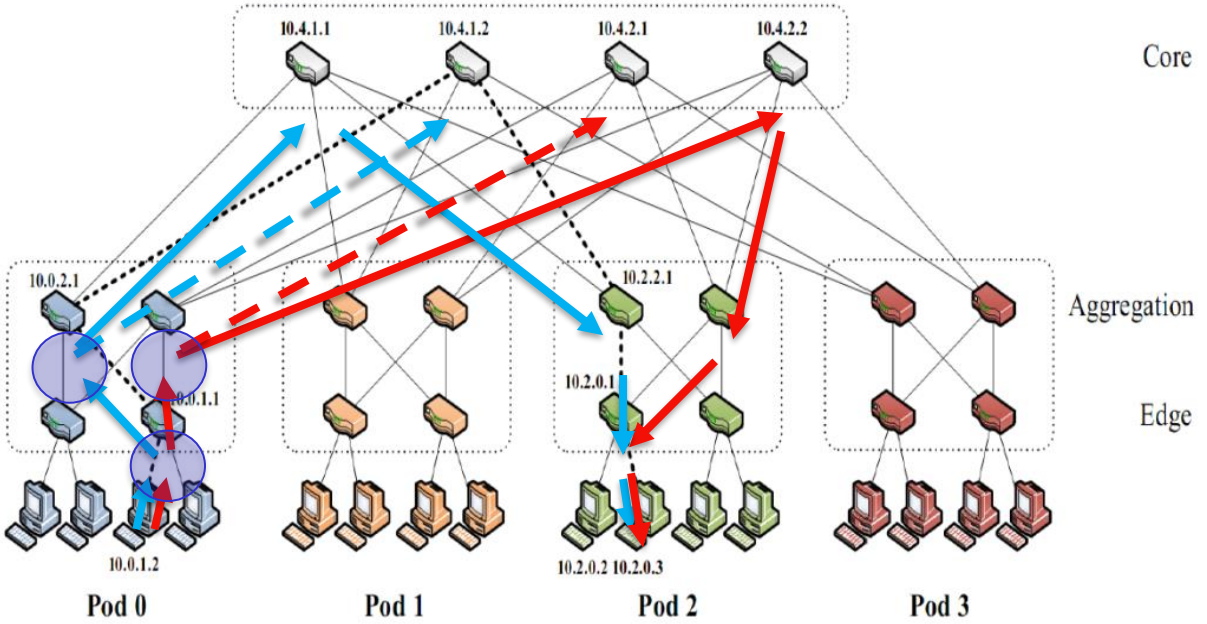
Source: Cloud Computing: automating the virtualized data center. Gustavo A. A. Santana. CISCO Press (2014)

Datacenter networks and alternate paths exploitation

- We have seen that DC networks are not single-root trees → alternate paths exist between any couple of hosts located in different racks
- If STP is used, the network topology is transformed into a loop-free tree by selectively disabling links → only a fraction of the network capacity is used
- To exploit alternate paths more complex solutions are needed that exploit some form of topological knowledge: ECMP
  - ECMP stands for Equal Cost MultiPath
- ECMP across L2 networks is supported by:
  - Transparent Interconnection of Lots of Links (TRILL)
  - Shortest-Path Bridging (SPB)
- TRILL: IETF standard born for complex campus networks
  - A layer-2.5 solution that can be incrementally deployed
  - TRILL combines techniques from bridging and routing

Multi-path routing: ECMP

- In a datacenter with multiple possible paths between any (source,destination) couple ECMP allows to randomly spread traffic over alternative paths



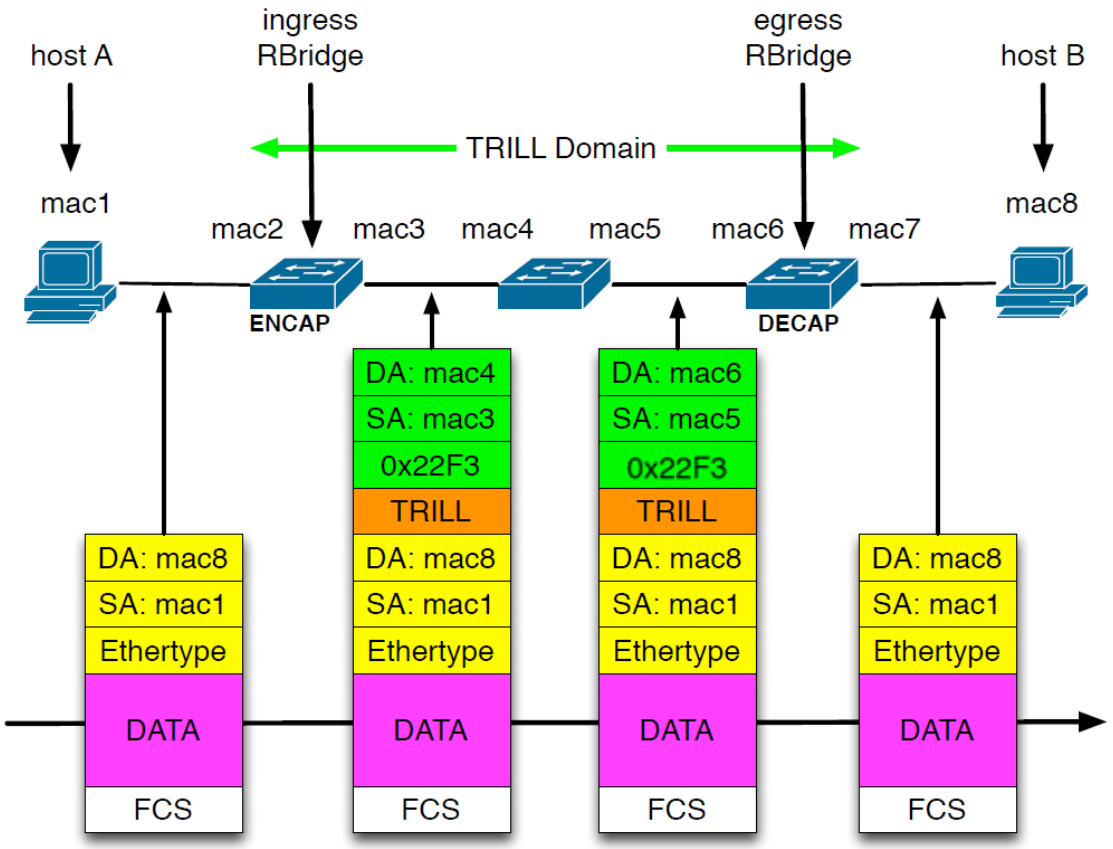
At the first edge switch, traffic from 10.0.1.2 to 10.2.0.3 is randomly routed either on the left path or on the right path
Also aggregation switches may randomly choose one among two different paths

- If upper layer switches have internal queues filled up differently, packets may arrive mis-ordered to destination → TCP performance degrades
- To avoid this problem, packets of the same flow need to be routed on the same path

TRILL: Transparent Interconnection of Lots of Links

- TRILL relies on special switches called R-Bridges
- R-Bridges run a link-state protocol:
  - learn the network topology through the exchange of *Link State Packets* (LSPs)
  - compute shortest path tree between them
- The link-state protocol used by TRILL is IS-IS
  - IS-IS was originally defined as an ISO/OSI standard (ISO/IEC 10589:2002) and later described in IETF RFC1142
  - IS-IS chosen because it runs directly over Layer 2, so it can be run without configuration
    - no IP addresses need to be assigned
- TRILL switches are identified by 6-byte IS-IS System ID and by 2-bytes nicknames
- TRILL is compatible with existing IP Routers: R-Bridges are transparent to IP routers
- R-Bridges encapsulate each packet they receive from hosts with a header bringing the ID of the next-hop R-Bridge in the shortest path to the destination
  - the R-bridge which is closest to the destination decapsulates the packet before delivering it to the destination
- TRILL data packets between R-Bridges have a **Local Link header** and a **TRILL header**
- For unicast packets:
  - Local Link header contains the addresses of the local source R-Bridge to the next hop R-Bridge
  - TRILL header specifies the first/ingress R-Bridge and the last/egress R-Bridge
- A 6-bits hop count is decreased at each R-Bridge

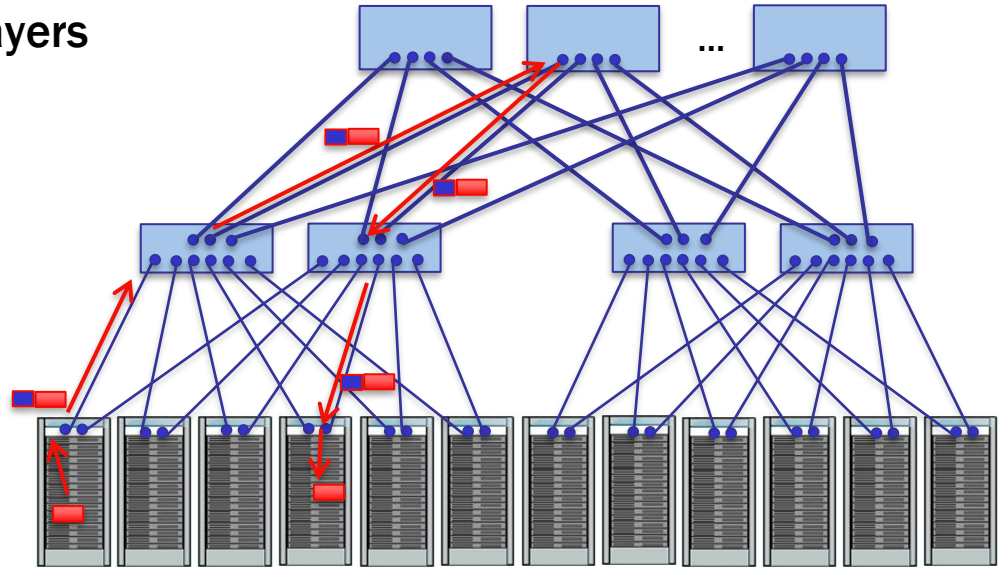
TRILL packet forwarding



- Figure by Ronald van der Pol, from "TRILL and IEEE 802.1aq Overview" (Apr.2012)

TRILL in the datacenter

- TRILL implemented in the switches at all layers
- Hosts speak “pure IEEE 803.3” (Ethernet)
- In order to learn end-hosts’ identity (i.e. association to edge R-Bridge), a directory service is needed
- Leaf switches encapsulate each packet they receive from hosts with a header bringing the ID of the next-hop R-Bridge in the shortest path to the destination



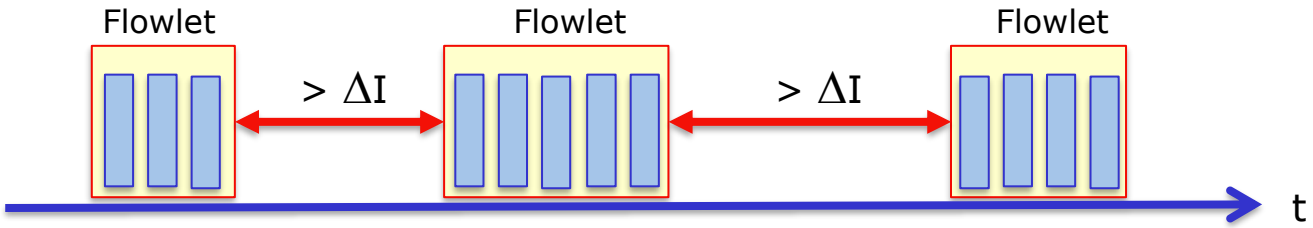
- The R-bridge which is closest to the destination decapsulates the packet before delivering it to the destination
- If multiple equal costs paths are presents towards a destination, an RBridge can distribute traffic over those multiple paths
- TRILL is implemented by two vendors: Cisco (FabricPath) and Brocade (VCS)
  - Cisco FabricPath: proprietary implementation of TRILL
  - Brocade Virtual Cluster Switching: uses TRILL data plane but a proprietary control plane

ECMP and flow hashing

- To avoid misordered delivery of packets belonging to the same flow, ECMP calculates the hash of the packet header to determine the output port at each switch
- In this manner, packets of the same flow, i.e. with same (source, destination), follow the same path and are not misordered
- Works well for a large number of small flows → traffic is evenly distributed
- If multiple long-lasting flows are mapped onto the same ports, this technique may lead to an unbalance of traffic flows
- This problem arises because, actually, the concept of flow above is too coarse
- To avoid this problem and achieve a more fine-grained balancing of traffic, randomization may occur at micro-flow or *flowlet* level

ECMP and flowlets

- A *flowlet* is a sequence of consecutive packets whose inter-arrival is smaller than the conservative estimate of latency difference between any two paths within the datacenter network
  - If two flowlets are routed along different paths, no misordered delivery may happen anyway



- Flowlet-based routing first proposed in FLARE in 2007
- Flowlet-to-path mapping is performed by using a hash table whose entries are

  (hash_key, last_seen_time, path_id)
- When a packet arrives, FLARE computes a hash of

  source IP, destination IP, source port, destination port
- and uses this as the key in the hash table

[FLARE] Srikanth Kandula, Dina Katabi, Shantanu Sinha, and Arthur Berger. *Dynamic load balancing without packet reordering*. ACM *SIGCOMM Comput. Commun. Rev.* 37, 2, pp. 51-62, March 2007

ECMP issues: local decisions

- One issue with ECMP is that it only takes local decisions without any knowledge of further links status
- In this example topology, once the path has been pinned to the core switch, there's no further alternative to a given destination (i.e. only one path)
- If a link fails, ECMP can do nothing to prevent upstream switches to select the path that contains that link, even if an alternative path exists