Link Aggregation – IEEE 802.3ad

Fulvio Risso

Politecnico di Torino

Based on chapter 8 of M. Baldi, P. Nicoletti, "Switched LAN", McGraw-Hill, 2002, ISBN 88-386-3426-2 and on an existing presentation of Mario Baldi and Piero Nicoletti

Link aggregation (1)

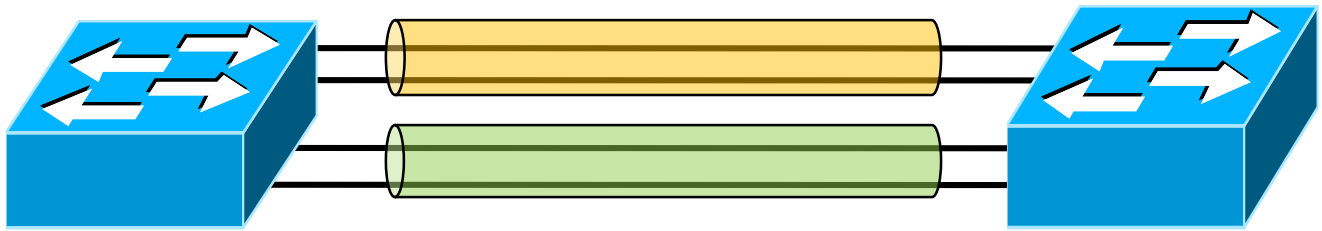
- Solution that aggregates several ports into a single channel
  - STP issues in case of multiple links between the same devices
  - Incremental increase in bandwidth capacity
    - 10x speed may be too costly or may not exist
    - Sometimes 10x speed not required
  - Useful to improve resiliency
    - Smooth decrease of bandwidth in case of fault
      - Unlikely to break all the cables at the same time
    - Avoids STP convergence time in case of a failure of a single link
- Normally used between switches, occasionally between a switch and a computer

Link aggregation (2)

- Several solutions existed before the 802.3ad standard
  - Static configuration
    - What about errors?
      - E.g., not all the links are between the same switches
      - E.g., the maximum number of channel is different in switches A and B
  - Proprietary solutions (e.g. Cisco EtherChannel)
- 802.3ad
  - Standard solution that aggregates several ports into a single channel
  - Autoconfiguration
    - Devices are able to recognize the number of available links in the channel, and whether the connection with the other party is correct (i.e. all the links are between the same switches)

802.3ad

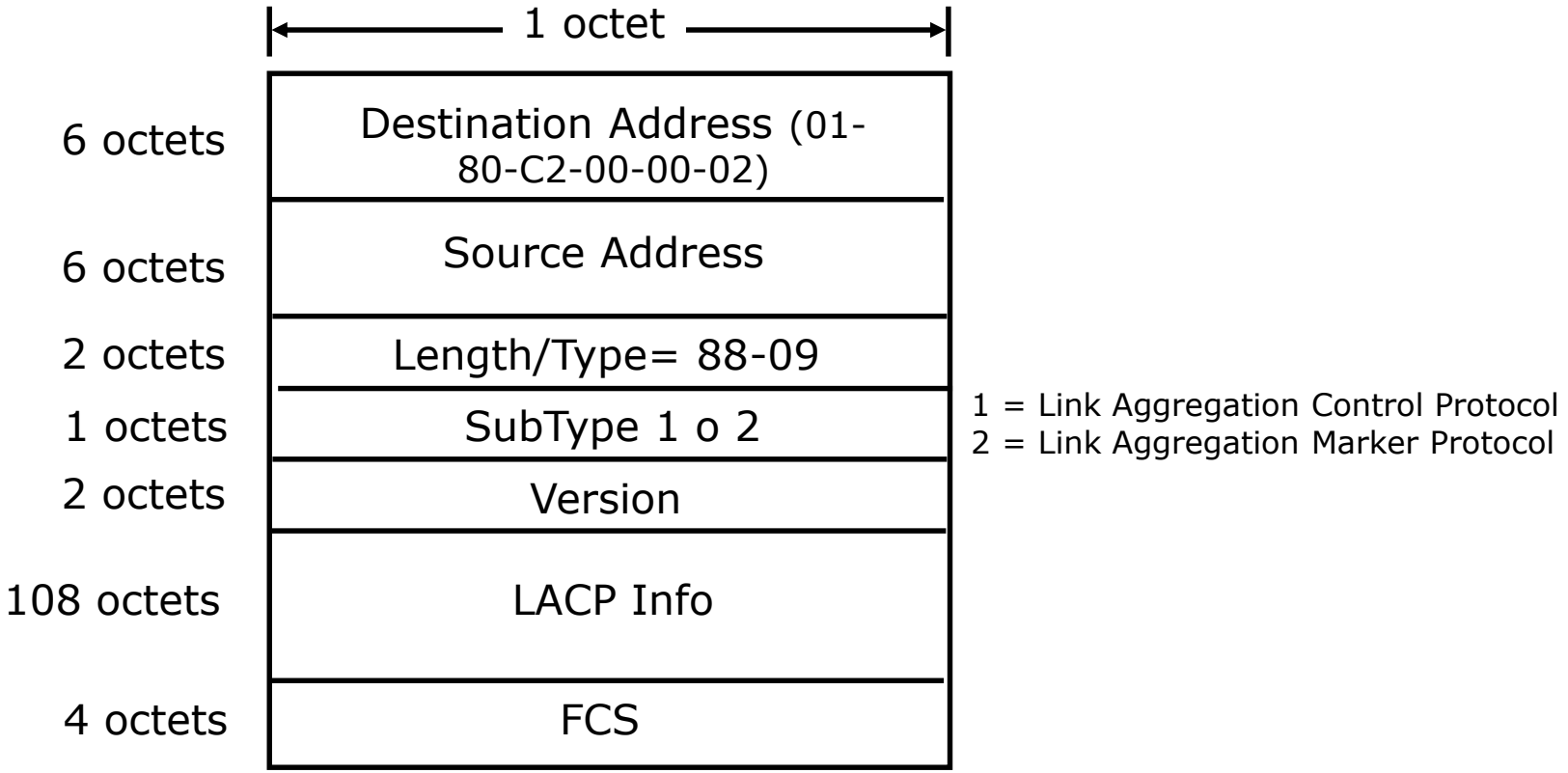
- Configuration details
  - Available only on full duplex links
  - All the links must terminate at the same devices
    - Not required that links are numerically in sequence
  - All the physical links within the same group must have the same speed
- Multiple physical links are grouped into a logical link
  - Usually 2-4 physical links per aggregate
    - The 10x technology may be a better choice if aggregate > 4 links
  - Multiple aggregations are possible
    - Only one aggregate will be active (due to STP)
- Fast convergence (usually less than 1s)



LACP

- Automatic aggregation configuration through LACP (Link Aggregation Control Protocol)
  - Multicast transmission of LACPDU
- LACPDUs exchanged periodically
  - When ports are connected
  - Every 3/90 seconds
  - When a link failure occurs
  - Contain both Actor and Partner parameters
- Each device will act as Actor and Partner at the same time

LACPDU


1 octet
6 octets
Destination Address (01-80-C2-00-00-02)
6 octets
Source Address
2 octets
Length/Type= 88-09
1 octets
SubType 1 o 2
2 octets
Version
108 octets
LACP Info
4 octets
FCS
1 = Link Aggregation Control Protocol
2 = Link Aggregation Marker Protocol

Link Aggregation Group Identifier (LAG ID) (1)

- Identifier of the aggregate
  - Determine if two ports share the same membership group
- Includes the following information
  - System Identifier (System Priority + MAC Address)
    - System Priority, default 32768
    - MAC address of the device or the MAC associated to the virtual switch and specified by the management
  - Operational Key
    - Associated to all the ports of the aggregate through explicit configuration by the network manager
  - Port Identifier (Port Priority + port number)
    - Unnecessary parameter in some cases, if so it is set to zero
- These parameters are also used for STP configuration

Link Aggregation Group Identifier (LAG ID) (2)

- Each devices send both its LAG ID and the LAG ID of the partner to the other endpoint

	Partner SKP	Partner TLQ
System Parameters (S, T)	System Priority = 0x8000 (see 43.4.2.2) System Identifier = AC-DE-48-03-67-80	System Priority = 0x8000 (see 43.4.2.2) System Identifier = AC-DE-48-03-FF-FF
Key Parameter (K, L)	Key = 0x0001	Key = 0x00AA
Port Parameters (P, Q)	Port Priority = 0x80 (see 43.4.2.2) Port Number = 0x0002	Port Priority = 0x80 (see 43.4.2.2) Port Number = 0x0002

The complete LAG ID derived from this information is represented as follows, for an Individual link:

[(SKP), (TLQ)] = [(8000,AC-DE-48-03-67-80,0001,80,0002), (8000,AC-DE-48-03-FF-FF,00AA,80,0002)]

Packets distribution on aggregate ports (1)

- Problems about reordering, which should be avoided
  - A bigger frame arriving first at the aggregate can terminate service after a shorter frame, send on another link
  - Packets belonging to the same conversation are sent on the same link
  - In case a conversation has to be redirected on another link (e.g. due to a link failure or configured out of the aggregation, or a new link added to the aggregation), better to introduce a small delay before starting the transmission of these packets

Packets distribution on aggregate ports (2)

- Standard does not define an algorithm to distribute packets
  - No packets segmentation and reassembling
  - More conversations over a port
  - A conversation can be moved to an another port because of load balancing or link failure
- The standard suggests possible packets distribution criteria over ports
  - Two switches of different vendors can use different packets distribution algorithms
    - Return path may be different from the outgoing path
  - It will be better if switches at the edge of an aggregation link belong to the same vendor

Packets distribution on aggregate ports (3)

- Conversation are assigned based on:
  - Source MAC Address
  - Destination MAC Address
  - Receiving port
  - Type of destination address (unicast, multicast, broadcast)
  - Length/Type value
  - Higher Layer Protocol (for example 4 Layer ports)
  - Mix of previous criteria
- More complex criteria
  - → more expensive switches
  - → we should add more intelligence (L3-L4 info on a L2 switch)
  - Not guaranteed that it performs better

Cisco distribution criteria

MAC address pairs **Source and Destination**	**Last** **2 bit**	**X-OR** **result**	**Chosen** **link**
Source MAC Address 00-00-00-00-00-01 Destination MAC Addr. 00-00-00-00-00-04	**01** 00	**01**	**Link 2**
Source MAC Address 00-00-00-00-00-02 Destination MAC Addr. 00-00-00-00-00-05	**10** 01	**11**	**Link 4**
Source MAC Address 00-00-00-00-00-03 Destination MAC Addr. 00-00-00-00-00-07	**11** 11	**00**	**Link 1**
Source MAC Address 00-00-00-00-00-06 Destination MAC Addr. 00-00-00-00-00-08	**10** 00	**10**	**Link 3**

Problems in packet distribution (1)

- Packet reordering must be avoided…
  - … hence packets belonging to the same conversation must be sent over the same physical link

Problems in packet distribution (2)

- Sometimes, conversations depend on a few MAC addresses
  - Case 1: a server receiving traffic from a router through two aggregated links
    - Both MAC src/dst are the same, hence probably one link unused
  - Case 2: a single, gigantic session (e.g. a nightly backup) cannot use the entire aggregate bandwidth
- In general, not always clear if “load balancing” works
  - Most of the time, the nubmer of MAC addresses in datacenters are limited, hence the balance among the different links may not be optimal
  - Often, servers send traffic out in round robin, but most routers (sending data toward the serrvers) still use the MAC-based criteria

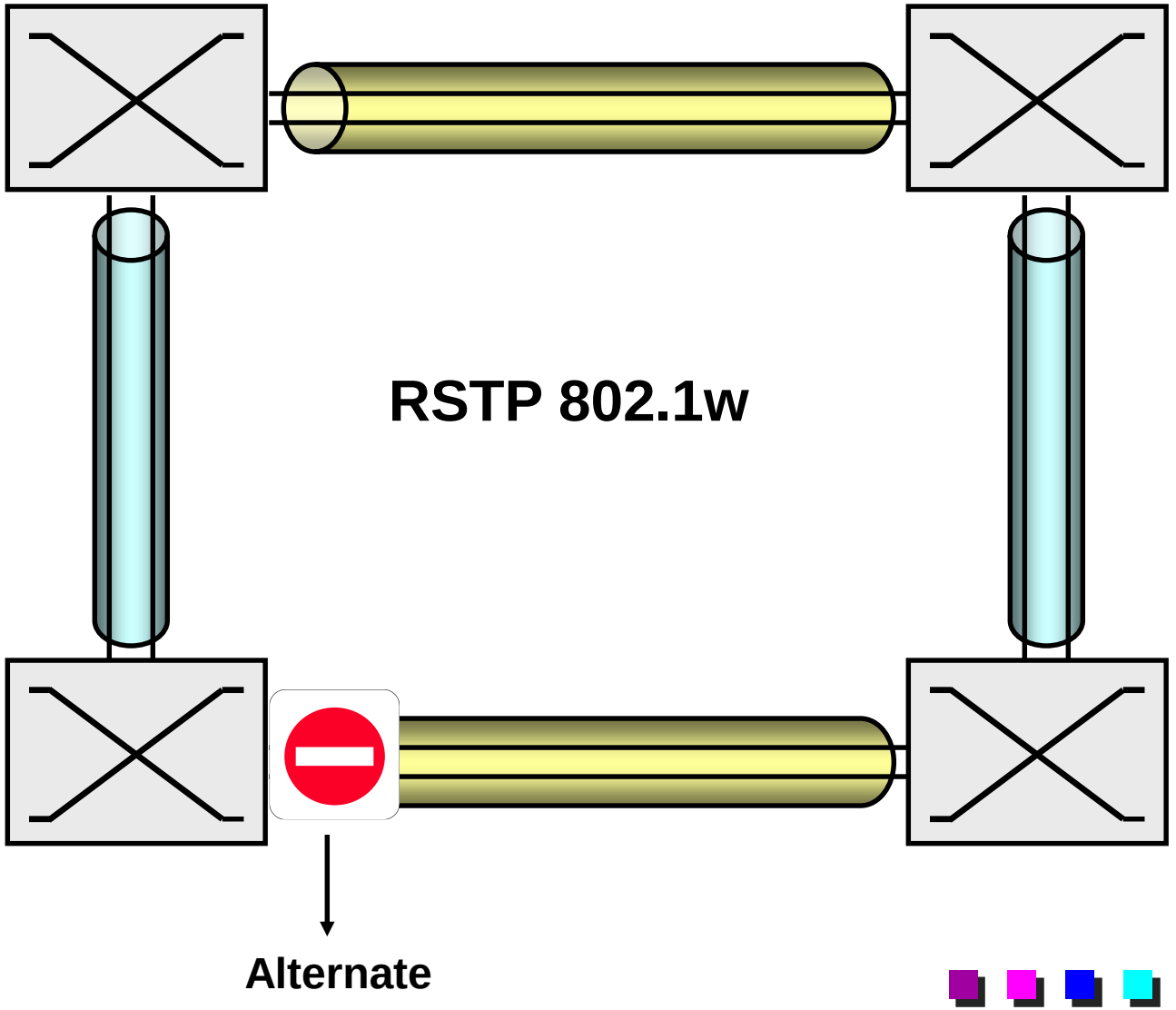
More advanced features: stand-by links

- Support the configuration of an aggregation of N links, but only M (M<N) are active
  - We do not want that a failure of a physical link belonging to the will reduce the bandwidth available in the aggregate
  - The other act as backup (or standby) links
- Possible through proper setting of link priorities
- Some more advanced configuration allows to create automatically load-balancing groups
  - Physical links can belong to the primary/secondary group dynamically
- Not really used in practice

Link Aggregation and STP/RSTP (1)

- Link aggregation is independent from STP
  - STP will no longer recognize physical links
  - Only aggregates are considered
- Be careful to set the Path Cost value to ports in order to reflect aggregate bandwidth value
  - Disable automatic path cost on ports
  - RSTP is better because of its wider range of Path Cost values

Link Aggregation and STP/RSTP (2)



Conclusions

- Advantages
  - May be useful
    - Do you need so much bandwidth?
  - Very common
  - Step-by -step increment / decrease in bandwidth (and cost)
  - Avoids STP convergence time
- Problems
  - It supports only connections between the same devices
    - A server cannot connect to two switches and achieve load sharing and resiliency
    - Some proprietary solutions (e.g. stackable switches, or Cisco Virtual Switching System) exist
  - We may not able to make use of the entire aggregate bandwidth
  - Interoperability (packet distribution algorithms)