intecs
the Brainware Company

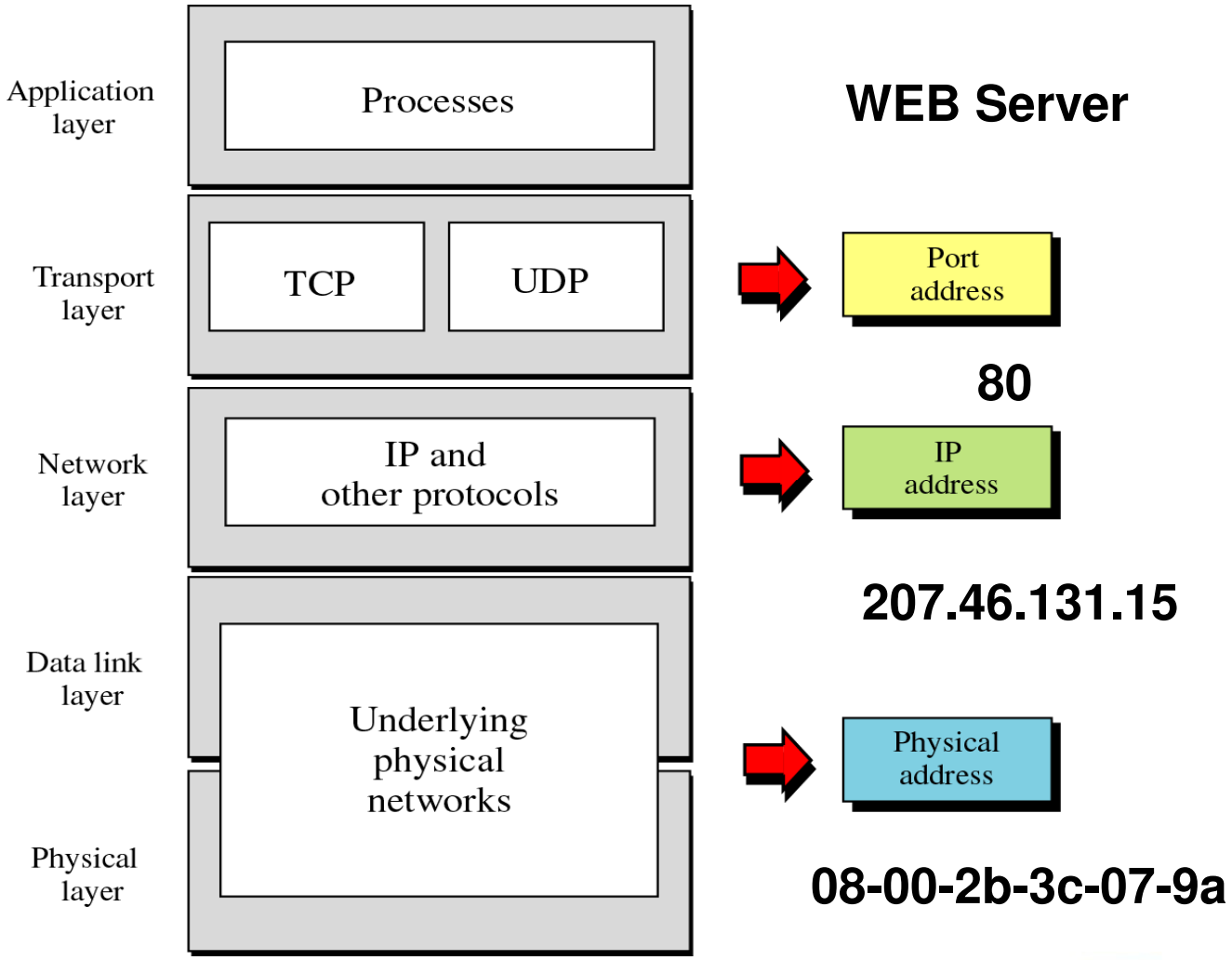
Problematiche di Switching di livello 2

TechnoLabs

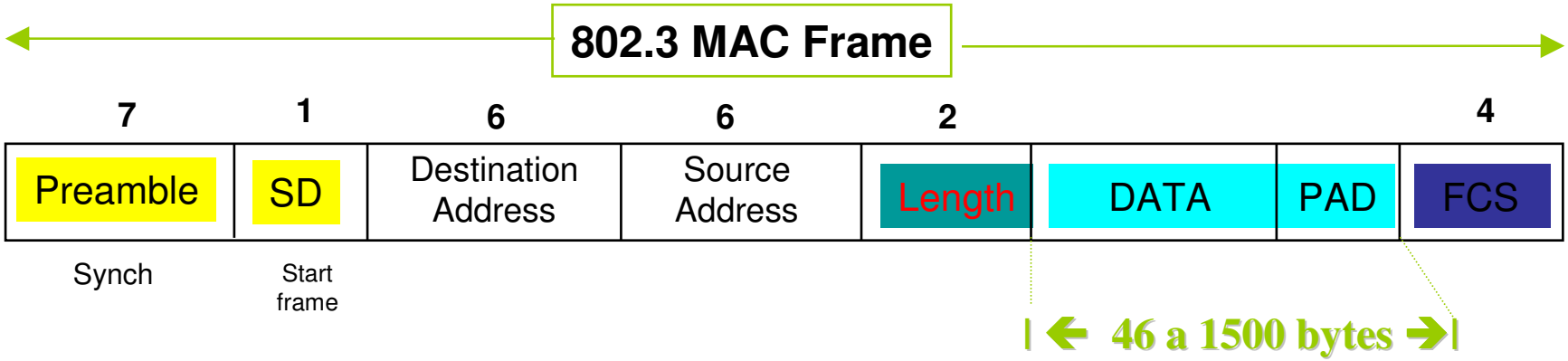
Indirizzamento e Layers

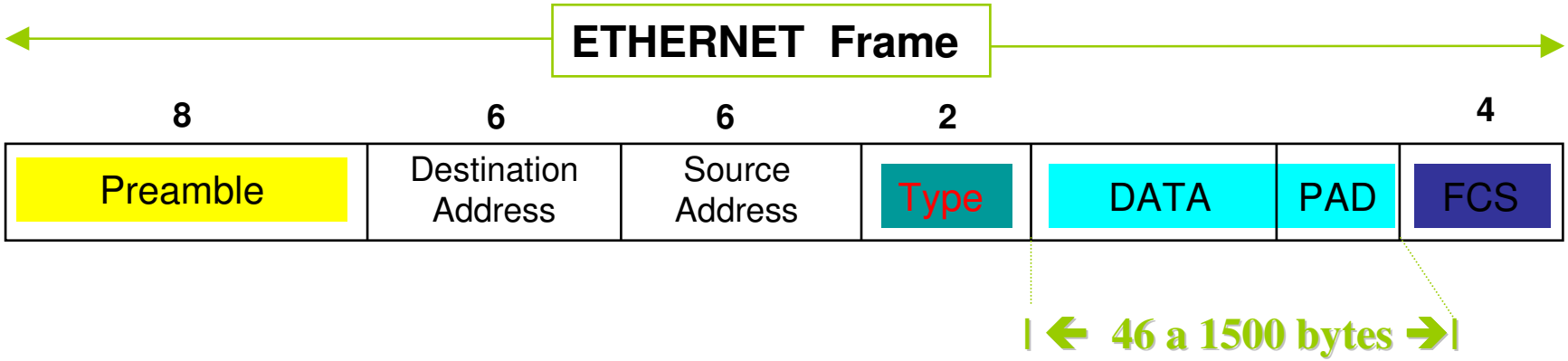
- Physical Layer: non c'è bisogno di indirizzamento
- Data Link Layer – l'indirizzamento deve garantire la possibilità di selezionare qualunque host sulla rete.
- Network Layer – l'indirizzamento deve fornire tutte le informazioni necessarie per il routing.
- Transport Layer – l'indirizzo deve identificare il processo destinazione.

Indirizzamento e Layers (cont.)



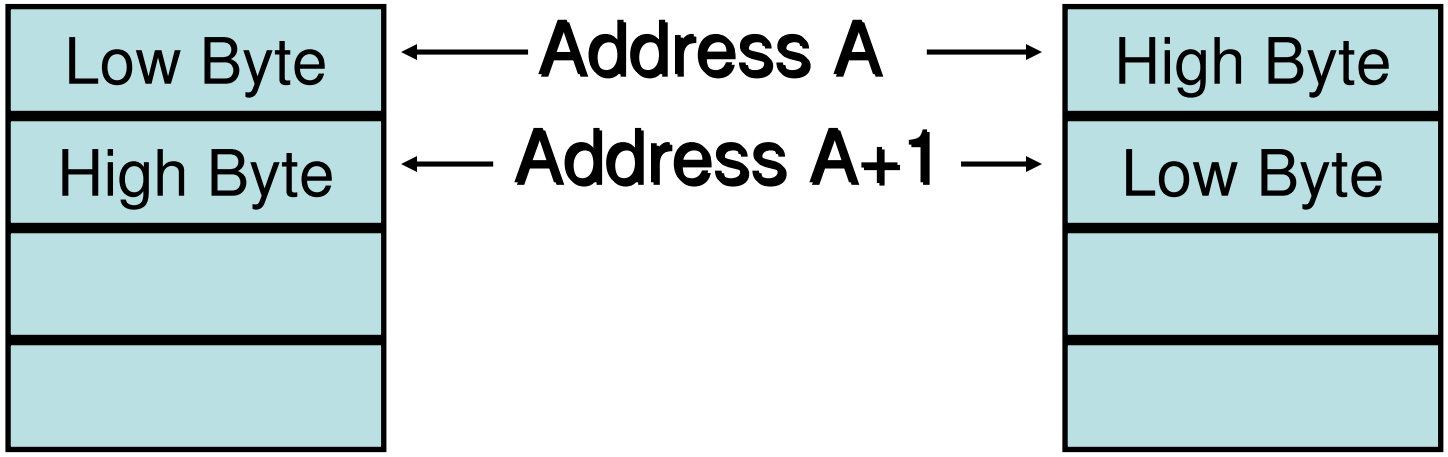
Ethernet vs 802.3





Byte Order

- Differenti architetture di computer usano differenti convenzioni per rappresentare valori che coinvolgono più byte.
- 16 bit integer:



Byte Ordering

Little-Endian

IBM 80x86

DEC VAX

DEC PDP-11

Big-Endian

IBM 370

Motorola 68000

Sun

Byte Order e Networking

- Supponiamo che una macchina Big Endian invii un 16 bit integer con il valore 2:

```
00000000 00000010
```

- Una macchina Little Endian penserà di avere ricevuto il numero 512:

```
00000010 00000000
```

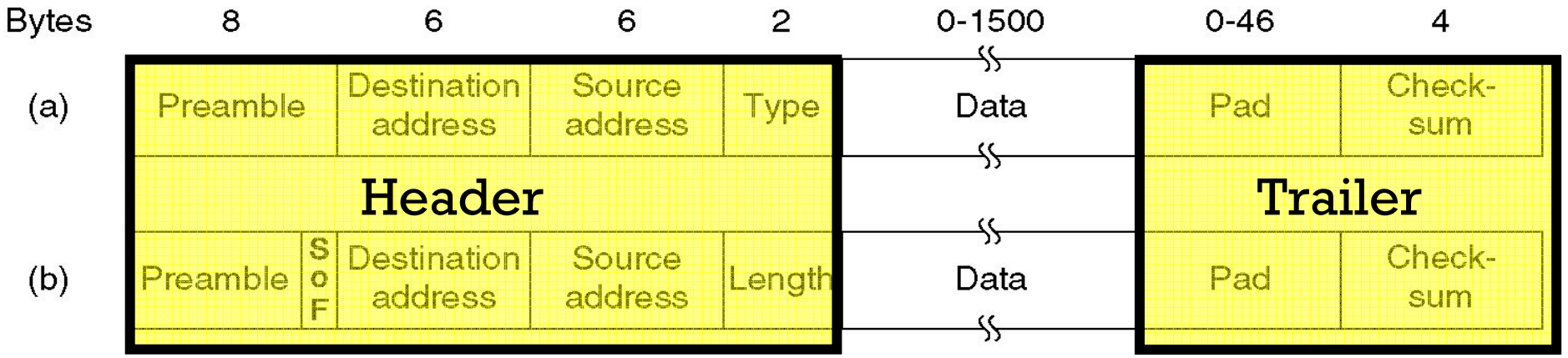
Network Byte Order

- La conversione dei dati per l' "application-level" è demandata al livello presentazione.
- Ma attenzione !!! Come comunicano i livelli più bassi se essi rappresentano i valori in modo differente ? (i.e.: il campo "data length" presente nell' header)
- É stato fissato un "byte order" (chiamato **network byte order**) per tutti i dati di controllo
  - **Il Network Byte Order è Big-Endian.**

intecs
the Brainware Company

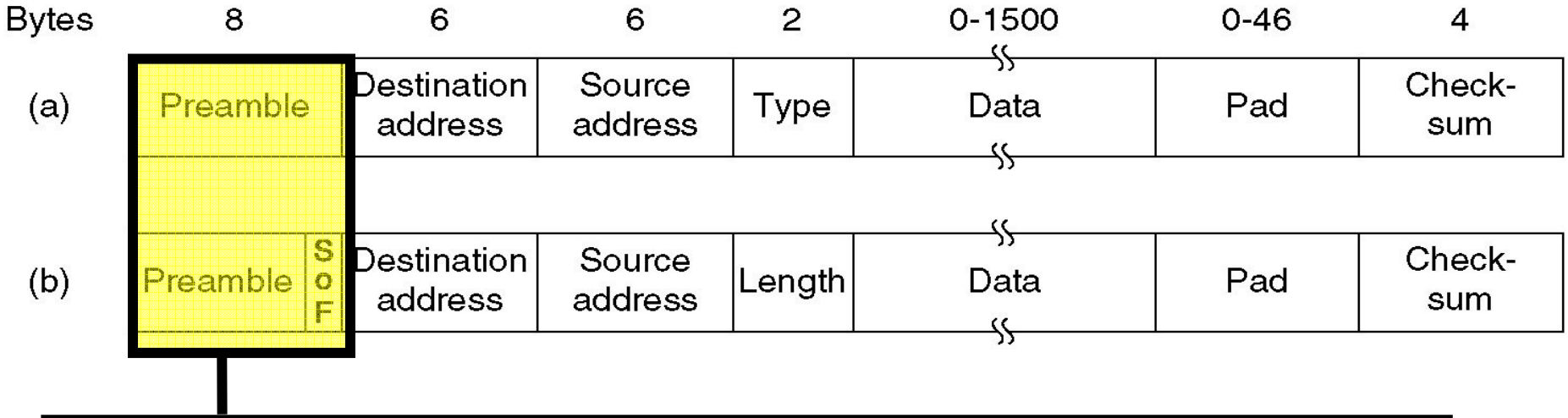
Ethernet

Ethernet MAC Protocol



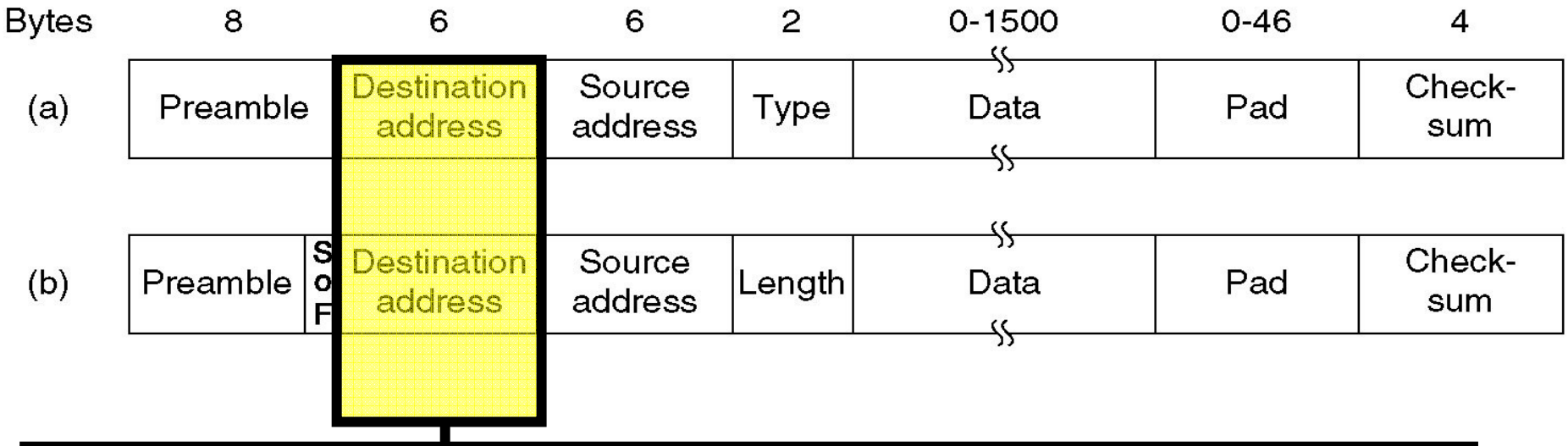
Frame formats: (a) DIX Ethernet versione 2,
(b) IEEE 802.3.

Ethernet MAC Protocol



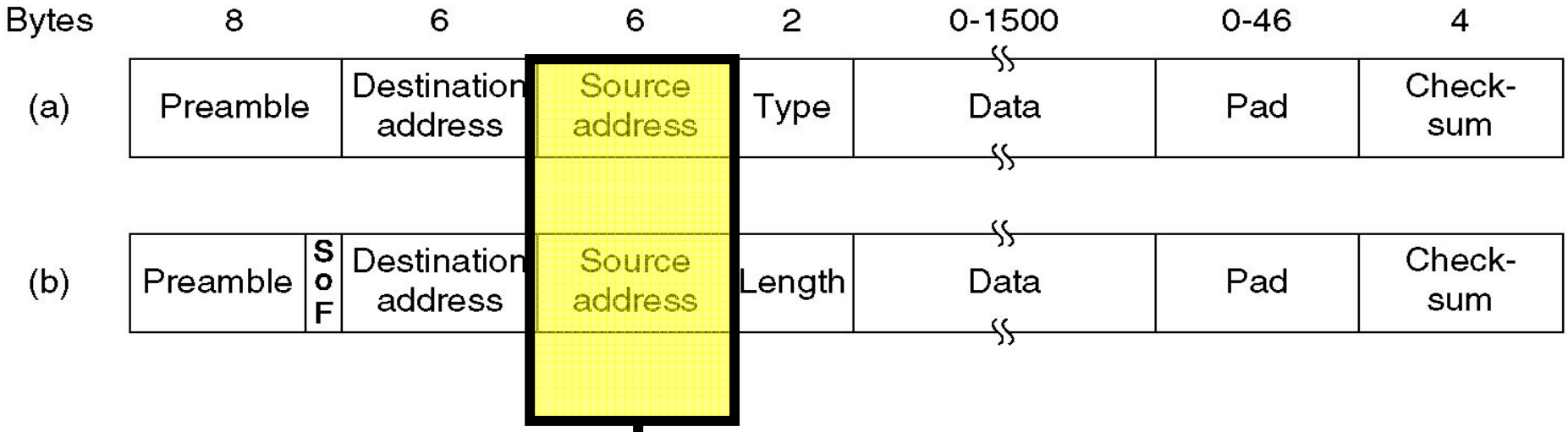
- Il Preambolo (10101010) serve a sincronizzare sia le trame Ethernet (a) che IEEE 802.3 (b). Ha una lunghezza minima (7 bytes) e termina con una violazione di alternanza (11 al posto di 10). Il preambolo Ethernet include un byte addizionale che equivale al campo IEEE 802.3 Start-of Frame (SOF=10101011). In parole povere cambiano le denominazioni ma i patterns sono identici nelle due trame.

Ethernet MAC Protocol



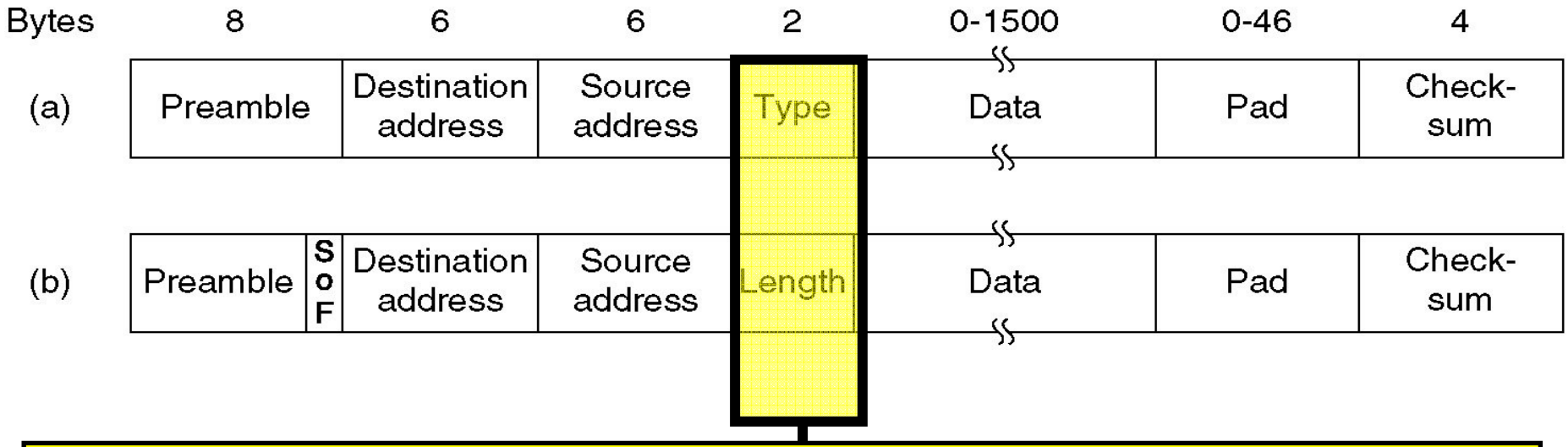
- Il campo Destination Address può essere unicast, multicast (gruppo) o broadcast (tutti i nodi).

Ethernet MAC Protocol



Il campo Source Address è sempre di tipo unicast.

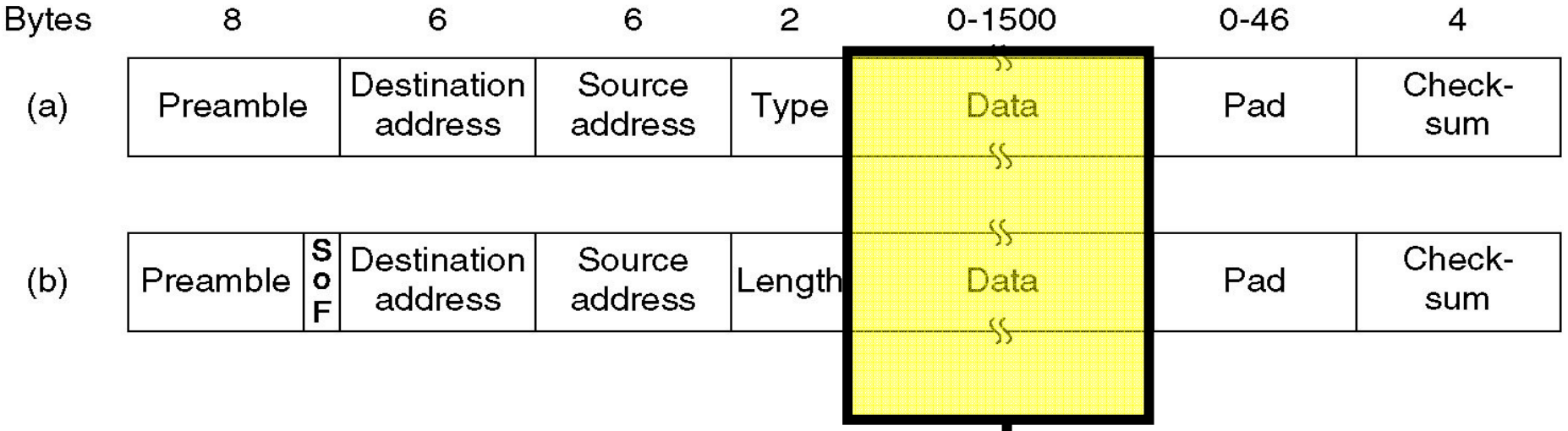
Ethernet MAC Protocol



Il campo Ethernet *Type* specifica il protocollo superiore trasportato (> 1536 0x600 esadecimale)

Il campo IEEE 802.3 *Length* specifica il numero di bytes di 'data' trasportati (che seguono questo campo).

Ethernet MAC Protocol

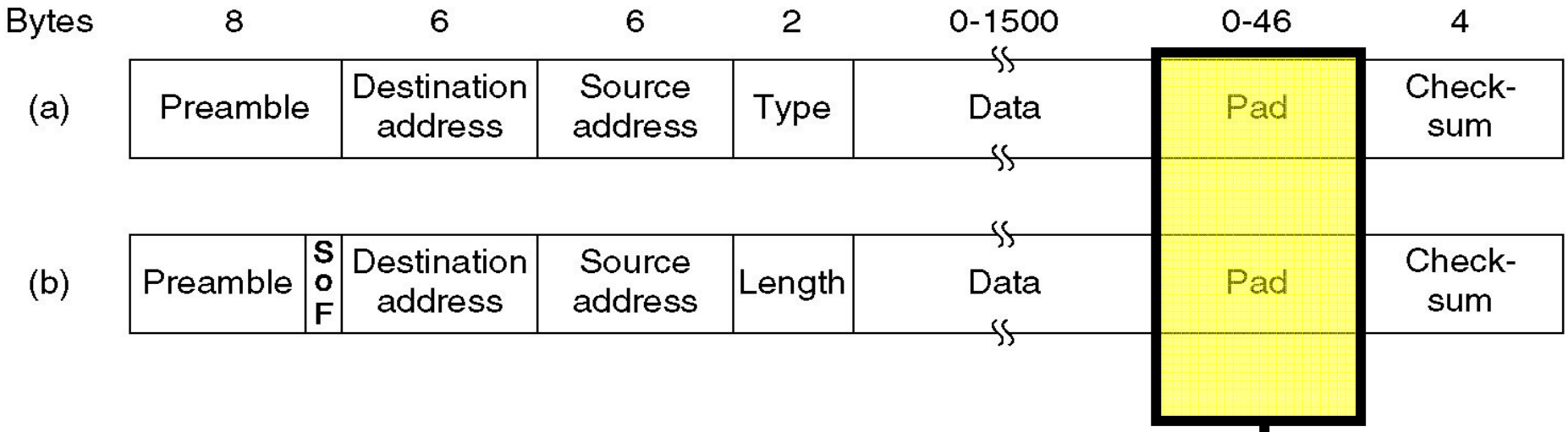


Il campo Data contiene il messaggio incapsulato.

Ethernet richiede che la frame sia non minore di 64 bytes e non maggiore di 1518 (escluso il preambolo).

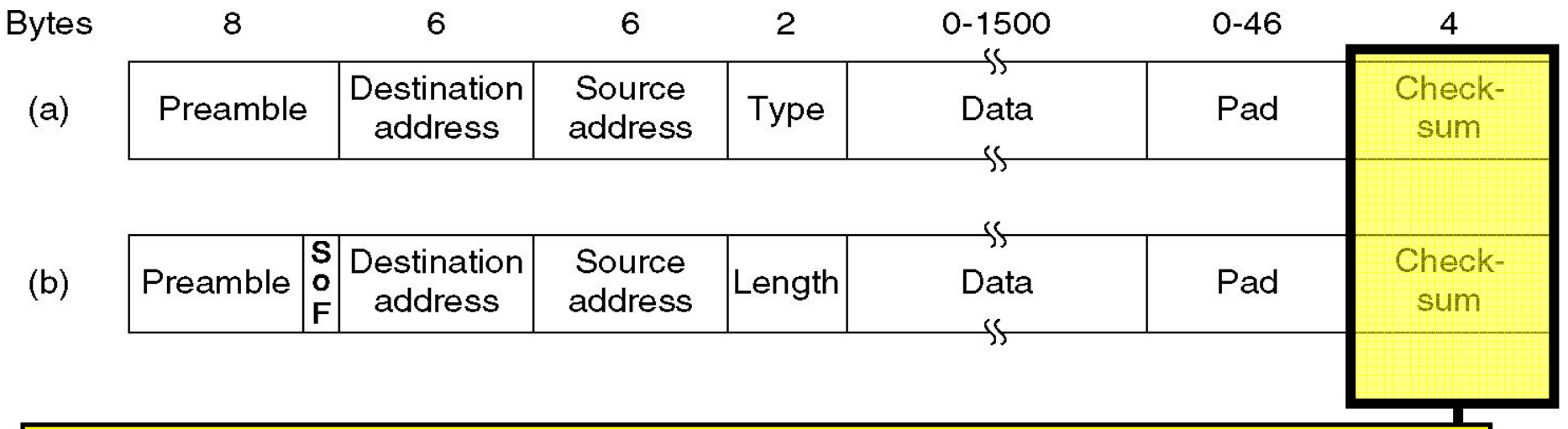
La MTU (maximum transmission unit) per Ethernet è 1500 ottetti, pertanto il campo DATA non deve eccedere questa limitazione.

Ethernet MAC Protocol



Il campo PAD può essere assente e serve a garantire la frame minima da 64 bytes.

Ethernet MAC Protocol



Il Campo Checksum o Frame Check Sequence (FCS) contiene un valore di 4-bytes di cyclic redundancy check (CRC).

- Il CRC è calcolato su tutta la frame Ethernet eccetto il preamble, SoF e FCS

$$x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}+x^{10}+x^{8}+x^{7}+x^{5}+x^{4}+x^{2}+x+1$$

Interframe spacing

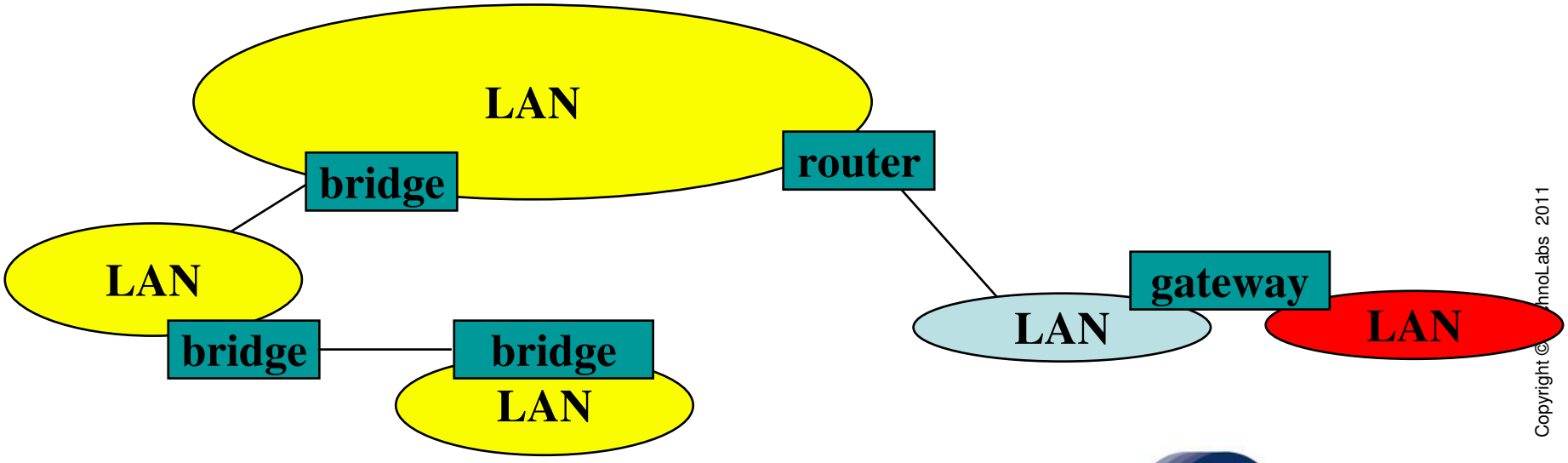
- Lo spazio minimo tra due frame che non-collidono è chiamato “interframe spacing”.
  - Dopo che un frame è stato inviato, tutte le stazioni su una 10-Mbps Ethernet devono aspettare un tempo minimo di 96 bit (9.6 microseconds) prima che qualunque stazione possa legalmente trasmettere il prossimo frame.
  - Questo intervallo è anche chiamato “spacing gap”.
  - Il gap serve per permettere a stazioni lente di processare il frame precedente e prepararsi al prossimo frame.
  - Se il livello MAC non riesce a spedire una frame dopo 16 tentativi, ci rinuncia e segnala un errore al network layer.
    - Questa occorrenza è abbastanza rara e si ha su reti estremamente cariche o con problemi fisici.

Gestione Errori

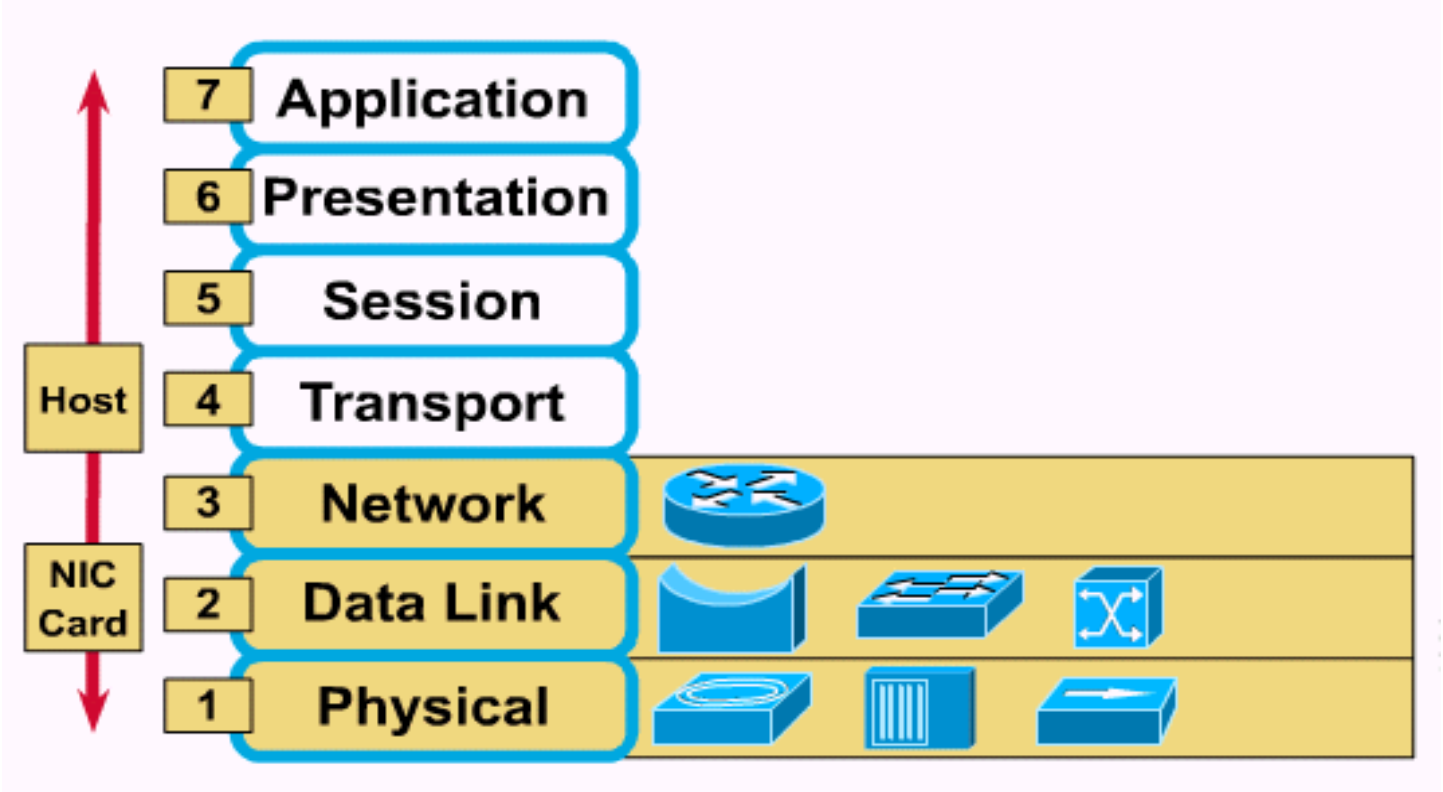
- La maggioranza delle collisioni occorre abbastanza presto nella frame, spesso prima dell' SFD.
  - Appena una collisione viene avvertita, la stazione trasmittente invia una sequenza da 32-bit detta di "jamming" che rafforza la collisione.
  - Il segnale di jam può essere composto da qualunque dato binario purchè non rispetti il campo FCS.
  - I messaggi parzialmente trasmessi sono chiamati "frammenti" o "runts".
  - Normali collisioni sono più corte di 64 ottetti e pertanto falliscono sia la lunghezza minima che il test FCS.

Apparati: Hub,bridge, router e gateway

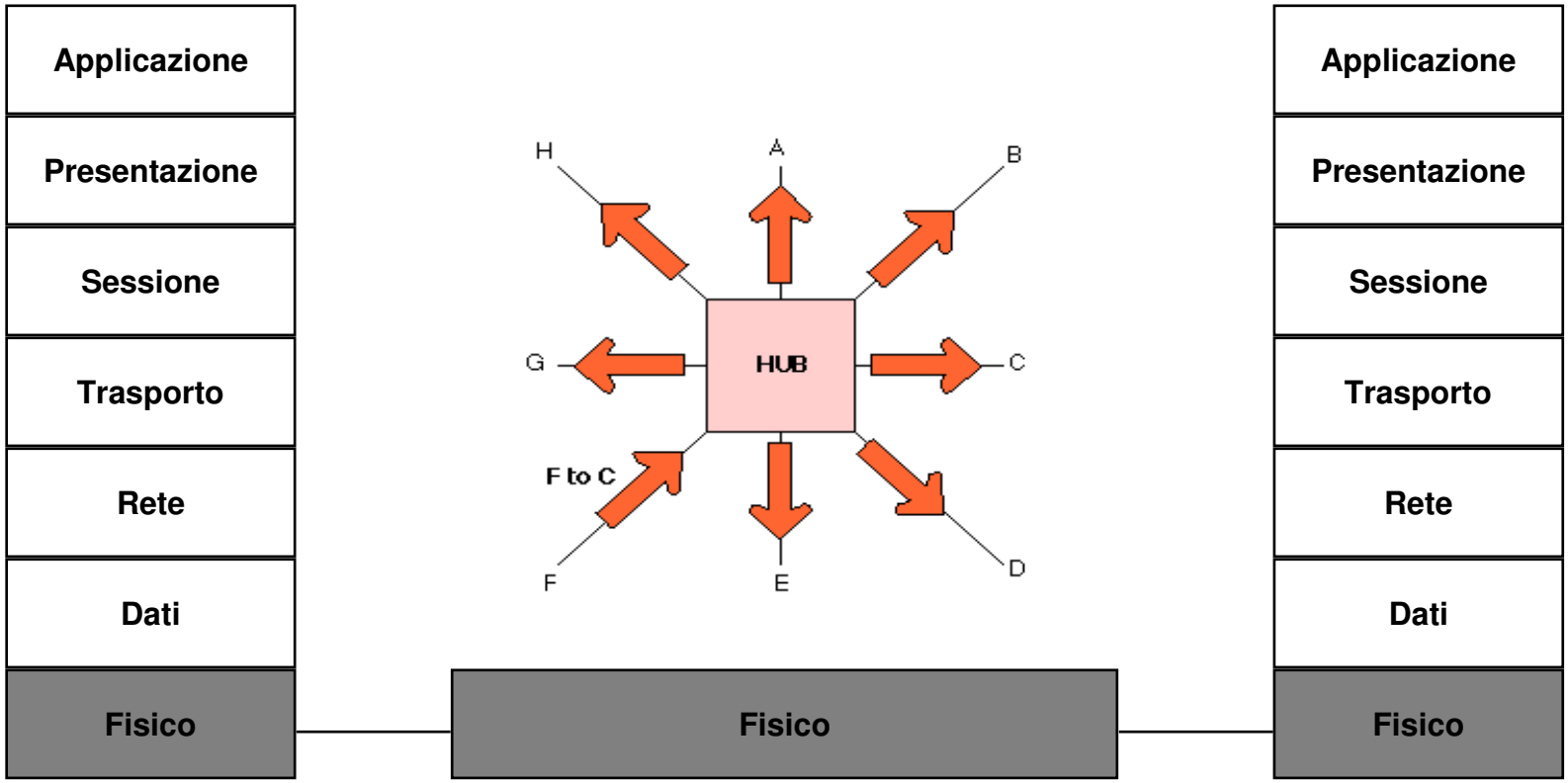
- Servono per interconnettere tra loro reti con diverse ...
  - tipologie (ad es. reti locali e geografiche)
  - tecnologie (ad es. Ethernet e token-ring)
  - architetture di rete (ad. es. SNA e TCP/IP)
  - e per aumentarne la dimensione (ad es. reti locali estese)



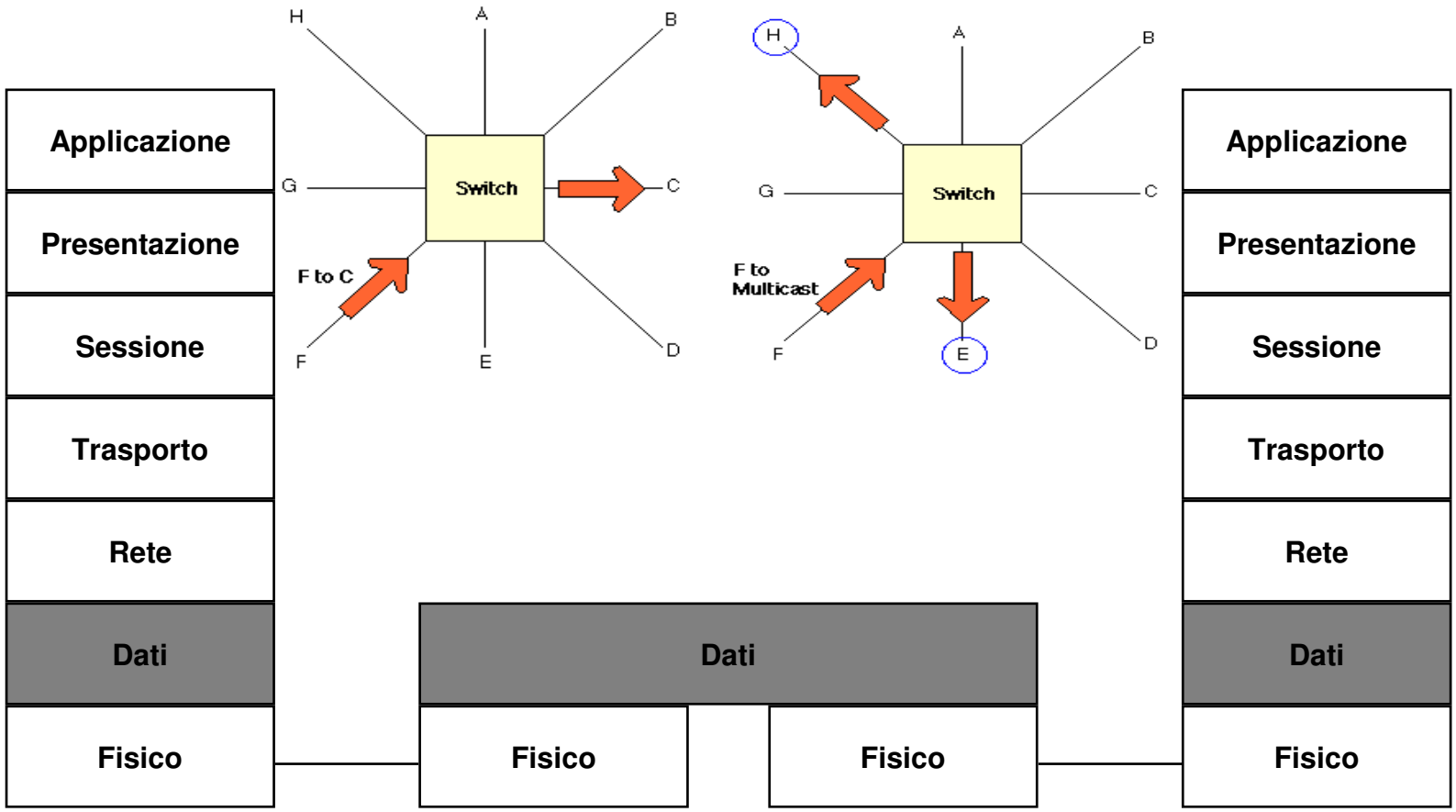
Devices Function at Layers



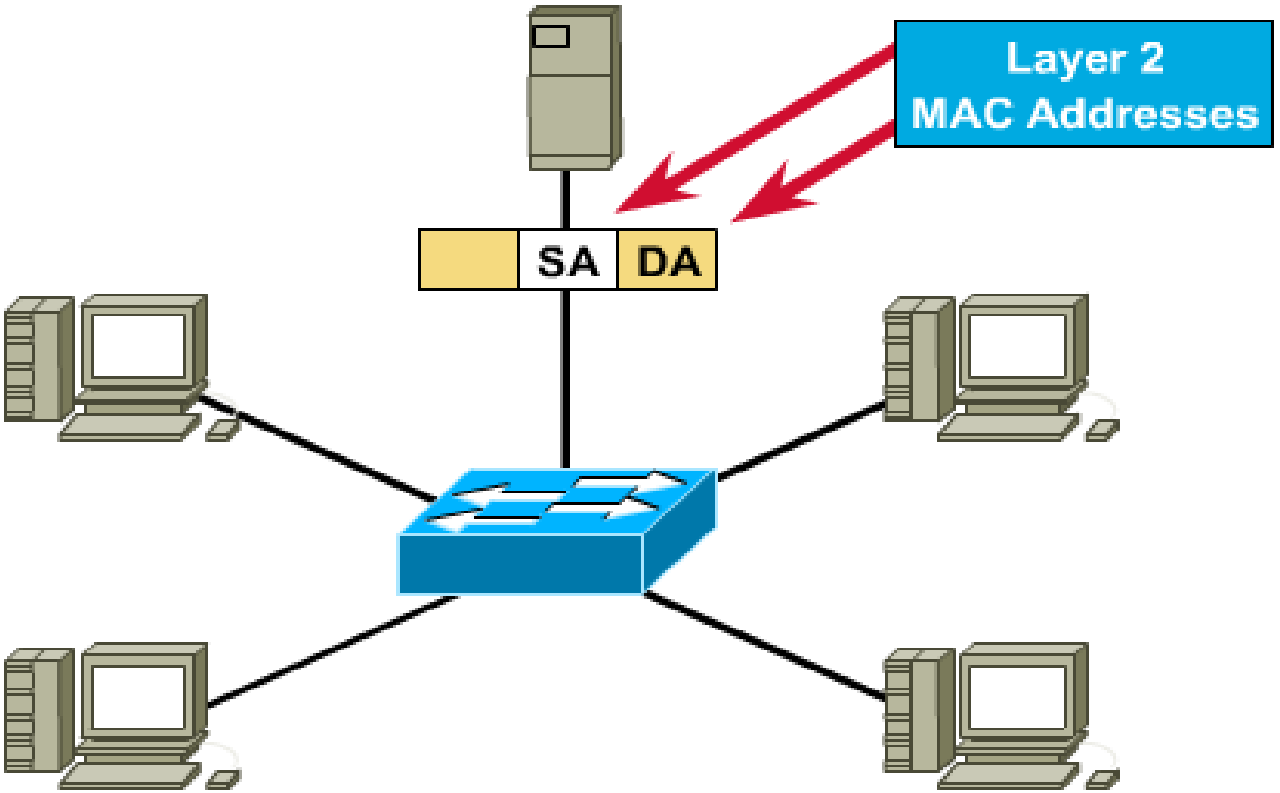
Repeater (E Hub)



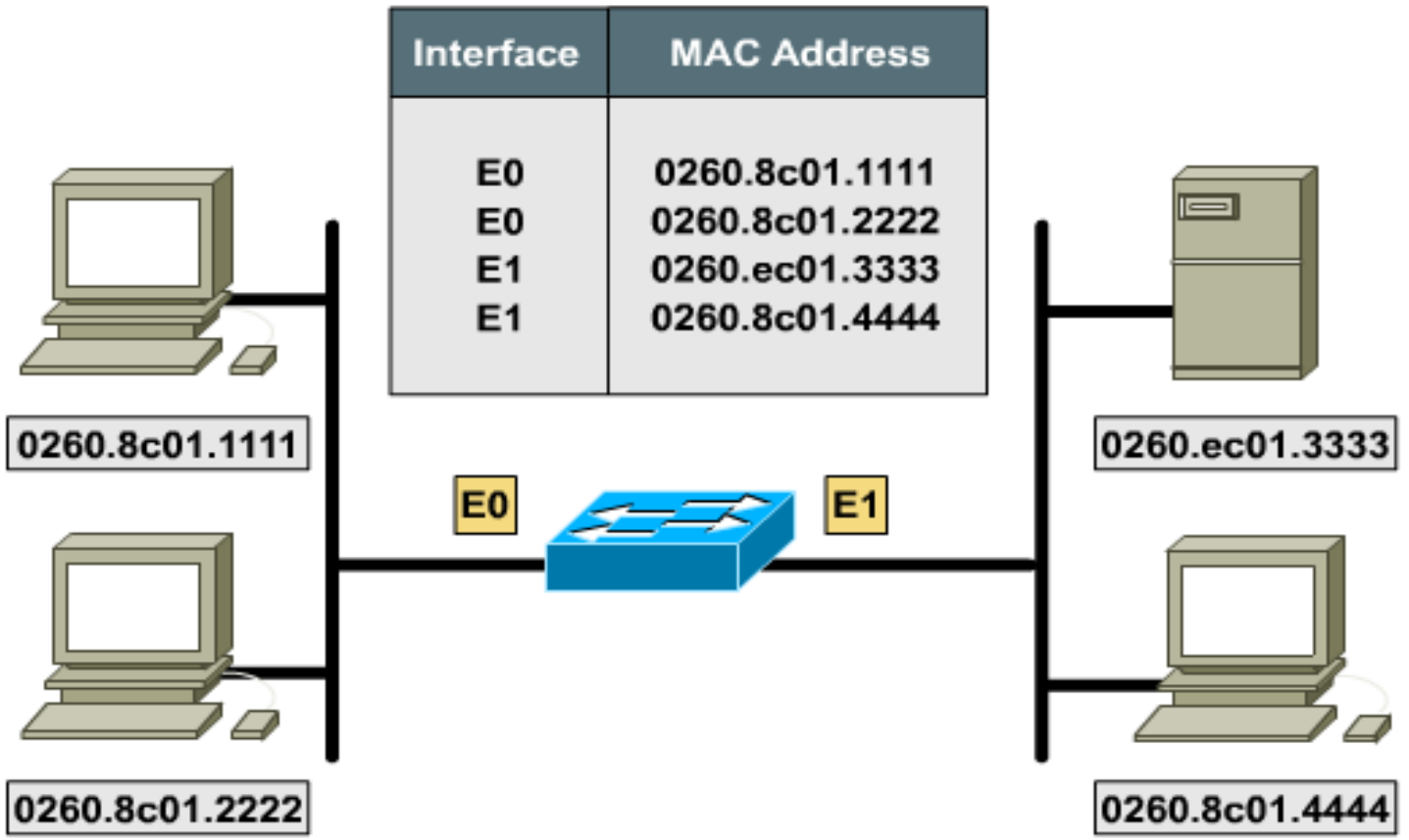
Bridge (Switch)



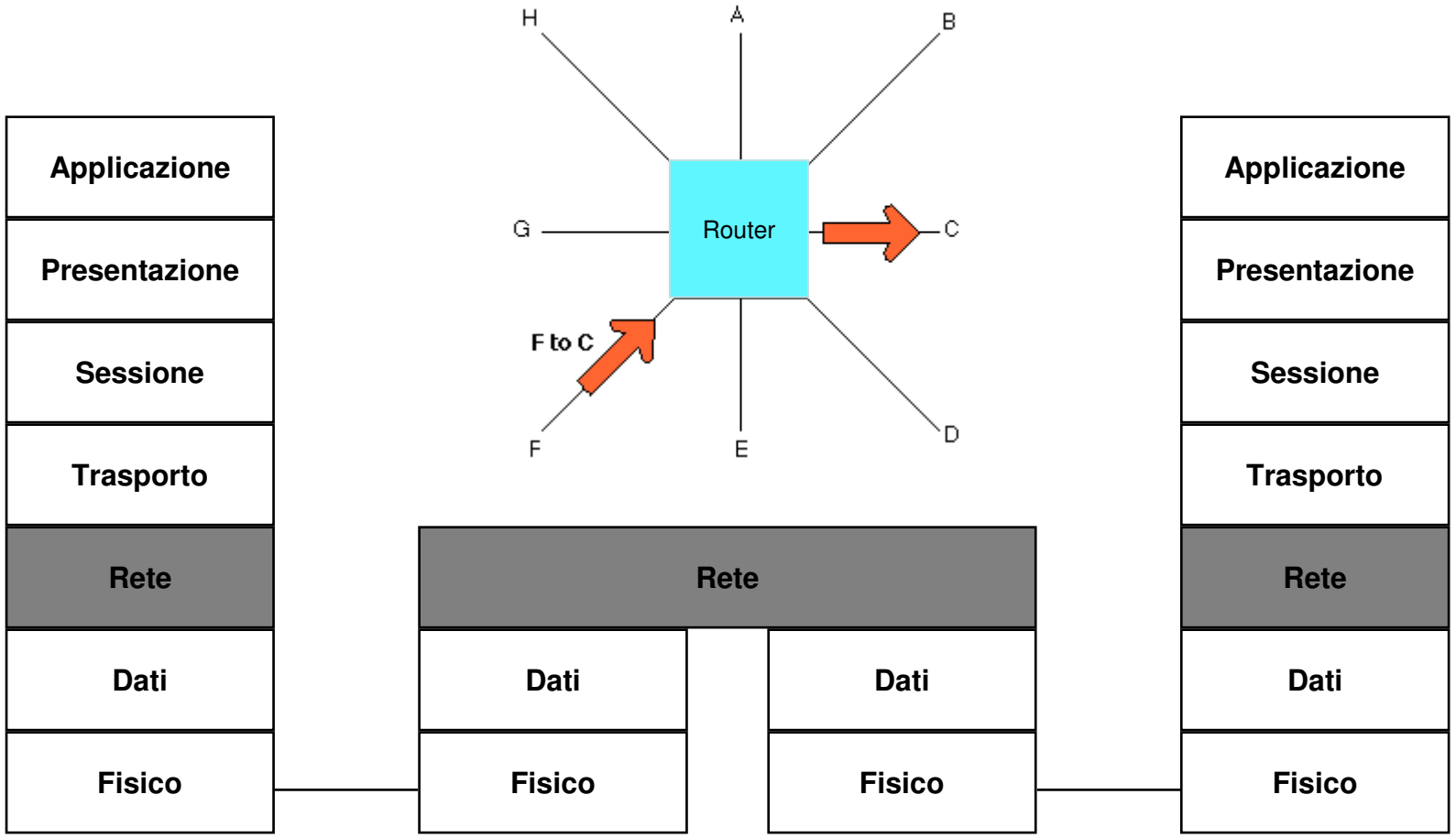
Switches



Switching Table



Router (e Laver-3 Switch)



TechnoLabs

Gateway

intecs the Brainware Company

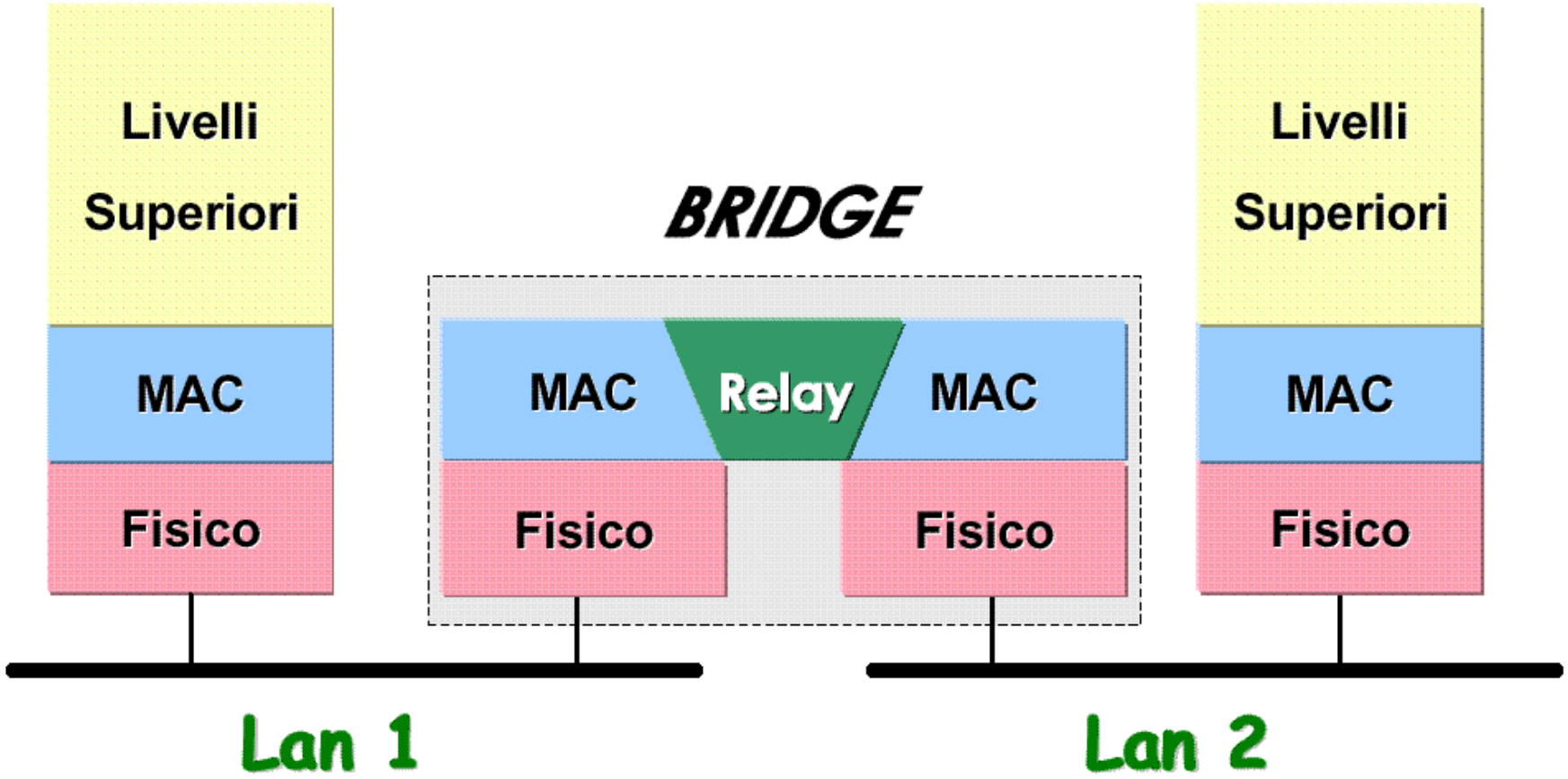
Data Flow

- Un device di Livello 1 inoltra una frame sempre
- Un device di Livello 2 inoltra una frame a meno che non gli venga specificato di non farlo
- Un device di Livello 3 blocca la frame a meno che non gli venga specificato di inoltrarla

intecs
the Brainware Company

L2-Switch

I Bridge



Proprietà di un bridge

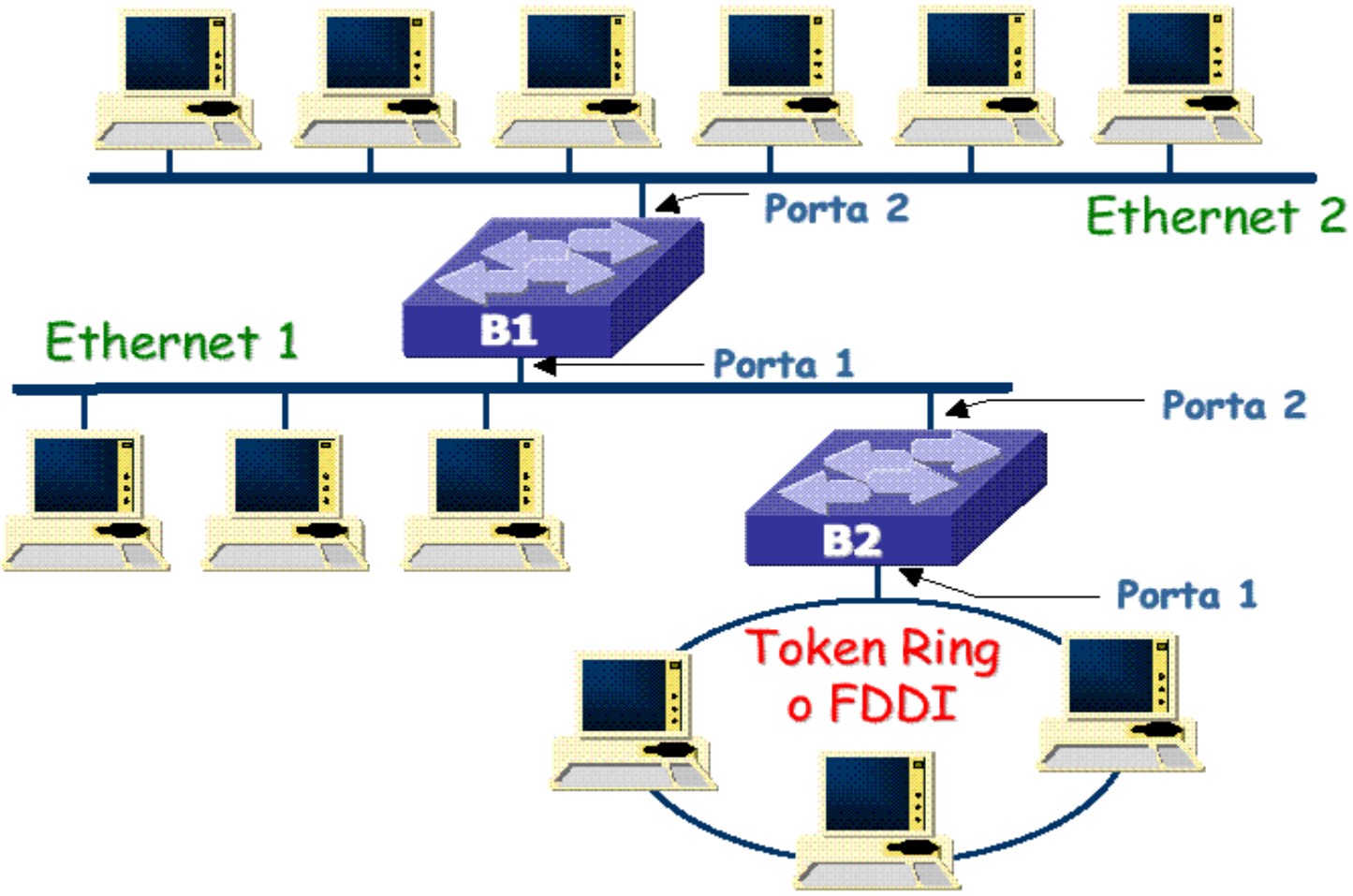
Instrada le trame di livello 2

- Utilizza lo Spanning Tree
  - per costruire un albero a partire da una maglia
- Utilizza il Filtering Database
  - per limitare la propagazione del traffico unicast
- Utilizza il flooding
  - Per il traffico Broadcast/Multicast/Unknown-unicast

Advanced features

- Multiple Spanning Tree
- VLANs
- User Priority
- IGMP snooping
- Trunking protocols
- Provider Bridge

Esempio di utilizzo di Bridge



IEEE 802.1D: Bridge Trasparenti

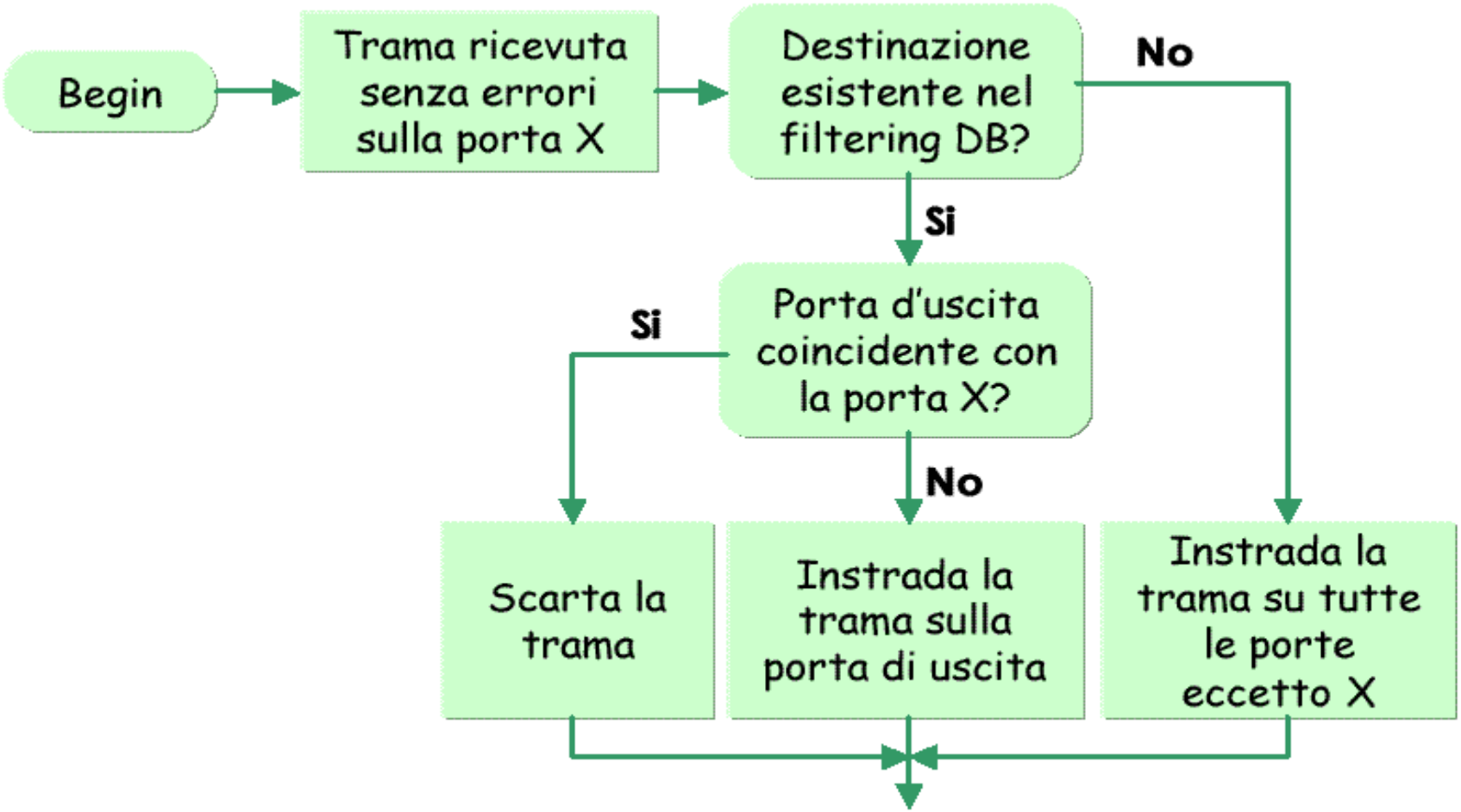
- Lo standard IEEE 802.1D definisce il funzionamento dei bridge cosìddetti Trasparenti (transparent spanning tree bridge)
  - sono derivati da Ethernet
  - hanno tabelle di instradamento locali
  - non necessitano di tabelle/modifiche sui nodi della LAN
- I transparent bridge svolgono tre funzioni base
  - forwarding di pacchetti
  - apprendimento della localizzazione di stazioni
  - risoluzione di possibili maglie
    - spanning tree protocol

Instradamento

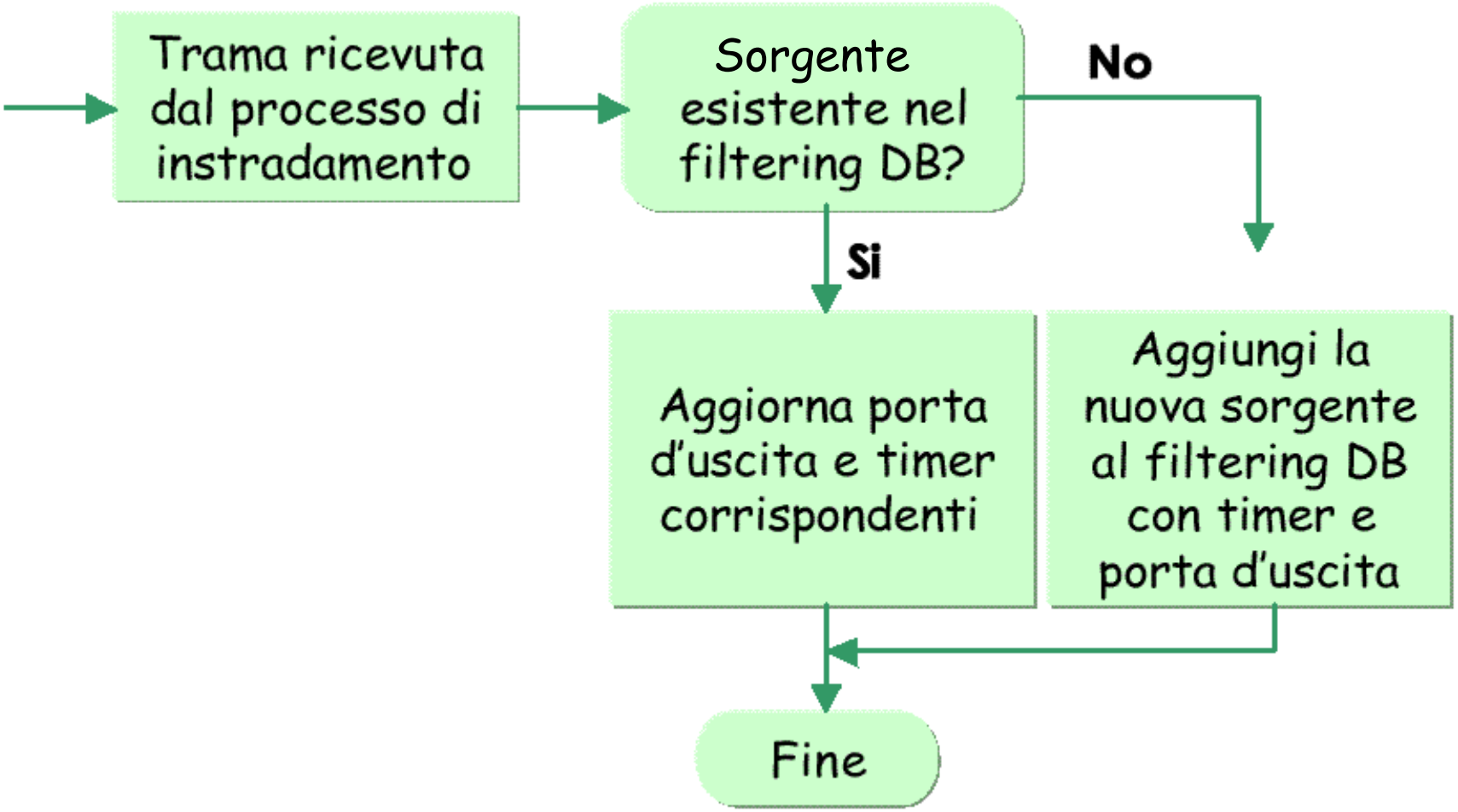
Le tabelle di instradamento sono calcolate tramite

- backward learning (algoritmo di routing isolato)
  - basato sull'osservazione degli indirizzi MAC
  - funziona solo su reti con topologia ad albero
  - le topologie magliate sono trasformate in topologie ad albero tramite il protocollo spanning tree
  - **Protocollo Spanning Tree**
    - opera periodicamente
    - decide quali porte porre in stato di ***forwarding*** e quali in stato di ***blocking***

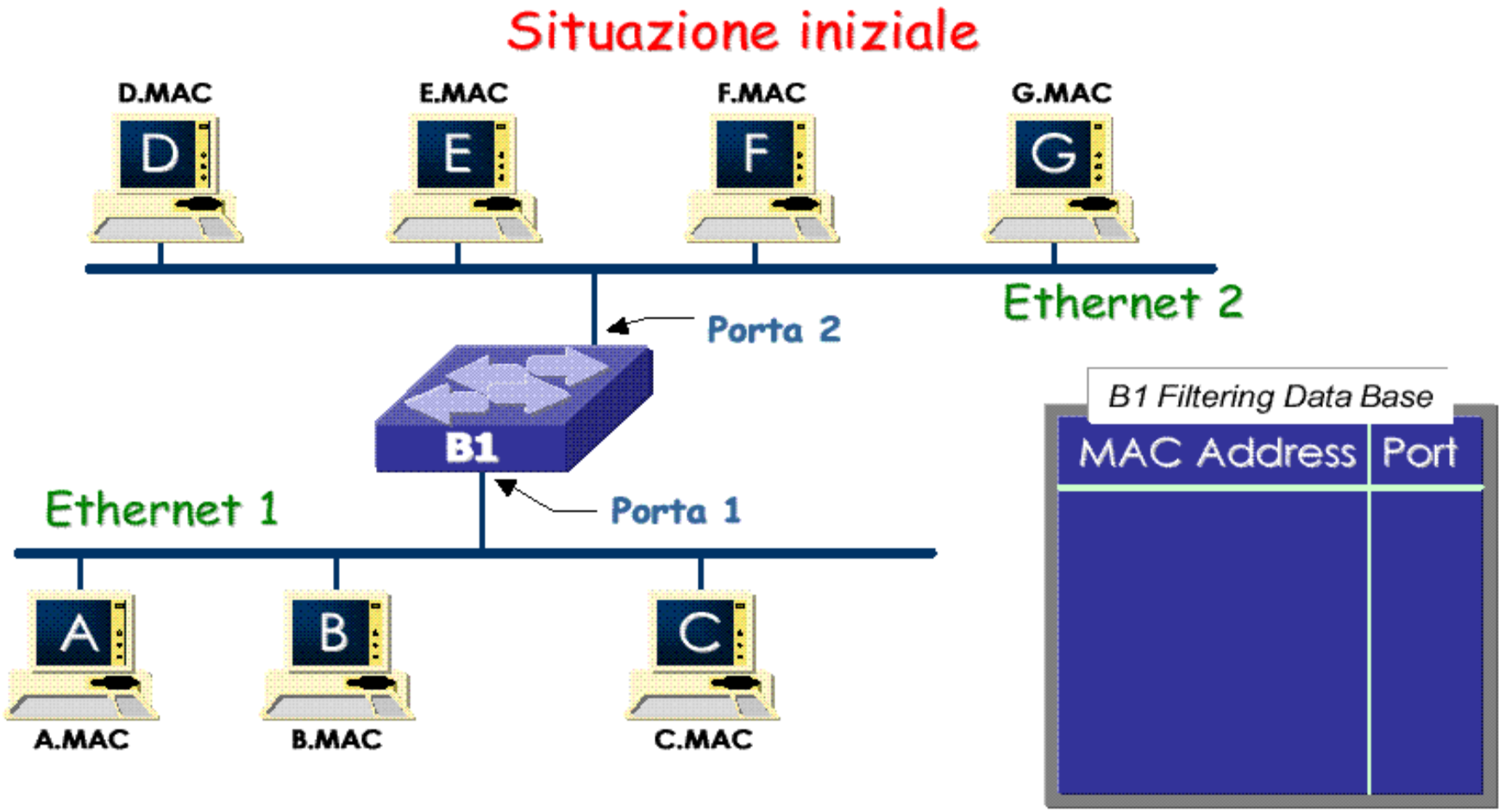
Bridge Forwarding



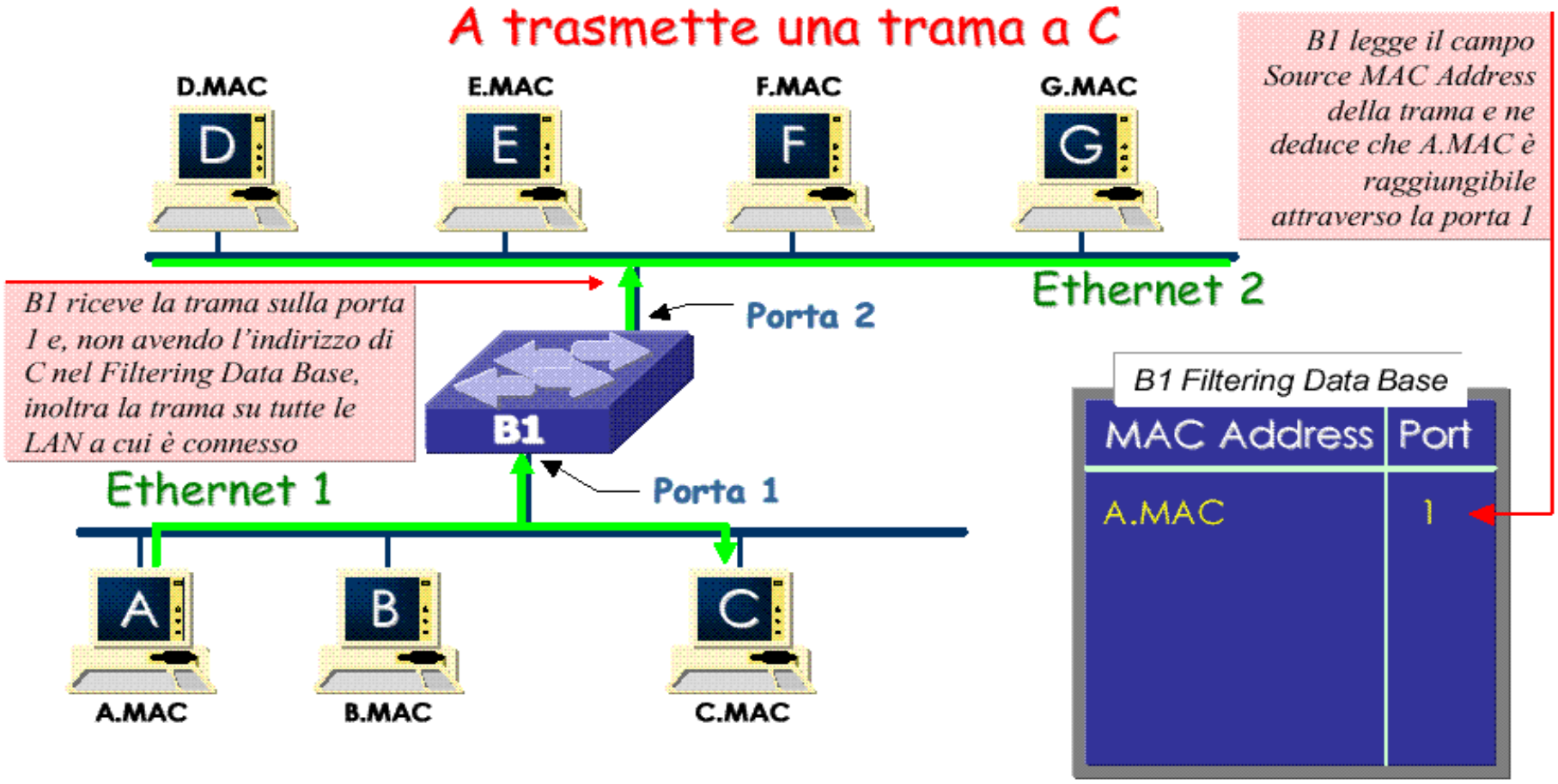
Bridge Learning



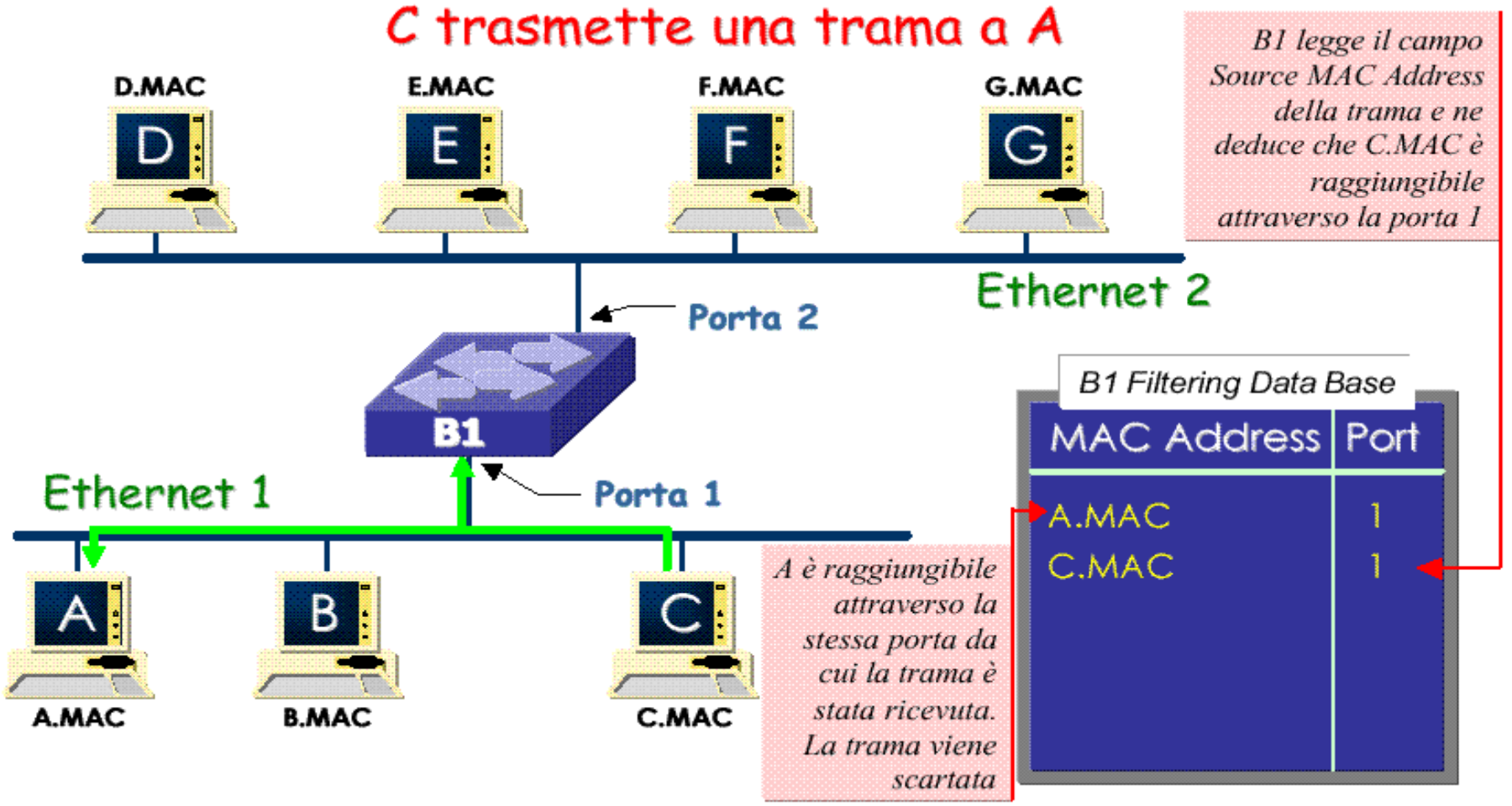
Esempio


Situazione iniziale
D.MAC
E.MAC
F.MAC
G.MAC
D
E
F
G
Ethernet 2
Porta 2
B1
Ethernet 1
Porta 1
A
B
C
A.MAC
B.MAC
C.MAC
B1 Filtering Data Base
MAC Address
Port

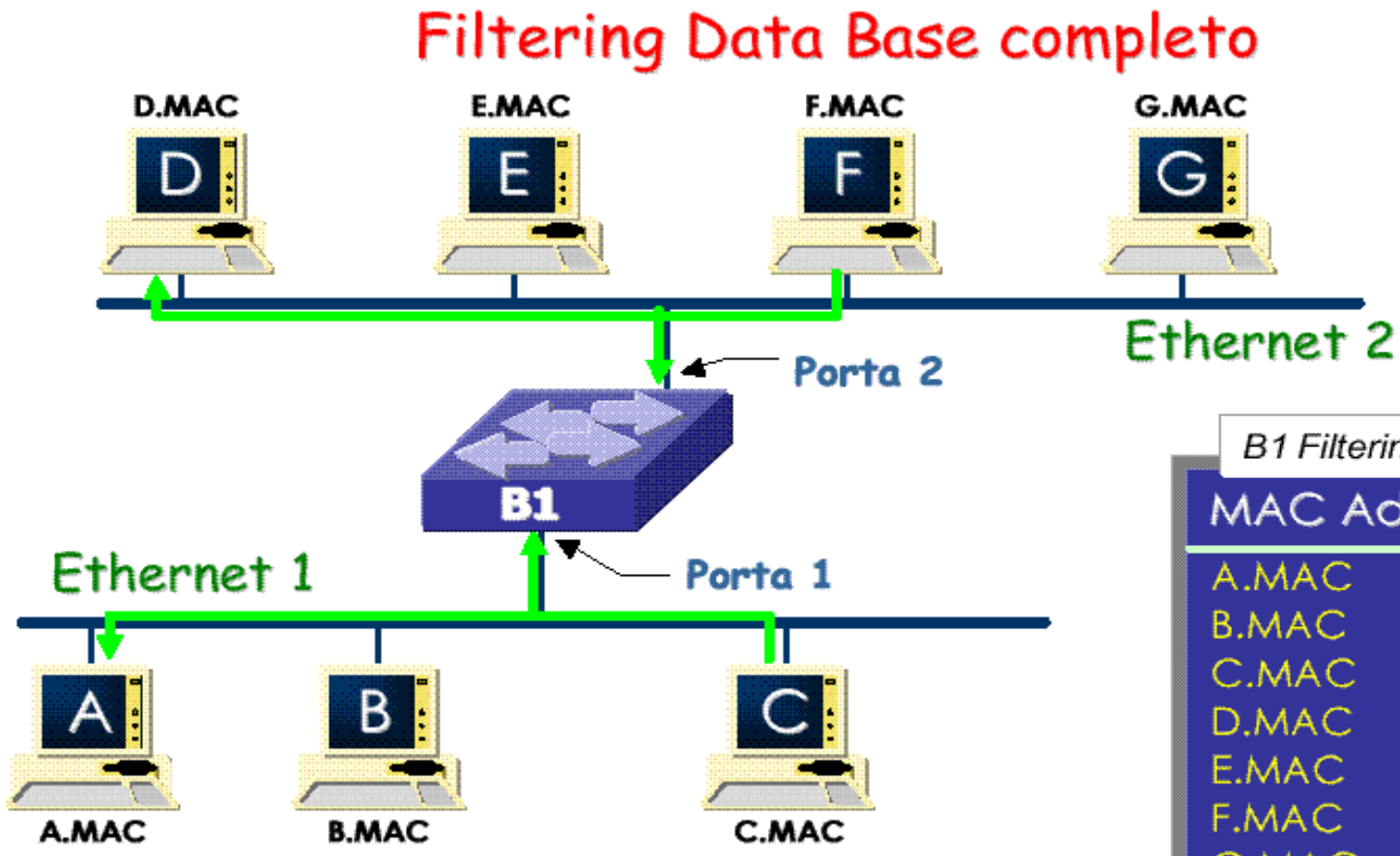
Esempio


A trasmette una trama a C
D.MAC
E.MAC
F.MAC
G.MAC
D
E
F
G
B1 legge il campo Source MAC Address della trama e ne deduce che A.MAC è raggiungibile attraverso la porta 1
Ethernet 2
Porta 2
B1 riceve la trama sulla porta 1 e, non avendo l'indirizzo di C nel Filtering Data Base, inoltra la trama su tutte le LAN a cui è connesso
B1
Ethernet 1
Porta 1
A
B
C
A.MAC
B.MAC
C.MAC
B1 Filtering Data Base
MAC Address
Port
A.MAC
1

Esempio


C trasmette una trama a A
D.MAC
E.MAC
F.MAC
G.MAC
D
E
F
G
Ethernet 2
Porta 2
B1
Ethernet 1
Porta 1
A
B
C
A.MAC
B.MAC
C.MAC
B1 legge il campo Source MAC Address della trama e ne deduce che C.MAC è raggiungibile attraverso la porta 1
B1 Filtering Data Base
MAC Address
Port
A.MAC
1
C.MAC
1
A è raggiungibile attraverso la stessa porta da cui la trama è stata ricevuta. La trama viene scartata

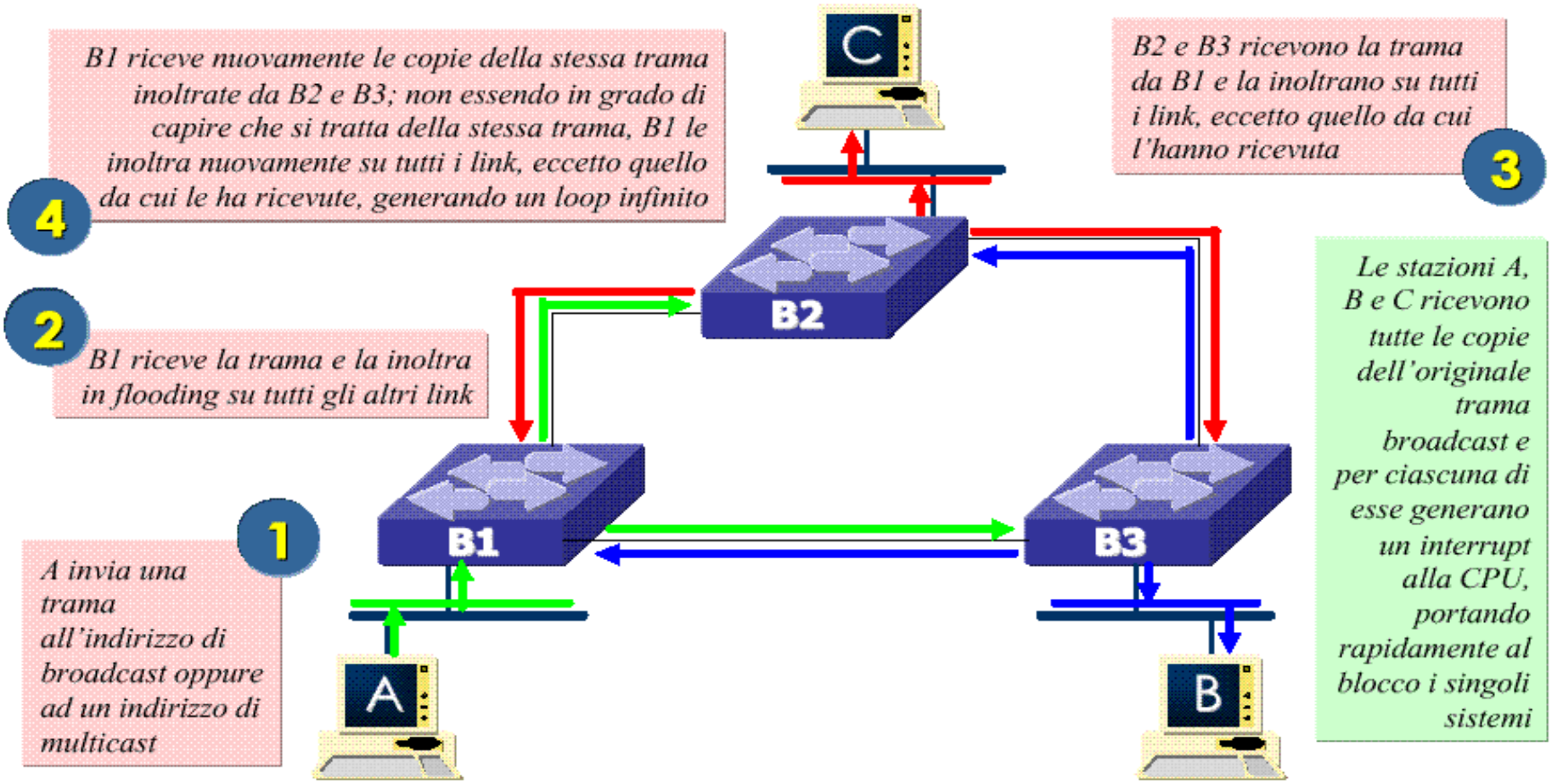
Esempio



B1 Filtering Data Base

MAC Address	Port
A.MAC	1
B.MAC	1
C.MAC	1
D.MAC	2
E.MAC	2
F.MAC	2
G.MAC	2

Problema delle maglie



Problema delle maglie

In presenza di maglie

- in pochi attimi si bloccano tutti i sistemi connessi alla rete
- fenomeno del broadcast storm

Creare un loop può essere più semplice di quanto non si creda

- è sufficiente sbagliare una permutazione in un patch panel
- e non basta accorgersene subito

Translating Bridge

- I bridge IEEE 802.1D devono essere translating
- Devono tradurre la busta di livello 2 ricevuta da una LAN nella busta di livello 2 da trasmettere sull'altra LAN
- Procedimento complicato quando si collegano LAN di tipo diverso (ad es. 802.3 con 802.5)
- Una delle cose più complesse è che si potrebbero dover trattare messaggi di lunghezza maggiore di quella supportata sulla rete di destinazione
  - La frammentazione dei messaggi è un compito tipico del livello 3!

Parametri operativi

- Aging time
  - I dati nel filtering database debbono invecchiare
    - Min= 10 s
    - Max= 1000000 s = 11.6 giorni
    - Rec= 300 s
- Bridge transit delay
  - I dati nel bridge debbono transitare entro:
    - Max= 4 s
    - Rec= 1 s
- FCS checking
  - I bridge debbono testare l'FCS e scartare il frame se non corretto
- Same Source and Destination Address
  - I bridge possono comportarsi come meglio gli aggrada

Bridge Address

- A rigore un bridge non ha bisogno di un indirizzo MAC, tuttavia lo standard richiede che un bridge abbia un MAC address per ogni porta più un MAC address che denoti il bridge nella sua interezza.
- Considerazioni aggiuntive hanno portato ad introdurre addizionali indirizzi MAC per supportare più istanze di spanning tree (1 MAC per istanza)

NOTA: Ci sono comunque in commercio bridge che hanno un solo MAC address

Indirizzi multicast riservati

INDIRIZZO	USO
01-80-C2-00-00-00	Spanning Tree Protocol
01-80-C2-00-00-01	Full Duplex PAUSE
01-80-C2-00-00-02	Link Aggregation
01-80-C2-00-00-03	Uso futuro
……	
……	
……	
01-80-C2-00-00-0F	Uso futuro

Il Bridge non deve mai fare Forward/Flood di frames destinate ai suddetti indirizzi.

Queste frames vanno processate dal bridge in questione oppure scartate.

Prestazioni di un Bridge

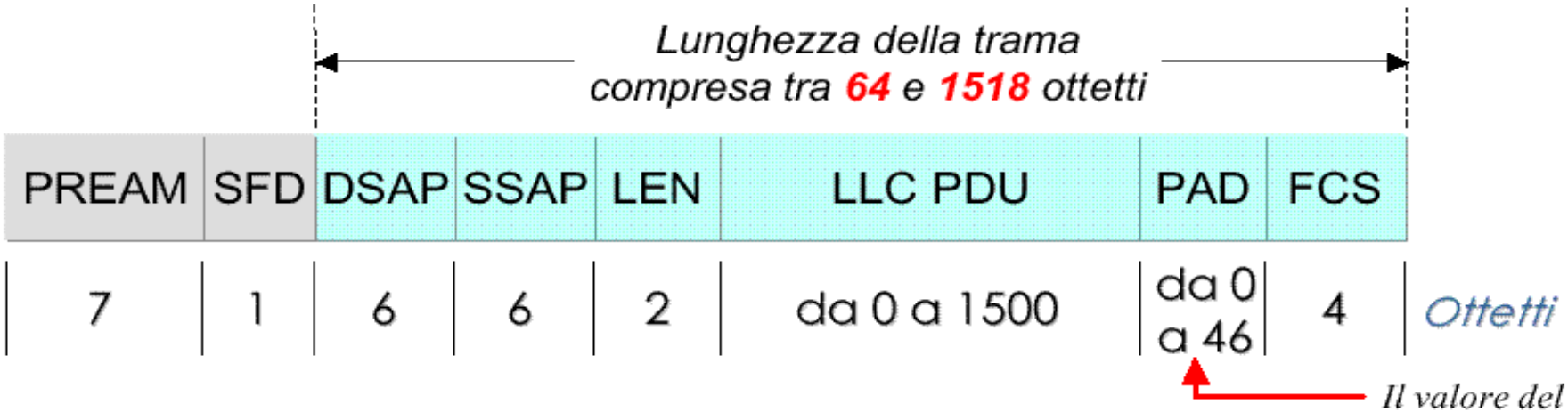
Un bridge è caratterizzato da due parametri

- **il numero di pacchetti/secondo che può ricevere e processare**
- **il numero di pacchetti/secondo che può inoltrare**

In generale il primo numero è maggiore del secondo

Si parla di bridge *wire-speed* quando questi due numeri sono uguali al massimo traffico che in teoria è possibile ricevere contemporaneamente da tutte le porte

MAX throughput Ethernet 10 Mb/s



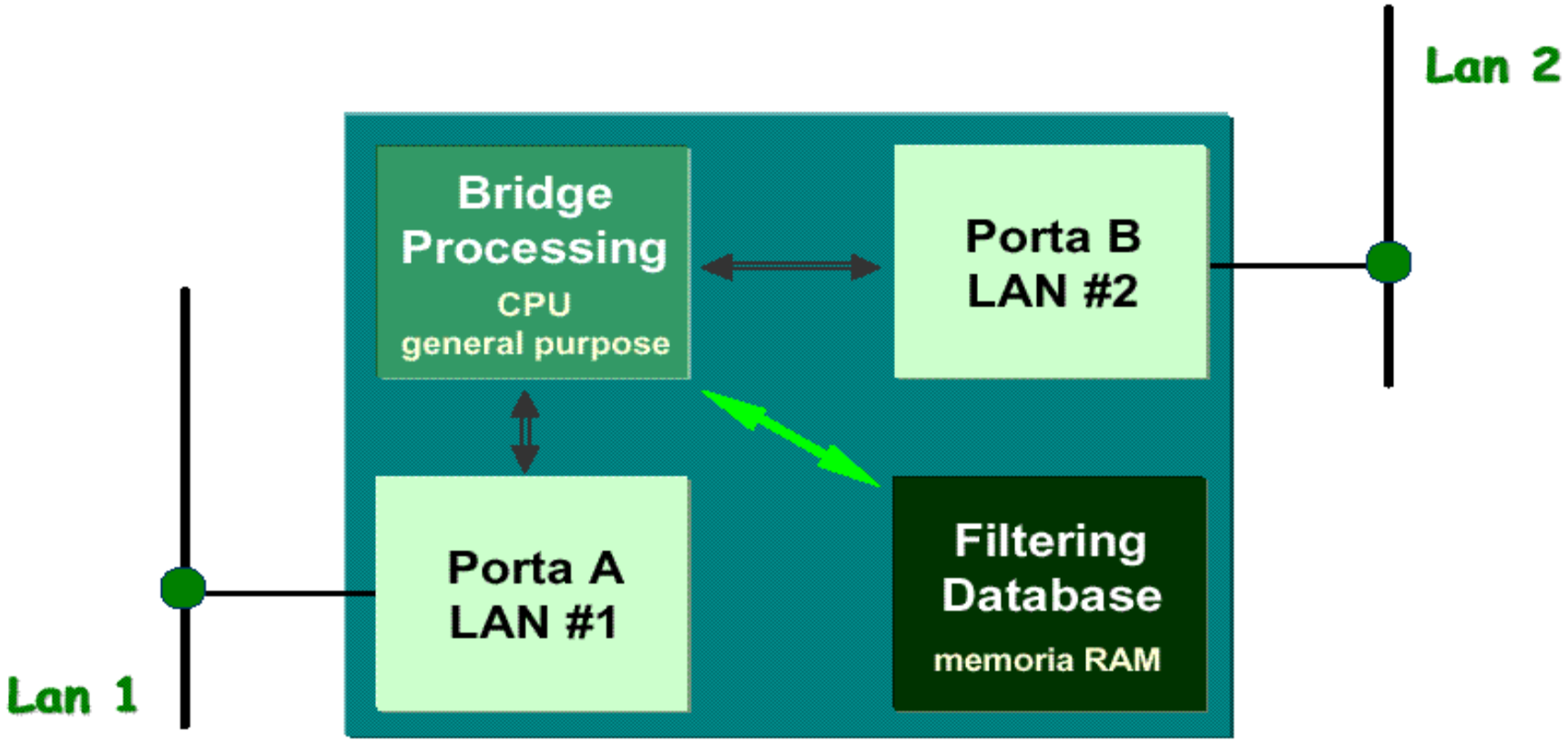
Caso peggiore: trame corte

- Dim. trama di livello MAC 64 ottetti = 512 bit
- Dim. totale a livello fisico 72 ottetti = 576 bit
- Inter-frame gap minimo 9,6ms = 96 bit
- Lunghezza totale di una trama 672 bit
- Velocità trasmissiva a livello fisico 10 Mb/s
- In un secondo si trasmettono 10 Mb/s ? 672 bit = 14880 pps

Prestazioni di un bridge 802.3

Dimensione pacchetto	Pacchetti al secondo	
	Carico 50%	Carico 100%
1518	403	812
1024	603	1206
512	1192	2385
256	2332	4664
128	4464	8928
64	8223	14880

Architettura di un Bridge

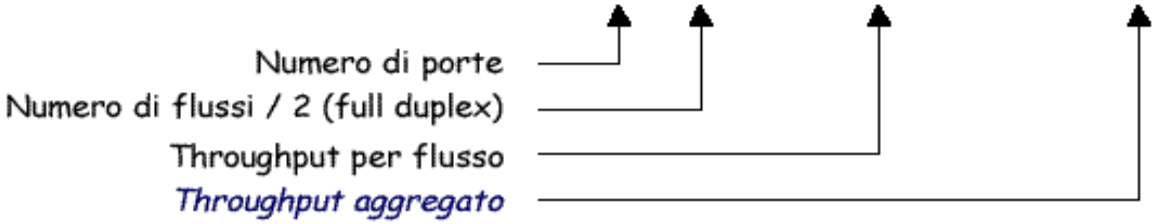


Prestazioni di uno switch

Tipologia di rete locale	Throughput in bps	Throughput in pps
Ethernet	10 Mbps	14 kpps
Fast Ethernet	100 Mbps	140 kpps
Gigabit Ethernet	1000 Mbps	1,4 Mpps

Le prestazioni di un dispositivo di interconnessione (di livello 2 o di livello 3) sono legate al numero di pacchetti al secondo che questo dispositivo deve inoltrare

Uno switch Fast Ethernet 8 porte full duplex wire speed deve essere in grado di gestire fino a 8 x 1 x 140kpps = 1,12 Mpps

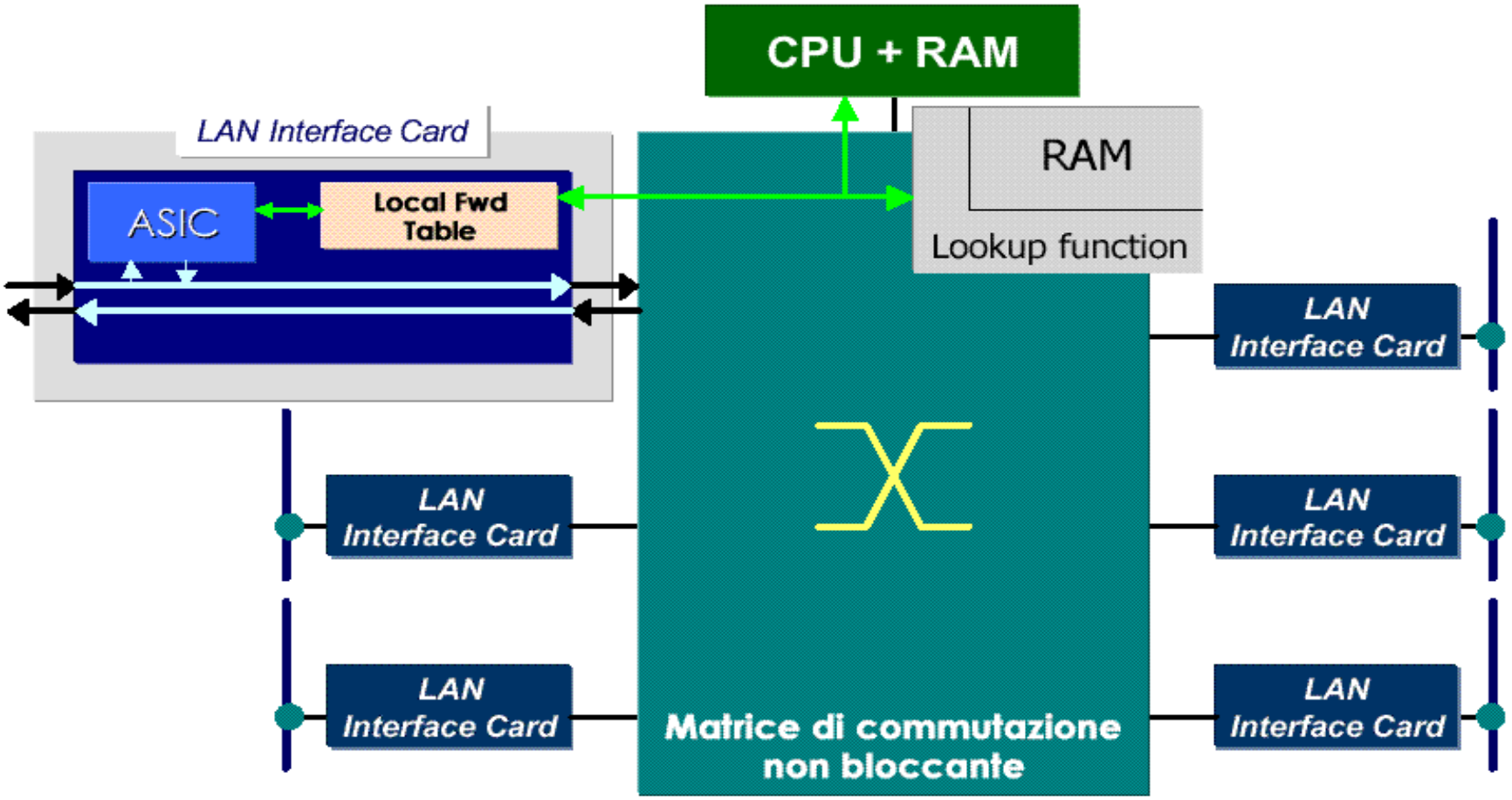


Prestazioni di uno switch

Con porte Fast Ethernet e Gigabit Ethernet diventa sempre più importante avere un throughput aggregato elevato

- Il problema non è la velocità in bps (che riguarda le singole schede di rete), ma il numero di trame (o pacchetti) da processare in un secondo
- Una CPU di tipo general purpose è in grado di processare fino a 500kpps
  - The rule of thumb is: 1 MIPS ==> 1kpps
- Si rendono necessarie architetture hardware specifiche, basate sull'utilizzo di ASIC (*Application Specific Integrated Circuit*) e di matrici di commutazione non bloccanti

Architettura di un moderno switch

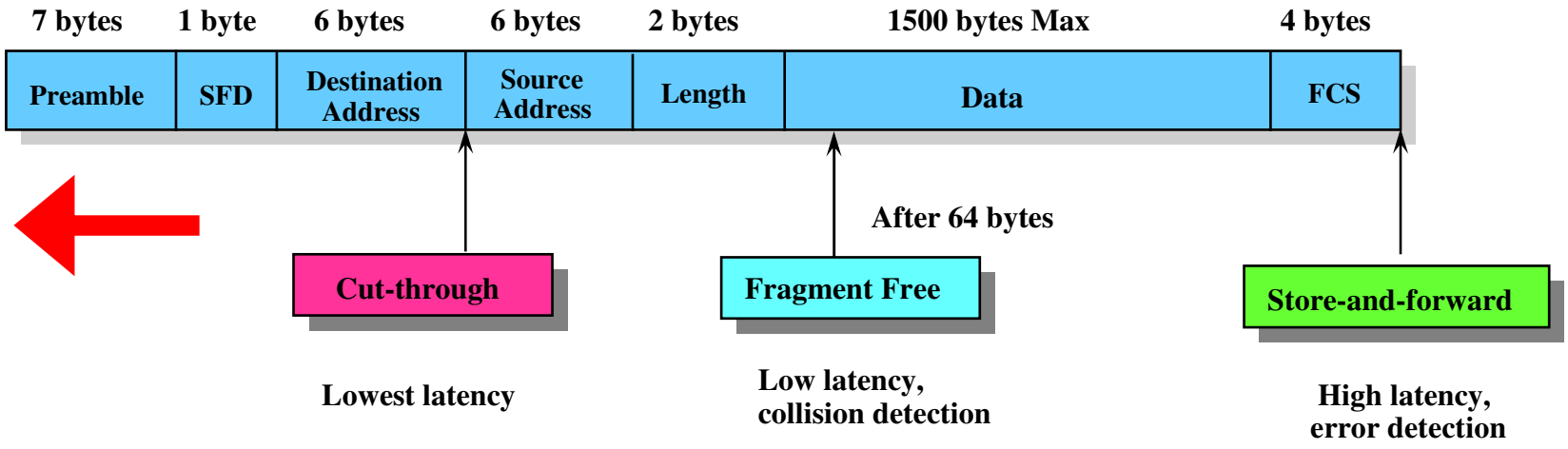


Tecniche di Switching

- **Store and forward**
  - Utilizzata dai bridge (prevista da IEEE 802.1D)
  - Si riceve la trama per intero e poi lo si ritrasmette
- **Cut through o on-the-fly switching**
  - Tecnica sviluppata da Kalpana
  - La decisione di inoltro viene presa durante il transito della trama nello switch
- **Fragment free**
  - Prima di iniziare a ritrasmettere la trama si aspetta comunque un tempo pari alla collision window (51.2 us nel caso di Ethernet 10Mb/s)

Ethernet Switch: schemi di Switching

- **Esistono 3 schemi di switching**
  - Store-and-forward
  - Cut-through
  - Fragment Free (Modified cut-through)



Store and Forward

- Opera come un bridge IEEE 802.1D multiporta ad alte prestazioni
- Può interconnettere MAC diversi
  - Ethernet, FDDI, ATM
- Può operare a velocità diverse
  - 10 Mb/s (802.3)
  - 100 Mb/s (802.3u)
- Non inoltra trame contenenti errori poiché controlla la FCS
- Non inoltra i frammenti di collisione

Cut through switching

Detto anche *On The Fly Switching*

- Ricevuto il campo MAC Destination Address lo switch decide se e dove ritrasmettere il pacchetto mentre la ricezione è ancora in corso
- Lascia passare eventuali frammenti di collisione poiché non aspetta che sia trascorsa la collision window
- Lascia passare eventuali pacchetti corrotti perché non può controllare la FCS

Architettura impiegata nel caso di reti ethernet

- I tempi di latenza sono molto bassi 10÷60 us

Limitazioni

Le tecniche cut through e fragment free possono essere utilizzate solo se

- su tutte le porte è presente lo stesso tipo di livello MAC
- tutte le porte hanno la stessa velocità trasmissiva
- la porta di destinazione è libera
- il pacchetto non è broadcast o multicast

In tutti gli altri casi occorre fare store and forward

Per i pacchetti corti le tre modalità sono equivalenti

Fault tolerance

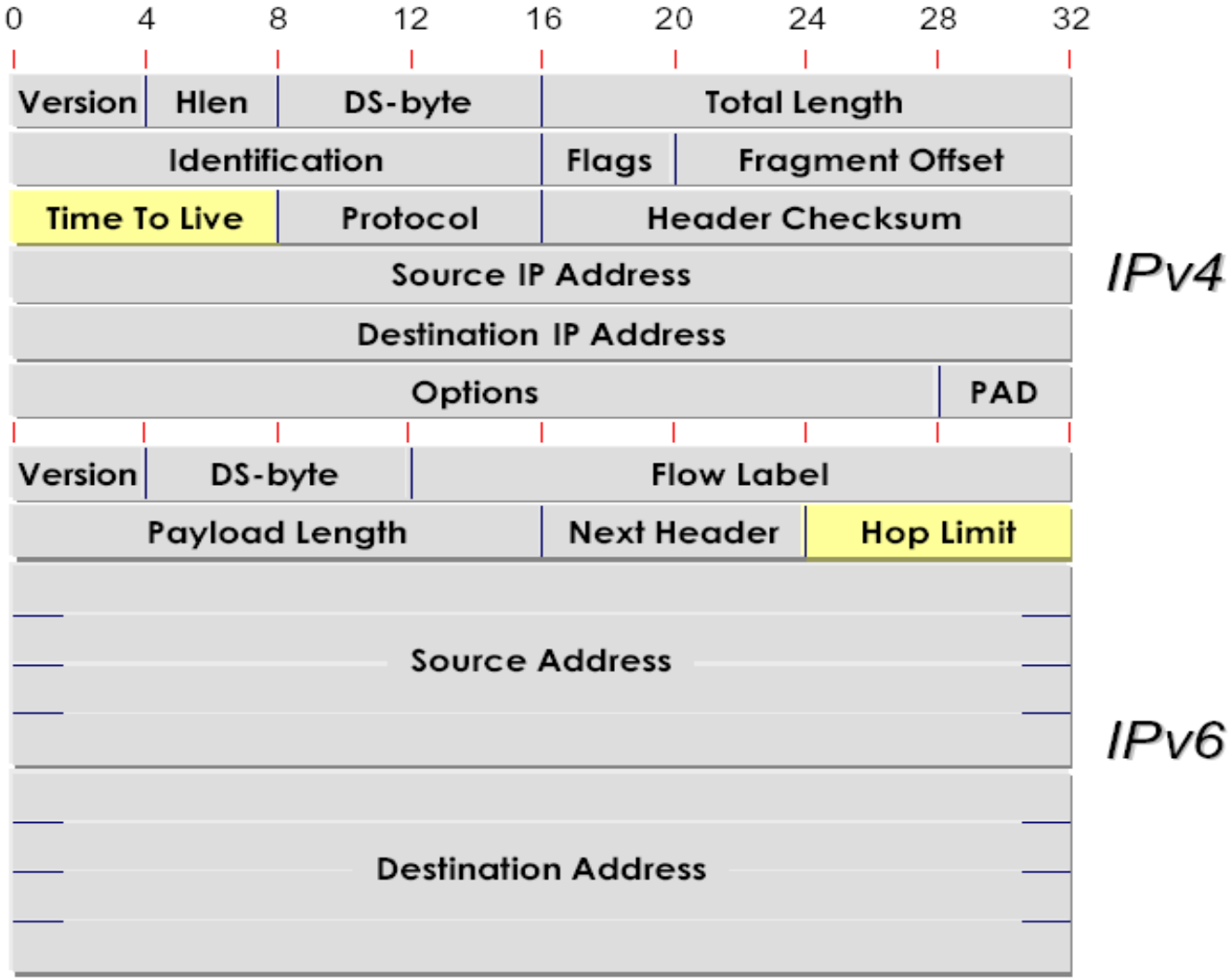
La topologia magliata è uno dei pochi modi per garantire davvero un buon livello di fault tolerance

- Ovunque in una rete sia presente una maglia, è anche presente un potenziale loop, ossia un anello chiuso lungo il quale potrebbero continuare a circolare pacchetti

A livello 3 i loop non rappresentano un problema grave

- In quasi tutti i protocolli di livello 3 è previsto un contatore che permette di scartare i pacchetti dopo un certo tempo, evitando che questi restino infinitamente all'interno del loop

TTL in IPv4 e HL in IPv6



Loop di livello 2

- A livello 2 non esistono i contatori presenti a livello 3
- È perciò necessario utilizzare un protocollo che eviti la formazione di loop
- Il protocollo utilizzato è lo Spanning Tree Protocol (STP):
  - Definito in IEEE 802.1D
  - Lo STP trasforma dinamicamente (periodicamente) la maglia in un albero

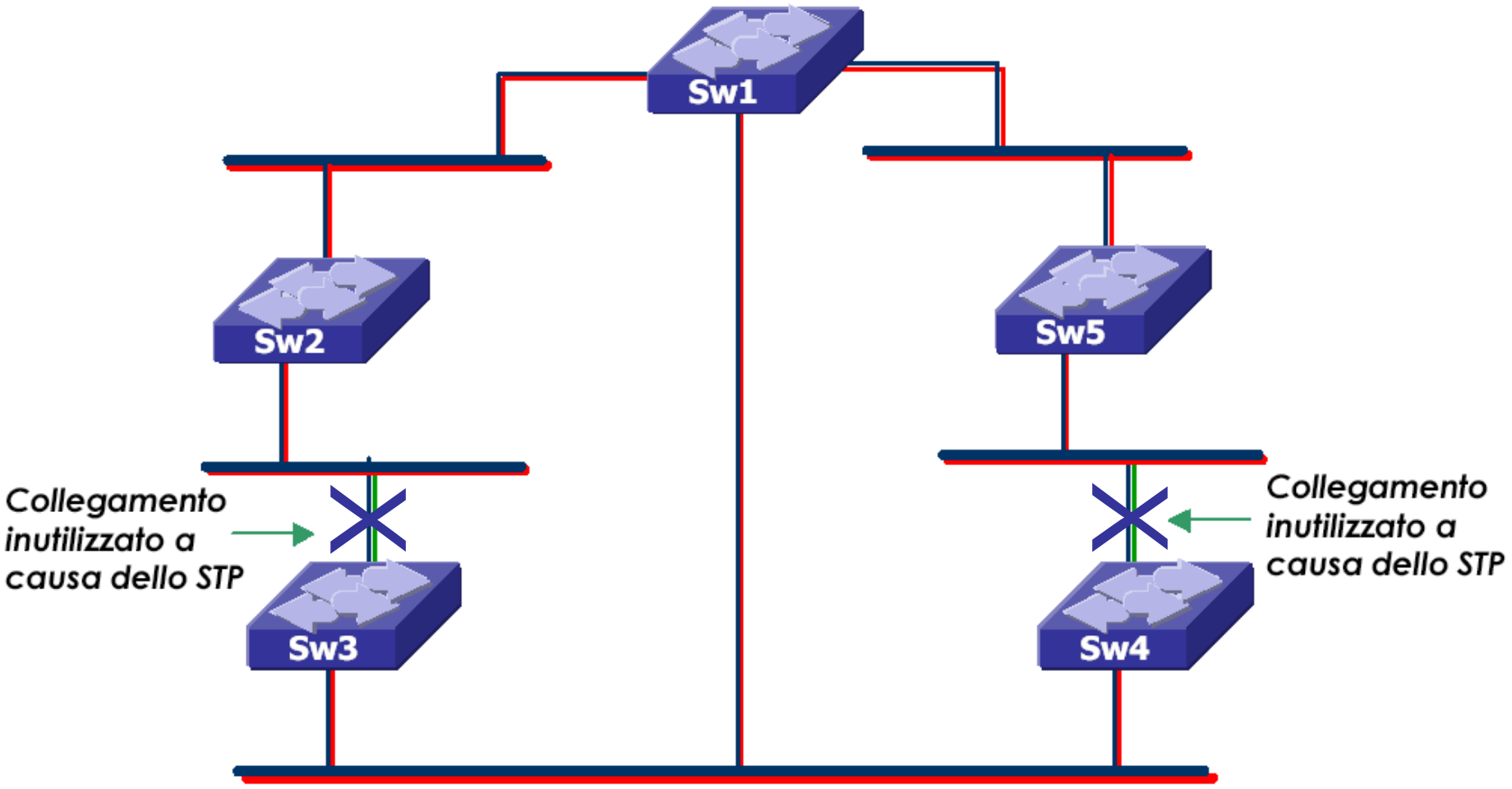
Spanning Tree

Lo STP non è un protocollo di routing

- Non serve per determinare i percorsi di instradamento,
- ma soltanto per ricavare, a partire da una topologia fisica magliata, un topologia logica ad albero (*lo spanning tree*) che rappresenta tutti i possibili cammini di instradamento nella rete

Tutti i bridge della rete devono partecipare allo STP in conformità a quanto specificato in IEEE 802.1D

Spanning Tree



Spanning Tree

Lo STP deve *convergere*, indipendentemente dalla dimensione della rete locale e la topologia attiva deve essere stabile e predicibile

- Fissato l'insieme dei parametri di configurazione e a parità di link e apparati correttamente funzionanti, la topologia attiva calcolata dall'algoritmo di spanning tree deve essere sempre la stessa

Lo STP deve essere *plug & play*

- Una volta configurato, deve funzionare correttamente senza richiedere l'intervento del network manager

Lo standard IEEE 802.1D contiene in appendice il codice C di funzionamento del protocollo stesso

- Questo ha semplificato i problemi di interoperabilità tra i costruttori

Spanning Tree

- L'algoritmo di spanning tree non blocca i collegamenti, ma solo le porte
- Una porta "bloccata" dallo STP lascia passare i messaggi del protocollo, ma non le trame contenenti i dati
- L'algoritmo opera nei seguenti passi
  - Elezione del Root Bridge
  - Selezione della Root Port
  - Selezione della Designated Bridge Port

TechnoLabs

Performance: Reliable, Fast Packet Delivery

- Wirespeed Performance
- How many packets per second ?
  - 14,880 pps on 10Mbps Ethernet
  - 148,800 pps on 100Mbps Ethernet
  - 1,488,000 pps on 1000Mbps Ethernet
- Non-Blocking Architecture
  - Each port can transmit data at wirespeed to another port
  - Multiple ports can transmit data at wirespeed to other multiple ports
- Forwarding Architecture: Store and Forward
- Broadcast Storm Control
  - Limits Network broadcasts on a per port basis
- Full/Half duplex per 10/100 port
- Full/Half duplex per Gigabit port

intecs the Brainware Company

TechnoLabs

Management: Simple Administration

- In-Band
  - Web Interface
  - Telnet
    - Concurrent Telnet sessions supported
- Out-of-Band
  - Local sessions with RS-232 DB9 console port
- SNMP

TechnoLabs

Broadcast/Multicast Storm Control

- A switch implements a global Broadcast/Multicast control algorithm which prevents misbehaving sources from using too many resources.
- When the Multicast/Broadcast traffic exceeds a certain threshold all broadcast and non-critical Multicast traffic is dropped:
  - For every ingress port the switch counts the broadcast frames in the ingress buffer to be processed. If there are more than 8 broadcast frames in a Fast Ethernet port or more than 16 broadcast frames in a GBE port, any further broadcast frame is dropped.

Spanning Tree Protocol

- A switch might support:
  - the regular spanning tree (IEEE802.1D),
  - the Multiple STP spanning tree (IEEE802.1s) and
  - the Rapid spanning tree specifications (IEEE802.1w).
- Because MSTP is a superset of RSTP and STP, each port can be forced by the operator to work in STP or RSTP or none modes.
- The software can handle n+1 (n=MSTI, 1=CST) Spanning trees instances at a time.

TechnoLabs

IEEE 802.3ad Link Aggregation

- Standards based technology
- Redundancy
  - Protect mission critical links
- Increased bandwidth
  - Economical way to create a high bandwidth port
- An aggregation of 2, 4 or 8 ports for FE ports and an aggregation of 2 or 4 ports for GbE ports can be specified.
  - Check CPU registers to prevent out-of-order delivery of packets when an aggregated link becomes live.
- Load balancing over all ports in group
- Continues to forward traffic if one link in group fails

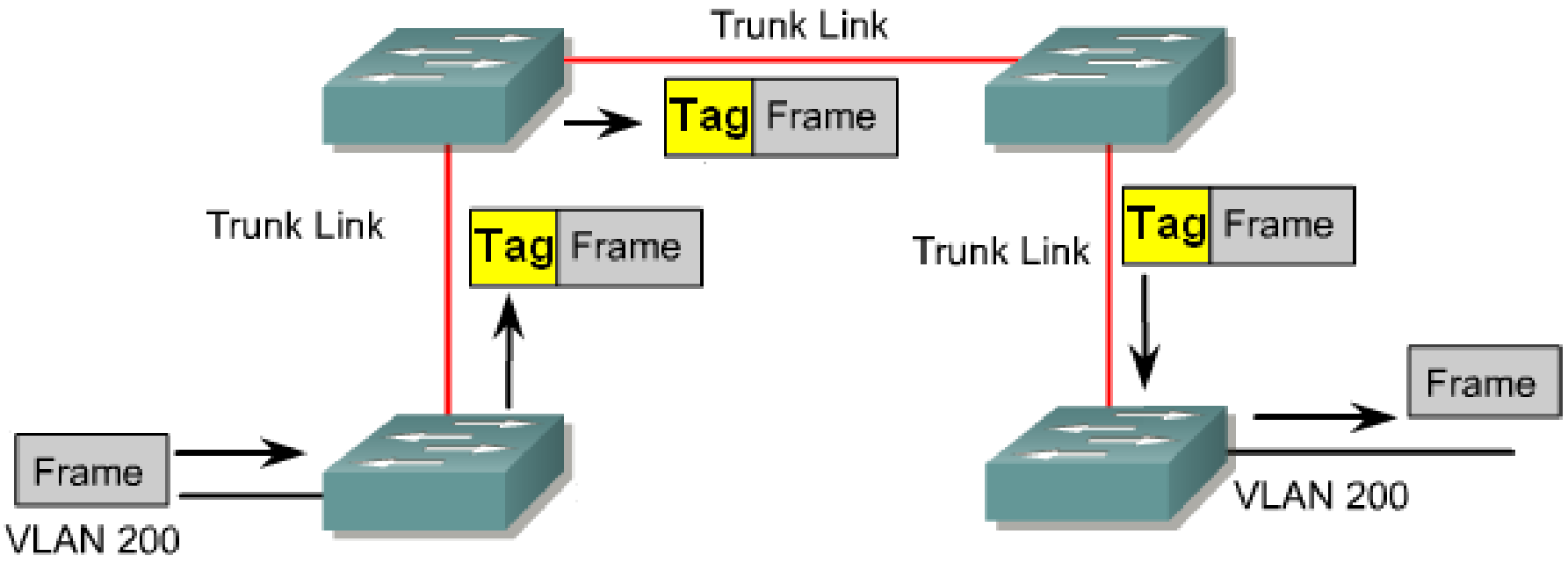
intecs
the Brainware Company

Spare

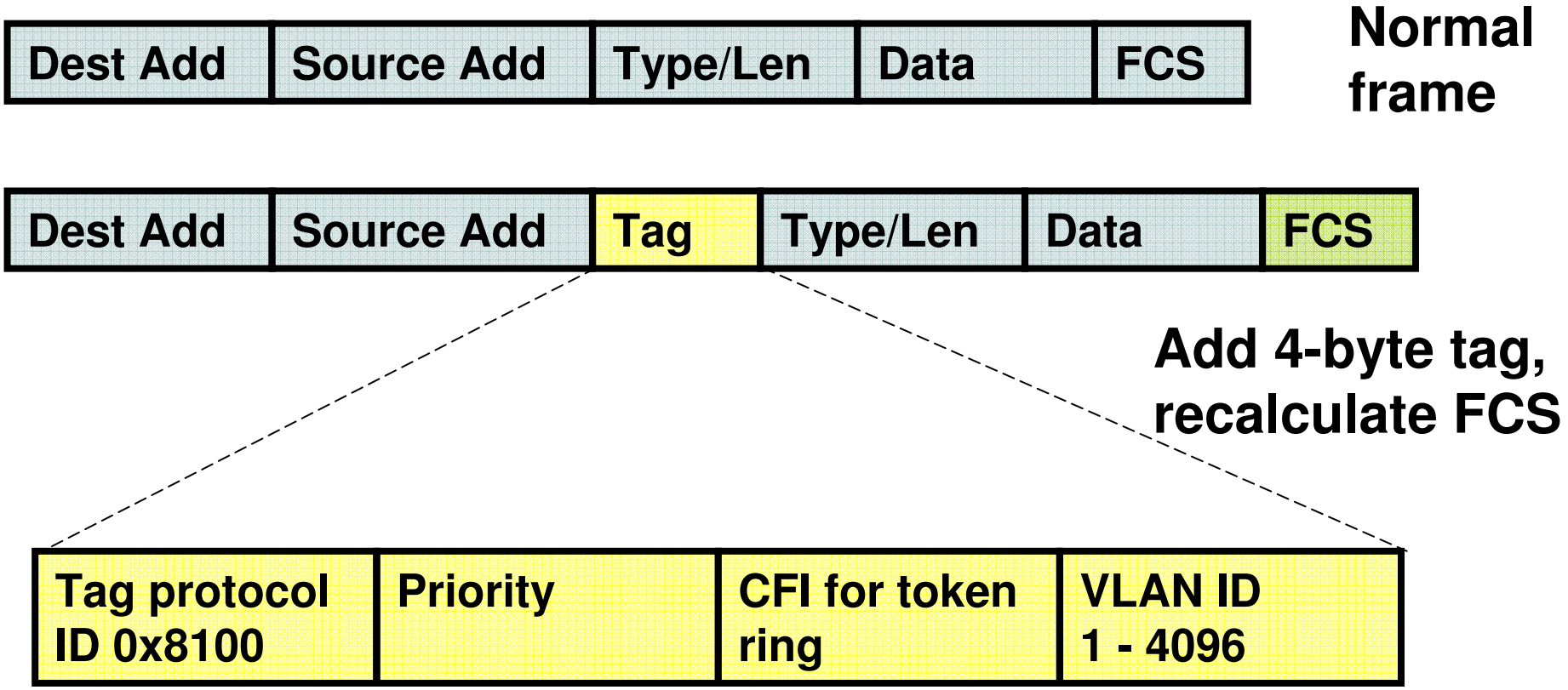
TechnoLabs

Tag to identify VLAN

- Tag is added to the frame when it goes on to the trunk
- Tag is removed when it leaves the trunk



Frame tagging IEEE 802.1Q



Port Mirroring

- Use as a debugging tool
- Provides the ability to study problematic packets

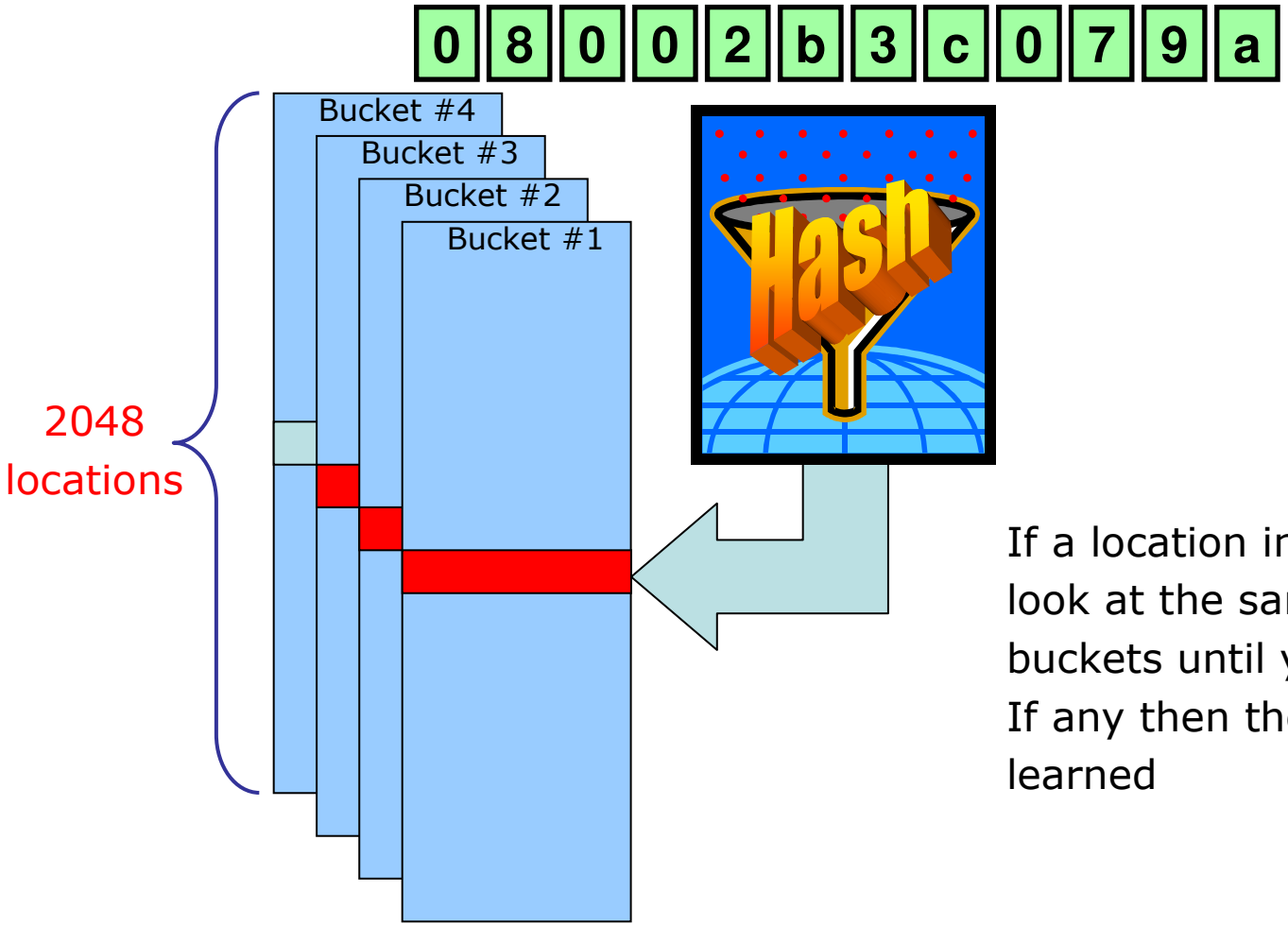
- Check the Port Mirroring Feature on all ports

TechnoLabs

IEEE 802.1P: Send Most Important Traffic First

- Priority Queuing
  - Four priority queues per port
  - WFQ and strict priority algorithm support on a port or 802.1p based criteria
  - Priority class handling: Priority to CoS conversion support (4 queues) according to 802.1p
- Read and write an IEEE 802.1p priority tag
  - If an IEEE 802.1p tag exists, the switch reads a packet's existing priority and maps it to the appropriate queue
  - If an IEEE 802.1p tag exists, the switch is able to remove the tag and regenerate the frame untagged
  - If an IEEE 802.1p tag does not exist, the switch is able to regenerate the frame tagged with a priority specified by the user

MAC Table organization



If a location in a bucket is full
look at the same location in the other buckets until you find an empty location
If any then the address will NOT be learned

Note: you can experience that at the 5° address if they collide on the same location