Politecnico di Torino
Laurea Magistrale in Ingegneria Informatica

appunti di

Progetto di reti locali

Autori principali: Luca Ghio
Docenti: Fulvio Giovanni Ottavio Risso, Guido Marchetto
Anno accademico: 2013/2014
Versione: 1.0.4.1
Data: 20 agosto 2017

Ringraziamenti

Speciali ringraziamenti vanno ad Andrea Marcelli per il suo contributo.

Oltre agli autori precedentemente citati, quest'opera può includere contributi da opere correlate su WikiAppunti e su Wikibooks, perciò grazie anche a tutti gli utenti che hanno apportato contributi agli appunti *Progetto di reti locali* e al libro *Progetto di reti locali*.

Informazioni su quest'opera

Quest'opera è pubblicata gratuitamente. Puoi scaricare l'ultima versione del documento PDF, insieme al codice sorgente LaTeX, da qui: http://lucaghio.webege.com/redirs/12

Quest'opera non è stata controllata in alcun modo dai professori e quindi potrebbe contenere degli errori. Se ne trovi uno, sei invitato a correggerlo direttamente tu stesso realizzando un commit nel repository Git pubblico o modificando gli appunti *Progetto di reti locali* su WikiAppunti, oppure alternativamente puoi contattare l'autore principale inviando un messaggio di posta elettronica a artghio@tiscali.it.

Licenza

Indice

Parte I

Nozioni di base sulle LAN

Capitolo 1

Introduzione alle Local Area Network

1.1 Origini

1.1.1 Definizione di LAN

Il gruppo di lavoro IEEE 802 definì la **rete locale** (LAN, dall'acronimo inglese "Local Area Network") come un sistema di comunicazione attraverso un mezzo condiviso, che permette a dispositivi indipendenti di comunicare tra loro entro un'area limitata, utilizzando un canale di comunicazione ad alta velocità e affidabile.

Parole chiave

- mezzo condiviso: tutti sono attaccati allo stesso mezzo di comunicazione;
- dispositivi indipendenti: tutti sono paritetici, cioè hanno gli stessi privilegi nel poter parlare (no interazione client-server);
- area limitata: tutti si trovano nella stessa area locale (ad es. azienda, campus universitario) e sono distanti al massimo qualche chilometro l'uno dall'altro (no attraversamento del suolo pubblico);
- ad alta velocità: all'epoca le velocità delle LAN si misuravano in Megabit al secondo (Mbps), mentre le velocità delle WAN in bit al secondo;
- affidabile: i guasti sono poco frequenti $\Rightarrow$ i controlli sono meno sofisticati a vantaggio delle prestazioni.

1.1.2 Confronto LAN vs. WAN

I protocolli per le **reti geografiche** (WAN, dall'acronimo inglese "Wide Area Network") e per le reti locali si evolverono in maniera indipendente fino agli '80 perché gli scopi erano differenti. Negli anni '90 la tecnologia IP permise finalmente di interconnettere questi due mondi.

WAN Le WAN nacquero negli anni '60 per connettere i terminali remoti ai pochi mainframe esistenti:

- mezzo fisico di comunicazione: linea dedicata punto punto su lunga distanza;
- proprietà del mezzo fisico: l'amministratore della rete deve affittare i cavi dal monopolio di Stato;

- modello di utilizzo: regolare, cioè piccola occupazione di banda per lunghi periodi di tempo (ad es. sessione di terminale);
- tipo di comunicazione: sempre unicast, più comunicazioni allo stesso tempo;
- qualità del mezzo fisico: elevata frequenza dei guasti, velocità basse, elevata presenza di disturbi elettromagnetici;
- costi: elevati, anche in termini di costi operativi (ad es. canone di affitto dei cavi);
- sistema di commutazione intermedio: richiesto per gestire le comunicazioni su larga scala (ad es. commutatori telefonici) ⇒ gli apparati di commutazione possono guastarsi.

LAN Le LAN comparvero alla fine degli anni '70 per condividere le risorse (come stampanti, dischi) tra piccoli gruppi di lavoro (ad es. dipartimenti):

- mezzo fisico di comunicazione: architettura a bus condiviso multi-punto su breve distanza;
- proprietà del mezzo fisico: l'amministratore della rete possiede i cavi;
- modello di utilizzo: bursty, cioè picchi di dati di breve durata (ad es. stampa di un documento) seguiti da lunghi silenzi;
- tipo di comunicazione: sempre broadcast, una sola comunicazione allo stesso tempo;
- qualità del mezzo fisico: maggiore affidabilità contro i guasti, velocità elevate, minore esposizione a disturbi esterni;
- costi: ragionevoli, concentrati principalmente nell'installazione della rete;
- sistema di commutazione intermedio: non richiesto ⇒ costo minore, velocità più alta, maggiore affidabilità, maggiore flessibilità nell'aggiungere e rimuovere stazioni.

1.1.3 Condivisione del mezzo di comunicazione

Prima dell'avvento di hub e bridge, il mezzo di comunicazione condiviso poteva essere implementato in due modi:

- broadcast fisico: tecnologie basate sul broadcast, come il bus: il segnale inviato da una stazione si propaga a tutte le altre stazioni;
- broadcast logico: tecnologie punto punto, come il token ring: il segnale inviato da una stazione arriva alla stazione successiva, la quale lo duplica verso la stazione ancora successiva, e così via.

Problemi

- privacy: tutti possono sentire quello che passa nel mezzo condiviso ⇒ è necessario realizzare un sistema di indirizzamento (oggi: gli indirizzi MAC);
- concorrenza: è possibile solo una comunicazione alla volta:
  - collisioni: se due stazioni trasmettono contemporaneamente, i dati inviati da una stazione potrebbero sovrapporsi ai dati inviati dall'altra ⇒ è necessario realizzare un meccanismo per il rilevamento delle collisioni e il recupero da esse (oggi: il protocollo CSMA/CD, si rimanda alla sezione 2.3);
  - monopolizzazione del canale: nella **trasmissione back to back**, una stazione può occupare il canale per un lungo periodo di tempo impedendo alle altre stazioni di parlare ⇒ è necessario realizzare una sorta di **multiplexing statistico**, cioè simulare più comunicazioni in contemporanea definendo un'unità di trasmissione massima detta **chunk** e alternando i chunk di una stazione con quelli di un'altra (oggi: le trame Ethernet).

1.2 Sottolivelli data-link

Nelle LAN il livello data-link è diviso in due sottolivelli:

- MAC: arbitra l'accesso al mezzo fisico, ed è specifico per ogni tecnologia di livello fisico (sezione 1.2.1);
- LLC: definisce l'interfaccia verso il livello rete, ed è comune in tutte le tecnologie di livello fisico (sezione 1.2.2).

1.2.1 MAC

Ogni scheda di rete è identificata in modo univoco da un **indirizzo MAC**. Gli indirizzi MAC hanno il seguente formato:

24	48
OUI	NIC ID

Tabella 1.1: Formato degli indirizzi MAC (6 byte).

dove i campi sono:

- campo Organization Unique Identifier (OUI) (3 byte): codice assegnato univocamente da IEEE per identificare il costruttore della scheda di rete:
  - primo bit meno significativo del primo byte:[1]
    - Individual (valore 0): l'indirizzo è associato a una singola stazione (unicast);
    - Group (valore 1): l'indirizzo fa riferimento a più stazioni (multicast/broadcast);
  - secondo bit meno significativo del primo byte:[1]
    - Universal (valore 0): l'indirizzo è assegnato univocamente;
    - Local (valore 1): l'indirizzo è personalizzato dall'utente;
- campo NIC Identifier (NIC ID) (3 byte): codice assegnato univocamente dal costruttore per identificare la specifica scheda di rete (detta anche "Network Interface Controller" [NIC]).

L'intestazione **Media Access Control** (MAC) ha il seguente formato:

48	96	112	da 46 a 1500 byte	4 byte
Destination Address	Source Address	Length	payload	FCS

Tabella 1.2: Formato dell'intestazione MAC (18 byte).

dove i campi sono:

- campo Destination Address (6 byte): specifica l'indirizzo MAC della destinazione.
  È messo prima dell'indirizzo MAC della sorgente perché così la destinazione lo può elaborare prima e scartare la trama se non è indirizzata ad essa;
- campo Source Address (6 byte): specifica l'indirizzo MAC della sorgente (sempre unicast);
- campo Length (2 byte): specifica la lunghezza del payload;

[1] Secondo l'ordine canonico (network byte order), che è l'ordine nativo in IEEE 802.3 (Ethernet) ma non in IEEE 802.5 (token ring) (si veda la sezione Bit-reversed notation nella voce *MAC address* su Wikipedia in inglese).

- campo Frame Control Sequence (FCS) (4 byte): contiene il codice CRC per il controllo di integrità sull'intera trama.
  Se il controllo del codice CRC fallisce, la trama arrivata è stata corrotta (ad es. a causa di una collisione) e viene scartata; i meccanismi di livello superiore (ad es. il TCP) si occuperanno di recuperare l'errore reinviando la trama.

Una scheda di rete quando riceve una trama:

- se l'indirizzo MAC di destinazione coincide con quello della scheda di rete o è di tipo broadcast ("FF-FF-FF-FF-FF-FF"), la accetta e la manda ai livelli superiori;
- se l'indirizzo MAC di destinazione non coincide con quello della scheda di rete, la scarta.

Una scheda di rete impostata in modalità promiscua accetta tutte le trame ⇒ serve per lo sniffing di rete.

1.2.2 LLC

L'intestazione **Logical Link Control** (LLC) ha il seguente formato:

8	16	24 o 32
DSAP	SSAP	CTRL

Tabella 1.3: Formato dell'intestazione LLC (3 o 4 byte).

dove i campi sono:

- campo DSAP (1 byte, di cui 2 bit riservati): identifica il protocollo di livello superiore utilizzato dalla destinazione;
- campo SSAP (1 byte, di cui 2 bit riservati): identifica il protocollo di livello superiore utilizzato dalla sorgente;
- campo Control (CTRL) (1 o 2 byte): deriva dal campo di controllo HDLC, ma è inutilizzato.

Problemi dei campi DSAP e SSAP

- intervallo di valori limitato: sono codificabili solo 64 protocolli;
- codici assegnati da ISO: corrispondono a dei codici solo i protocolli pubblicati da una organizzazione degli standard internazionalmente riconosciuta, mentre sono esclusi i protocolli definiti da altri organi o spinti da alcuni fornitori (ad es. IP);
- ridondanza di codici: non c'è alcun motivo di avere due campi per definire i protocolli, perché la sorgente e la destinazione parlano sempre lo stesso protocollo (ad es. entrambe IPv4 o entrambe IPv6).

SNAP

Il **Subnetwork Access Protocol** (SNAP) è una particolare implementazione dell'LLC per i protocolli che non hanno un codice standard.

L'intestazione LLC SNAP ha il seguente formato:

8	16	24	48	64
DSAP (0xAA)	SSAP (0xAA)	CTRL (3)	OUI	Protocol Type

Tabella 1.4: Formato dell'intestazione LLC SNAP (8 byte).

dove i campi sono:

- campi DSAP, SSAP, CTRL: i campi LLC sono fissi per indicare la presenza dell'intestazione SNAP;
- campo Organization Unique Identifier (OUI) (3 byte): identifica l'organizzazione che ha definito il protocollo.
  Se è uguale a 0, il valore nel campo "Protocol Type" corrisponde a quello utilizzato in Ethernet DIX;
- campo Protocol Type (2 byte): identifica il protocollo di livello superiore (ad es. 0x800 = IP, 0x806 = ARP).

In realtà, l'intestazione LLC SNAP non è molto utilizzata a causa del suo spreco di byte, a vantaggio del campo "Ethertype" di Ethernet DIX (si rimanda alla sezione 2.1.1).

Capitolo 2

Ethernet

Ethernet è la tecnologia oggi più utilizzata nelle LAN cablate con architettura a bus condiviso, perché è una soluzione semplice e poco costosa rispetto ad altre tecnologie per le LAN come token ring e token bus.

2.1 Formato della trama Ethernet

Esistono due versioni di Ethernet, con formati delle trame diversi:

- **Ethernet II DIX** (1982): versione sviluppata da DEC, Intel e Xerox (da qui l'acronimo "DIX") (sezione 2.1.1);
- standard **IEEE 802.3** (1983): versione standardizzata dal gruppo di lavoro IEEE 802 (sezione 2.1.2).

A causa della presenza di due versioni di Ethernet, esiste una notevole disomogeneità nell'imbustamento dei protocolli di livello superiore:

- i protocolli più vecchi (ad es. l'IP) e i protocolli più lontani da IEEE utilizzano l'imbustamento Ethernet II DIX;
- i protocolli standardizzati sin dall'inizio da IEEE (ad es. lo STP) utilizzano l'imbustamento IEEE 802.3.

2.1.1 Ethernet II DIX

Il pacchetto[1] Ethernet II DIX ha il formato seguente:

7 byte	1 byte	6 byte	6 byte	2 byte	da 46 a 1500 byte	4 byte	12 byte
preambolo	SFD	indirizzo MAC di destinazione	indirizzo MAC sorgente	EtherType	payload	FCS	IFG
		trama Ethernet II DIX (da 64 a 1518 byte)					

Tabella 2.1: Formato del pacchetto Ethernet II DIX (da 84 a 1538 byte).

dove i campi più significativi sono:

- preambolo (7 byte): sequenza di bit per recuperare la sincronizzazione tra il clock del trasmettitore e il clock del ricevitore.
  Il preambolo può accorciarsi ogniqualvolta il pacchetto attraversa un hub ⇒ non è possibile collegare più di 4 hub in cascata (si rimanda alla sezione 3.1);

[1] Lo standard chiama "pacchetto Ethernet" la trama Ethernet + i campi "Preamble", "SFD" e "IFG" del livello fisico.

- campo Start of Frame Delimiter (SFD) (1 byte): sequenza di bit che identifica l'inizio della trama;
- campo EtherType (2 byte): identifica il protocollo di livello superiore utilizzato nel payload (è un numero maggiore o uguale a 1500);
- campo Inter-Frame Gap (IFG) (12 byte): silenzio, cioè assenza di segnale, che identifica la fine della trama.

2.1.2 IEEE 802.3

Il pacchetto IEEE 802.3 può avere uno dei due formati seguenti:

7 byte	1 byte	14 byte	3 byte	da 0 a 1497 byte	da 0 a 43 byte	4 byte	12 byte
preambolo	SFD	intestazione MAC	intestazione LLC	payload	padding	FCS	IFG
		trama IEEE 802.3 (da 64 a 1518 byte)					

Tabella 2.2: Formato del pacchetto IEEE 802.3 con intestazione LLC (da 84 a 1538 byte).

7 byte	1 byte	14 byte	8 byte	da 0 a 1492 byte	da 0 a 38 byte	4 byte	12 byte
preambolo	SFD	intestazione MAC	intestazione LLC SNAP	payload	padding	FCS	IFG
		trama IEEE 802.3 (da 64 a 1518 byte)					

Tabella 2.3: Formato del pacchetto IEEE 802.3 con intestazione LLC SNAP (da 84 a 1538 byte).

Osservazioni

- le trame Ethernet II DIX e IEEE 802.3 hanno le stesse lunghezze minima e massima, perché IEEE doveva specificare un formato di trama compatibile con la vecchia versione di Ethernet;
- una trama Ethernet II DIX e una trama IEEE 802.3 si possono distinguere guardando il valore nel campo che segue l'indirizzo MAC sorgente:
  - se è minore o uguale a 1500 (campo "Length"), la trama è IEEE 802.3;
  - se è maggiore o uguale a 1536 (campo "EtherType"), la trama è Ethernet II DIX;
- nella trama IEEE 802.3 il campo "Length" renderebbe superfluo il campo "Inter-Frame Gap" (IFG), ma è presente per mantenere la compatibilità con la trama Ethernet II DIX;
- nella trama Ethernet II DIX il livello superiore deve trasmettere almeno 46 byte, mentre nella trama IEEE 802.3 la trama può essere allungata alla dimensione minima con del padding a seconda delle necessità;
- le intestazioni LLC e LLC SNAP della trama IEEE 802.3 sprecano molti più byte rispetto al campo "EtherType" della trama Ethernet II DIX pur servendo alla stessa funzione di indicare il protocollo di livello superiore, e per questo motivo lo standard IEEE 802.3 non è stato adottato in larga scala a vantaggio di Ethernet II DIX.

2.2 Livello fisico

Ethernet a 10 Mbps può viaggiare sui seguenti mezzi fisici di trasmissione:

- **cavo coassiale**: (sezione 2.2.1)
  - 10Base5: cavo spesso (max 500 m);
  - 10Base2: cavo sottile (max 185 m);
- **doppino di rame**: (sezione 2.2.2)
  - 10BaseT: cavo con 4 coppie intrecciate di cui solo 2 usate (max 100 m):
    - Unshielded (UTP): non schermato;
    - Shielded (STP): schermato con singola calza globale;
    - Foiled (FTP): schermato con singola calza globale + una calza per ogni coppia;
- **fibra ottica** (max 1-2 km) (sezione 2.2.3)

2.2.1 Cavo coassiale

Agli inizi il bus condiviso era fisicamente realizzato con un cavo coassiale:

- prese a vampiro: ogni scheda di rete è collegata a un cavo coassiale spesso tramite una presa a vampiro, che permetteva la propagazione elettrico per contatto fisico (continuità galvanica) ⇒ collegamento scomodo;
- connettori a T: ogni scheda di rete è collegata a un cavo coassiale sottile tramite un connettore a T ⇒ la connessione e la disconnessione di un host richiede di scollegare l'intera rete.

2.2.2 Doppino di rame

Con l'introduzione del doppino di rame, il cablaggio (cioè la posa dei cavi negli edifici) acquisì una maggiore flessibilità: ogni host può essere collegato a una presa a muro RJ45 tramite l'apposito connettore RJ45, e tutte le prese sono collegate a loro volta a un armadio.

Inoltre, alla presa a muro RJ45 è possibile collegare anche il connettore RJ11 usato dalla telefonia ⇒ nel cablaggio è possibile mettere delle prese RJ45 nell'intero edificio e poi decidere di volta in volta se collegare una scheda Ethernet o un telefono, commutando tra il collegamento dati e il collegamento telefonico nell'armadio.

2.2.3 Fibra ottica

Caratteristiche

- no sensibilità alle interferenze elettromagnetiche
- distanze maggiori
- costi più elevati
- flessibilità minore

2.3 CSMA/CD

Si ha una **collisione** quando due o più nodi nello stesso dominio di collisione[2] trasmettono contemporaneamente e i loro segnali si sovrappongono. Il protocollo **Carrier Sense Multiple Access with Collision Detection** (CSMA/CD) specifica come riconoscere una collisione (CD) e come recuperare una collisione (ritrasmissione).

Il CSMA/CD è un **protocollo ad accesso casuale** (cioè non deterministico) semplice e distribuito: non prevede apparati intermedi né meccanismi di sincronizzazione particolari, a differenza del token ring dove il meccanismo di sincronizzazione è il token stesso ⇒ il protocollo CSMA/CD è efficiente in termini di throughput perché non c'è l'overhead per la sincronizzazione, in termini di ritardi e di occupazione del canale.

In modalità full duplex non è più necessario abilitare il protocollo CSMA/CD (si rimanda alla sezione 3.2.1).

2.3.1 Rilevamento delle collisioni

Anziché trasmettere l'intera trama e soltanto alla fine verificare la collisione, il nodo può usare il **Collision Detection** (CD): durante la trasmissione ogni tanto cerca di capire se si è verificata una collisione ("listen while talking"), e in caso affermativo interrompe subito la trasmissione, evitando di sprecare il canale per una trasmissione inutile.

Nel mondo reale, il rilevamento delle collisioni avviene in due modi diversi a seconda del tipo di mezzo trasmissivo:

- cavo coassiale: c'è un singolo canale sia per la trasmissione sia per la ricezione ⇒ basta misurare la DC media sul link;
- doppino di rame, fibra ottica: ci sono due canali, uno per la trasmissione e l'altro per la ricezione:
  - stazioni trasmittenti: possono accorgersi che si è verificata una collisione rilevando attività sul canale di ricezione durante la trasmissione;
  - stazioni non trasmittenti: possono accorgersi che si è verificata una collisione solo rilevando un codice CRC errato sulla trama ricevuta.
    La **sequenza di jamming** è un segnale potente che viene mandato da chi si è accorto di una collisione per garantire che il codice CRC sia non valido e per massimizzare la probabilità che tutti gli altri nodi abbiano capito che è avvenuta una collisione.

2.3.2 Riduzione del numero di collisioni

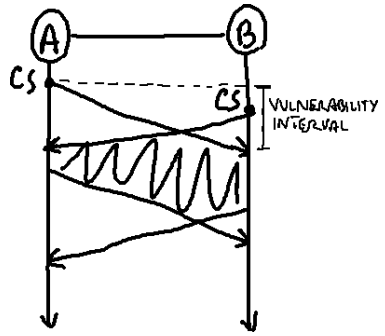
Il **Carrier Sense** (CS) permette di ridurre il numero di collisioni: il nodo che vuole trasmettere ascolta il canale prima della trasmissione:

- se sente che il canale è libero: il nodo trasmette la trama;
- se sente che il canale è occupato:
  - CSMA 1-persistente: il nodo continua a verificare se il canale è libero e trasmette appena si libera;
  - CSMA 0-persistente: il nodo riprova dopo un tempo casuale;
  - CSMA p-persistente: il nodo con probabilità $1 - p$ aspetta un tempo casuale (0-persistente), con probabilità p riverifica subito (1-persistente).

In una LAN nel caso peggiore l'occupazione del canale è pari al 30-40% della banda disponibile ⇒ Ethernet implementa il CSMA/CD 1-persistente perché è pensato per reti mediamente scariche con ridotta probabilità di collisioni.

[2] Si veda la sezione 3.1.

Limiti del CSMA Tuttavia, con il doppino di rame o la fibra ottica il CSMA non è in grado di evitare del tutto le collisioni (altrimenti non servirebbe il CD): se si tiene conto dei tempi di propagazione, un nodo lontano può sentire il canale libero, anche se in realtà è occupato ma la trasmissione non ha ancora raggiunto il nodo lontano.



Si dice **intervallo di vulnerabilità** l'intervallo di tempo in cui l'avvio di una trasmissione da parte del nodo lontano creerebbe una collisione (è pari al ritardo di propagazione sul canale), e questo intervallo è tanto grande quando la distanza è maggiore ⇒ questo protocollo funziona bene su reti piccole.

2.3.3 Recupero delle collisioni

Dopo che si è verificata una collisione, la trama deve essere ritrasmessa. Se le stazioni coinvolte nella collisione ritrasmettessero subito, si verificherebbe un'altra collisione ⇒ l'**algoritmo di back-off** inserisce nell'attesa un elemento di casualità esponenziale nelle ritrasmissioni:

- 1ª ritrasmissione: il nodo aspetta un tempo T scelto a caso tra 0 e 1 slot time;
- 2ª ritrasmissione: il nodo aspetta un tempo T scelto a caso da 0 a 3 slot time;
- 3ª ritrasmissione: il nodo aspetta un tempo T scelto a caso da 0 a 7 slot time;

e così via, secondo la formula:

$$T = r \cdot \tau, \quad 0 \leq r < 2^k,\ k = \min(n, 10),\ n \leq 16$$

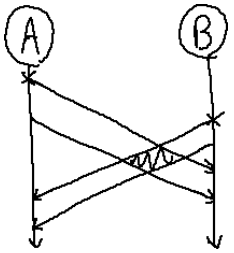
dove:

- n è il numero di collisioni avvenute sulla trama corrente;
- τ è lo **slot time**, ovvero il tempo richiesto per inviare una trama Ethernet di dimensione minima (64 byte), equivalente a 51,2 μs.

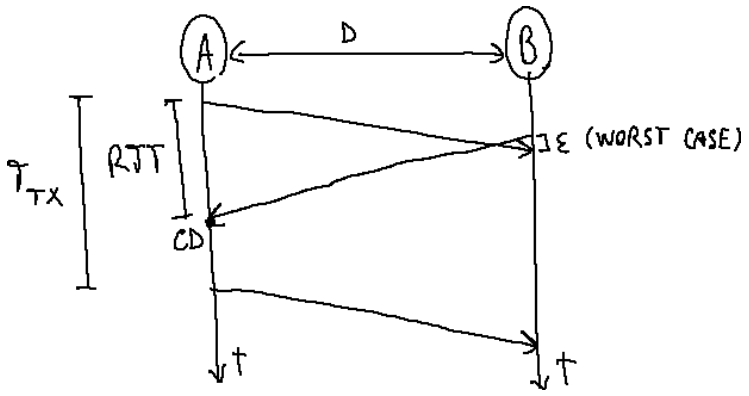
Al termine di ogni attesa, il nodo ascolta di nuovo il canale con il CS.

2.3.4 Vincolo tra la dimensione della trama e il diametro di collisione

Siccome l'accesso al canale è conteso, quando si riesce a ottenere l'accesso alla rete conviene trasmettere pacchetti grandi. È necessario stabilire una dimensione minima per le trame: se la trama è troppo piccola e la trasmissione collisa dura troppo poco, può avvenire che nessun nodo si accorga della collisione:



Esiste un vincolo tra la dimensione della trama L_{PDU} e il diametro di collisione D affinché tutte le collisioni siano riconosciute: il rilevamento delle collisioni funziona solo se il round trip time RTT, cioè il tempo di andata e ritorno, è minore del tempo di trasmissione T_{TX}:



$$\text{RTT} < T_{\text{TX}} \Rightarrow 2\frac{D}{V_{\text{PR}}} < \frac{L_{\text{PDU}}}{V_{\text{TX}}} \Rightarrow \begin{cases} L_{\text{PDU}} > \dfrac{V_{\text{TX}} \cdot 2D}{V_{\text{PR}}} \\ D < \dfrac{V_{\text{PR}} \cdot L_{\text{PDU}}}{2V_{\text{TX}}} \end{cases}$$

dove V_{TX} è la velocità di trasmissione e V_{PR} è la velocità di propagazione.

Aumentare la velocità di trasmissione significa aumentare la dimensione minima delle trame, oppure a parità di dimensione minima significa diminuire la distanza massima tra i nodi, ma trame troppo grandi aumenterebbero la probabilità di errore della trasmissione e intaserebbero la rete.

In Ethernet DIX il diametro di collisione teorico non può superare i 5750 metri:[3]

$$\begin{cases} L_{\text{PDU}_{\min}} = 72 \text{ byte} \\ V_{\text{TX}} = 10 \text{ Mbps} \\ V_{\text{PR}} = c = 200000 \text{ km/s} \end{cases} \Rightarrow D_{\max} = \frac{V_{\text{PR}} \cdot L_{\text{PDU}_{\min}}}{2V_{\text{TX}}} = 5750 \text{ m}$$

In assenza di hub le dimensioni massime della rete sono piuttosto limitate dalle distanze massime supportate dai mezzi trasmissivi (ad es. a causa dell'attenuazione del segnale). Grazie agli hub è possibile estendere la dimensione della rete (anche se al massimo a 3 km per via delle non idealità del dispositivo): l'hub, tipicamente posto come centro stella di una topologia a stella, rigenera il segnale (repeater) e simula internamente il bus condiviso permettendo di collegare tra loro più stazioni attraverso il doppino di rame (si rimanda alla sezione 3.1).

[3]Per la lunghezza della trama L_{PDU} si considerano il preambolo e lo SFD, ma non l'IFG.

Capitolo 3

Repeater e bridge

3.1 Interconnessione a livello fisico

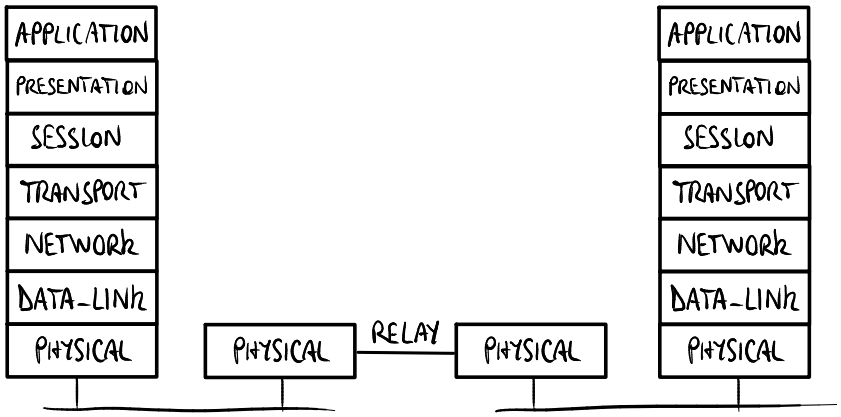


Figura 3.1: Interconnessione a livello fisico nella pila OSI.

Il **repeater** e l'**hub**[1] sono dispositivi di rete per l'interconnessione a livello fisico, che si limitano a ricevere e propagare una sequenza di bit. I canali di livello fisico interconnessi possono anche essere di tecnologia diversa (ad es. da doppino a fibra ottica), ma tutti i livelli superiori devono essere uguali (ad es. solo Ethernet, solo FDDI).

Il repeater svolge anche la funzione di recupero della degradazione del segnale: si sincronizza con il segnale a onda quadra e lo rigenera in modo da pulirlo. Il preambolo che precede la trama serve per la sincronizzazione, cioè per il riconoscimento del segnale, e ogni volta che il segnale attraversa un repeater viene così "mangiata" una parte di questo preambolo ⇒ non è possibile collegare più di 4 repeater in cascata.

Un **dominio di collisione** è l'insieme di nodi che concorrono per accedere allo stesso mezzo trasmissivo ⇒ la trasmissione contemporanea provoca collisione. L'interconnessione di due **segmenti di rete** crea un unico dominio di collisione: il repeater non è in grado di riconoscere le collisioni che vengono propagate su tutte le porte ⇒ è un limite alla dimensione del dominio fisico.

[1] La differenza tra hub e repeater sta nel numero di porte: il repeater ha due porte, l'hub ha più di due porte.

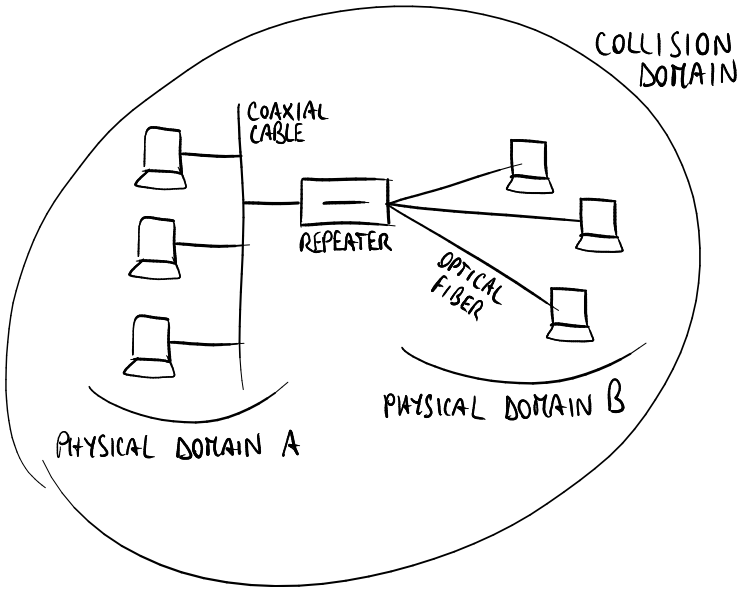


Figura 3.2: Esempio di interconnessione a livello fisico.

3.2 Interconnessione a livello data-link

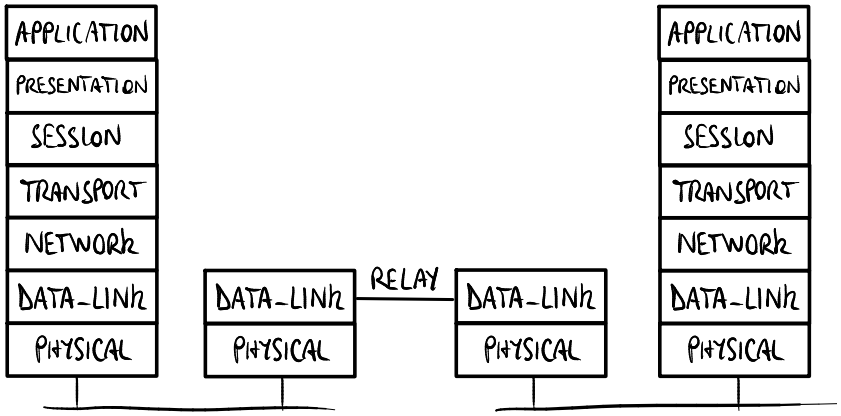


Figura 3.3: Interconnessione a livello data-link nella pila OSI.

Il **bridge** e lo **switch** sono dispositivi di rete per l'interconnessione a livello data-link, che memorizzano (modalità store and forward) e poi rigenerano la trama. Anche i domini di livello data-link interconnessi possono essere di tecnologia diversa (ad es. da Ethernet a FDDI).

Problema della dimensione massima delle trame In pratica è spesso impossibile interconnettere due tecnologie di livello data-link differenti, a causa di problemi legati alla dimensione massima delle trame: per esempio, in una rete basata su Ethernet avente MTU = 1518 byte interconnessa con una rete basata su token ring avente MTU = 4 KB, che cosa succede se dalla rete token ring giunge una trama maggiore di 1518 byte? Inoltre la frammentazione a livello data-link non esiste.

Il bridge disaccoppia il dominio di broadcast dal dominio di collisione:

- "spezza" il dominio di collisione: implementa il protocollo CSMA/CD per rilevare le collisioni, evitando di propagarle sulle altre porte;

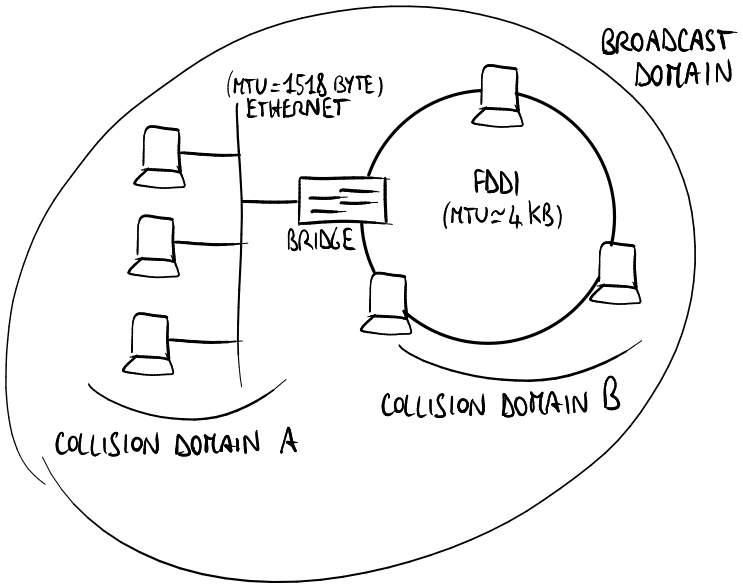


Figura 3.4: Esempio di interconnessione a livello data-link.

- estende il **dominio di broadcast**: le trame inviate in broadcast vengono propagate su tutte le porte.

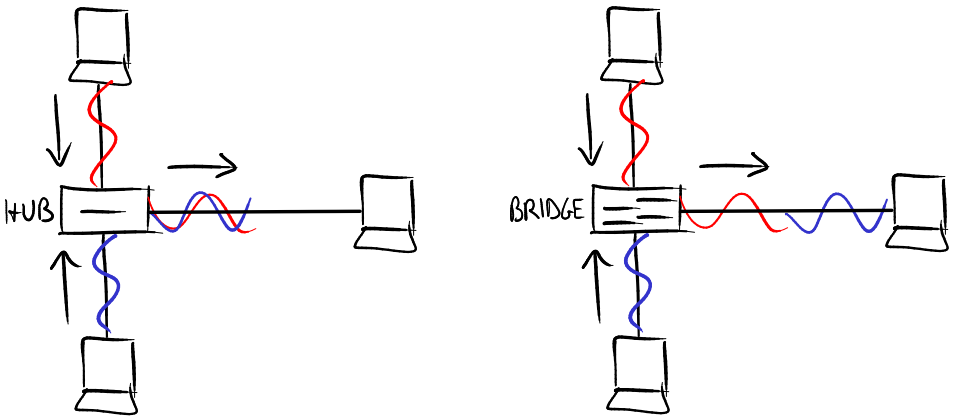


Figura 3.5: I bridge evitano di creare delle collisioni grazie alla modalità store and forward.

3.2.1 Modalità half duplex e full duplex

Un link punto-punto a livello data-link tra due nodi (ad es. un host e un bridge) può essere effettuato in due modi:

- **modalità half duplex**: i due nodi sono collegati con un unico filo bidirezionale ⇒ ogni nodo non può trasmettere e ricevere allo stesso tempo, perché si verificherebbe una collisione;
- **modalità full duplex**: i due nodi sono collegati con due fili unidirezionali separati ⇒ ogni nodo può trasmettere e ricevere allo stesso tempo, grazie alla separazione dei domini di collisione.

Vantaggi della modalità full duplex

- maggiore larghezza di banda: il throughput tra i due nodi raddoppia;
- assenza di collisioni:
  - non è più necessario abilitare il protocollo CSMA/CD;
  - non è più necessario il vincolo sulla dimensione minima delle trame Ethernet;
  - non esiste più il limite sul diametro massimo del dominio di collisione (l'unico limite di distanza è quello fisico del canale).

3.2.2 Transparent bridge

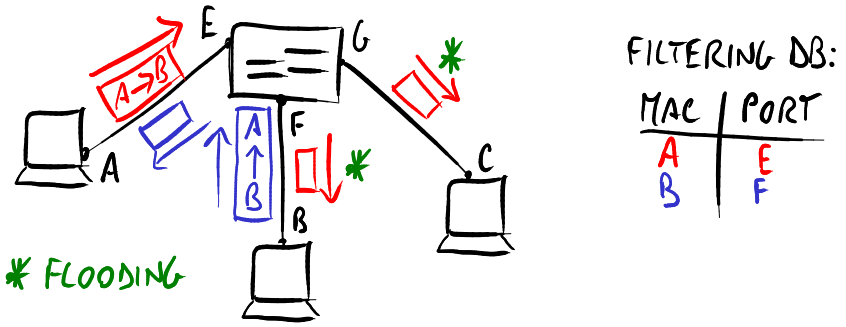


Figura 3.6: Esempio di funzionamento degli algoritmi di apprendimento.

L'instradamento è svolto in maniera trasparente: il bridge cerca di imparare la posizione dei nodi ad esso collegati riempiendo una tabella di inoltro chiamata **filtering database**, le cui entry hanno il seguente formato:

<indirizzo MAC> <porta di destinazione> <ageing time>

dove la porta di destinazione è la porta del bridge, appresa con gli algoritmi di apprendimento, da cui far uscire le trame dirette verso l'indirizzo MAC di destinazione associato.

Algoritmi di apprendimento

- **frame forwarding**: l'apprendimento è basato sull'indirizzo MAC di destinazione: quando arriva una trama la cui destinazione non è ancora presente nel filtering database, il bridge manda la trama in broadcast su tutte le porte tranne la porta d'ingresso (**flooding**) e attende la risposta che molto probabilmente la destinazione invierà e su cui entrerà in azione l'algoritmo di backward learning;
- **backward learning**: l'apprendimento è basato sull'indirizzo MAC sorgente: quando arriva una trama a una certa porta, il bridge verifica se la sorgente è già presente ed associata a quella porta nel filtering database, e se necessario lo aggiorna.

Il **processo di inoltro intelligente** aumenta la banda aggregata della rete: le trame non vengono più propagate sempre in broadcast su tutte le porte ma vengono inoltrate solo sulla porta verso la destinazione, lasciando gli altri link liberi di trasportare degli altri flussi nello stesso tempo.

Mobilità

L'**ageing time** consente di mantenere il filtering database aggiornato: esso viene impostato a 0 quando la entry viene creata o aggiornata dall'algoritmo di backward learning, e viene incrementato con il passare del tempo fino a quando non supera il tempo di scadenza e la entry non

viene rimossa. In questo modo il filtering database contiene informazioni sulle sole stazioni che sono effettivamente presenti nella rete, liberandosi dalle informazioni sulle vecchie stazioni.

Le reti di livello data-link supportano nativamente la **mobilità**: se la stazione viene spostata, rimanendo nella stessa LAN, in modo da essere raggiungibile da un'altra porta, il bridge va "notificato" dello spostamento tramite l'invio di un qualsiasi trama broadcast (ad es. ARP Request), in modo che l'algoritmo di backward learning possa correggerne il filtering database. I sistemi Windows tendono a essere più "loquaci" dei sistemi UNIX.

Esempi di stazioni che si possono spostare sono:

- cellulari;
- macchine virtuali nei datacenter: durante il giorno possono essere distribuite su più server Web per distribuire il carico di lavoro, durante la notte possono essere concentrate sullo stesso server Web perché il traffico è minore consentendo un risparmio di energia;
- stazioni collegate alla rete con due link, uno primario usato in condizioni normali e uno secondario tollerante ai guasti: quando il link primario si guasta, l'host può ripristinare la connettività inviando una trama in broadcast attraverso il link secondario.

3.2.3 Switch

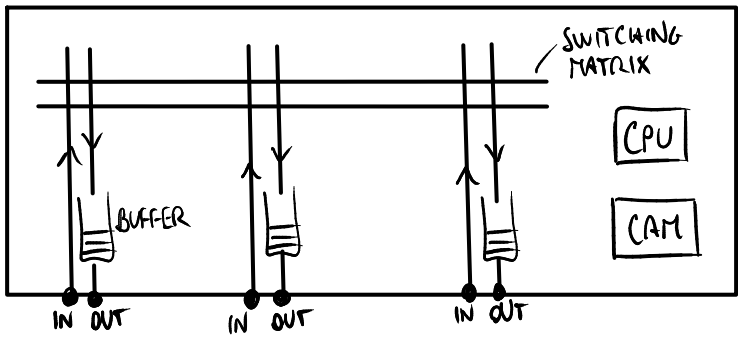


Figura 3.7: Architettura interna di uno switch.

"**Switch**" è il nome commerciale dato ai bridge dotati di funzionalità avanzate per enfatizzare le loro maggiori prestazioni:

- uno switch è un bridge multi-porta: uno switch è dotato di molte più porte, tipicamente tutte in modalità full duplex, rispetto a un bridge;
- il processo di inoltro intelligente non è più un componente software ma è implementato in un chip hardware (l'algoritmo di spanning tree rimane implementato in software perché più complesso);
- la ricerca della porta associata a un dato indirizzo MAC nel filtering database è più rapida grazie all'uso delle Content Addressable Memory (CAM), che però hanno un costo maggiore e un consumo energetico maggiore;
- gli switch supportano la tecnologia di inoltro cut-through più veloce della modalità store and forward: una trama può essere inoltrata sulla porta di destinazione (a meno che non già occupata) subito dopo la ricezione dell'indirizzo MAC di destinazione.

3.2.4 Problemi

Scalabilità

I bridge presentano dei problemi di scalabilità perché non sono in grado di organizzare il traffico, e perciò non sono adatti per reti complesse (come le reti geografiche):

- nessun filtraggio del traffico broadcast ⇒ su una rete ampia con molti host le trame broadcast rischiano di intasare la rete;
- l'algoritmo di spanning tree rende completamente inutilizzati alcuni link che formerebbero anelli in topologia sfavorendo il bilanciamento del carico (si rimanda alla sezione 6.6.2).

Sicurezza

Sono possibili alcuni attacchi al filtering database:

MAC flooding attack La stazione attaccante genera trame con indirizzi MAC sorgente casuali ⇒ il filtering database si riempie di indirizzi MAC di stazioni inesistenti, mentre quelli delle stazioni esistenti vengono buttati fuori ⇒ il bridge manda in flooding (quasi) tutte le trame provenienti dalle stazioni esistenti perché non riconosce più gli indirizzi MAC di destinazione ⇒ la rete viene rallentata e (quasi) tutto il traffico sulla rete viene ricevuto dalla stazione attaccante.[2]

Tempesta di pacchetti La stazione attaccante genera trame con indirizzi MAC di destinazione casuali ⇒ il bridge manda in flooding tutte le trame provenienti dalla stazione attaccante perché non riconosce gli indirizzi MAC di destinazione ⇒ la rete viene rallentata.

[2]Non tutto il traffico può essere intercettato: quando arriva una trama da una stazione esistente il bridge ne salva l'indirizzo MAC sorgente nel filtering database, così se subito dopo arriva una trama diretta a quella stazione, prima che la relativa entry venga cancellata, essa non verrà mandata in flooding.

Capitolo 4

Evoluzioni di Ethernet

Con il successo di Ethernet sorsero nuovi problemi:

- necessità di maggiore velocità: Ethernet II DIX supportava una velocità di trasmissione pari a 10 Mbps, mentre FDDI, utilizzata nel backbone, supportava una velocità molto più alta (100 Mbps), ma sarebbe stato troppo costoso cablare gli edifici con la fibra ottica;
- necessità di interconnettere più reti: reti di tecnologie diverse (ad es. Ethernet, FDDI, token ring) erano difficili da interconnettere perché avevano delle MTU diverse ⇒ avere la stessa tecnologia dappertutto avrebbe risolto questo problema.

4.1 Fast Ethernet

Fast Ethernet, standardizzato come IEEE 802.3u (1995), innalza la velocità di trasmissione a 100 Mbps e accorcia di conseguenza di 10 volte il massimo diametro di collisione (∼200-300 m), mantenendone lo stesso formato di trama e lo stesso algoritmo CSMA/CD.

4.1.1 Livello fisico

Il livello fisico di Fast Ethernet è completamente diverso dal livello fisico di Ethernet a 10 Mbps: deriva in parte da standard esistenti nel mondo FDDI, tanto che Fast Ethernet e FDDI sono compatibili a livello fisico, e abbandona definitivamente il cavo coassiale:

- 100BASE-T4: doppino di rame che utilizza 4 coppie;
- 100BASE-TX: doppino di rame che utilizza 2 coppie;
- 100BASE-FX: fibra ottica (solo nel backbone).

4.1.2 Adozione

Quando fu introdotto Fast Ethernet, il suo tasso di adozione fu piuttosto basso a causa di:

- limite di distanza: le dimensioni della rete erano limitate ⇒ Fast Ethernet non era appropriato per il backbone;
- colli di bottiglia nel backbone: il backbone realizzato in tecnologia FDDI a 100 Mbps aveva la stessa velocità delle reti di accesso in tecnologia Fast Ethernet ⇒ rischiava di non riuscire a smaltire tutto il traffico proveniente dalle reti di accesso.

Fast Ethernet cominciò a essere adottato più ampiamente con:

- l'introduzione dei bridge: interrompono il dominio di collisione superando il limite di distanza;
- l'introduzione di Gigabit Ethernet nel backbone: evita i colli di bottiglia nel backbone.

4.2 Gigabit Ethernet

Gigabit Ethernet, standardizzato come IEEE 802.3z (1998), innalza la velocità di trasmissione a 1 Gbps e introduce due funzionalità, la "Carrier Extension" e il "Frame Bursting", per mantenere funzionante il protocollo CSMA/CD.

4.2.1 Carrier Extension

La decuplicazione della velocità di trasmissione ridurrebbe il massimo diametro di collisione di altre 10 volte portandolo a qualche decina di metri, troppo pochi per il cablaggio ⇒ per mantenere invariato il massimo diametro di collisione, sarebbe necessario aumentare la dimensione minima della trama a 512 byte[1].

L'allungamento della trama minima però genererebbe un problema di incompatibilità: nell'interconnessione di una rete Fast Ethernet e di una rete Gigabit Ethernet con un bridge, le trame di dimensione minima provenienti dalla rete Fast Ethernet non potrebbero entrare nella rete Gigabit Ethernet ⇒ invece di allungare la trama fu allungato lo slot time, cioè l'unità temporale minima di trasmissione: in fondo a tutte le trame più corte di 512 byte viene aggiunta una **Carrier Extension** composta da bit fittizi di padding (fino a 448 byte):

7 byte	1 byte	da 64 a 1518 byte	da 0 a 448 byte	12 byte
preambolo	SFD	trama Ethernet II DIX/IEEE 802.3	Carrier Extension	IFG
		da 512 a 1518 byte		

Tabella 4.1: Formato del pacchetto Gigabit Ethernet (da 532 a 1538 byte).

Svantaggi

- la Carrier Extension occupa il canale con dei bit inutili.
  Ad esempio con trame da 64 byte il throughput utile è molto basso:

$$\frac{64 \text{ byte}}{512 \text{ byte}} \cdot 1 \text{ Gbit/s} = 125 \text{ Mbit/s}$$

- nelle reti moderne commutate pure è attiva la modalità full duplex ⇒ il CSMA/CD è disabilitato ⇒ la Carrier Extension è inutile.

4.2.2 Frame Bursting

La massima dimensione della trama di 1518 byte è ormai obsoleta: in Ethernet a 10 Mbps l'occupazione del canale era pari a 1,2 ms, un tempo ragionevole per garantire il multiplexing statistico[2], mentre in Gigabit Ethernet l'occupazione del canale è pari a 12 μs ⇒ le collisioni sono molto meno frequenti ⇒ per diminuire l'overhead delle intestazioni in rapporto ai dati utili migliorando l'efficienza, si potrebbe aumentare la dimensione massima della trama.

L'allungamento della trama massima però genererebbe un problema di incompatibilità: nell'interconnessione di una rete Fast Ethernet e di una rete Gigabit Ethernet con un bridge, le trame di dimensione massima provenienti dalla rete Gigabit Ethernet non potrebbero entrare nella rete Fast Ethernet ⇒ il **Frame Bursting** consiste nella concatenazione di diverse trame di dimensione standard una dopo l'altra, senza rilasciare il canale:

- solo la prima trama viene eventualmente estesa con la Carrier Extension, per assicurarsi che la finestra di collisione sia riempita; nelle trame successive la Carrier Extension non serve perché, se avvenisse una collisione, essa sarebbe individuata già con la prima trama;

[1] Teoricamente la trama sarebbe da allungare di 10 volte, quindi a 640 byte, ma lo standard decise diversamente.
[2] Si veda la sezione 1.1.3.
[3] preambolo + SFD + trama Ethernet II DIX/IEEE 802.3

trama 1^3 + Carrier Extension	FILL	trama 2^3	FILL	...	FILL	ultima trama^3	IFG
burst limit (8192 byte)							

Tabella 4.2: Formato del pacchetto Gigabit Ethernet con Frame Bursting.

- l'IFG tra una trama e l'altra viene sostituito con una "Filling Extension" (FILL) per delimitare i byte e annunciare che seguirà un'altra trama;
- la stazione trasmittente mantiene un contatore di byte: quando arriva al byte numero 8192, la trama correntemente in trasmissione deve essere l'ultima ⇒ con il Frame Bursting è possibile inviare fino a 8191 byte + 1 trama.

Vantaggi

- il numero di occasioni di collisione è diminuito: una volta che la prima trama è trasmessa senza collisioni, tutte le altre stazioni rilevano che il canale è occupato grazie al CSMA;
- le trame successive alla prima non richiedono la Carrier Extension ⇒ il throughput utile aumenta soprattutto in caso di trame piccole, grazie al risparmio della Carrier Extension.

Svantaggi

- il Frame Bursting non risponde all'obiettivo originario di ridurre l'overhead delle intestazioni: è stato scelto di mantenere in ogni trama tutte le intestazioni (compresi il preambolo, lo SFD e l'IFG) per semplificare l'hardware di elaborazione;
- tipicamente una stazione che utilizza il Frame Bursting deve inviare tanti dati ⇒ le trame grosse non richiedono la Carrier Extension ⇒ non c'è il risparmio della Carrier Extension;
- nelle reti moderne commutate pure è attiva la modalità full duplex ⇒ il CSMA/CD è disabilitato ⇒ il Frame Bursting non ha alcun vantaggio e perciò è inutile.

4.2.3 Livello fisico

Gigabit Ethernet può viaggiare sui seguenti mezzi fisici di trasmissione:

- doppino di rame:
  - Shielded (STP): lo standard 1000BASE-CX utilizza 2 coppie (25 m);
  - Unshielded (UTP): lo standard 1000BASE-T utilizza 4 coppie (100 m);
- fibra ottica: gli standard 1000BASE-SX e 1000BASE-LX utilizzano 2 fibre, e possono essere:
  - Multi-Mode Fiber (MMF): meno pregiate (275-550 m);
  - Single-Mode Fiber (SMF): ha una lunghezza massima di 5 km.

GBIC Gigabit Ethernet introdusse per la prima volta i **gigabit interface converter** (GBIC), che sono una soluzione comune per avere la possibilità di aggiornare il livello fisico senza dover aggiornare il resto dell'apparecchiatura: la scheda Gigabit Ethernet non ha il livello fisico integrato a bordo, ma include solo la parte logica (dal livello data-link in su), e l'utente può collegare negli appositi alloggiamenti della scheda il GBIC desiderato che implementa il livello fisico.

4.3 10 Gigabit Ethernet

10 Gigabit Ethernet, standardizzato come IEEE 802.3ae (2002), innalza la velocità di trasmissione a 10 Gbps e abbandona finalmente la modalità half duplex, eliminando tutte le problematiche derivanti dal CSMA/CD.

Non è ancora usato nelle reti di accesso, ma è principalmente usato:

- nei backbone: viaggia su fibra ottica (da 26 m a 40 km) perché il doppino di rame non è più sufficiente a causa dei limiti di attenuazione del segnale;
- nei datacenter: oltre alle fibre ottiche, si usano anche cavi molto corti per collegare i server agli switch top of the rack (TOR):[4]
  - Twinax: cavi coassiali, inizialmente usati perché le unità per la trasmissione sui doppini di rame consumavano troppa energia;
  - 10GBase T: doppini di rame schermati, aventi una latenza molto alta;
- nelle reti metropolitane (MAN) e nelle reti geografiche (WAN): 10 Gigabit Ethernet può essere trasportato sulle infrastrutture MAN e WAN già esistenti, anche se a una velocità di trasmissione ridotta a 9,6 Gb/s.

4.4 40 Gigabit Ethernet e 100 Gigabit Ethernet

40 Gigabit Ethernet e **100 Gigabit Ethernet**, entrambi standardizzati come IEEE 802.3ba (2010), innalzano la velocità di trasmissione rispettivamente a 40 Gbps e a 100 Gbps: per la prima volta l'evoluzione della velocità di trasmissione non è più a 10×, ma è stato deciso di definire uno standard a una velocità intermedia per via degli costi ancora elevati di 100 Gigabit Ethernet. Inoltre, 40 Gigabit Ethernet può essere trasportato sull'infrastruttura DWDM già esistente.

Queste velocità sono usate solo nel backbone perché non sono ancora adatte non solo per gli host, ma anche per i server, perché sono molto vicine alle velocità interne delle unità di elaborazione (bus, memoria, ecc.) ⇒ il collo di bottiglia non è più la rete.

[4] Si rimanda alla sezione 14.4.2.

Capitolo 5

Funzionalità avanzate sulle reti Ethernet

5.1 Autonegoziazione

L'**autonegoziazione** è una funzione orientata al plug-and-play: quando una scheda di rete si connette a una rete, manda degli impulsi con una codifica particolare per provare a determinare le caratteristiche della rete:

- modalità: half duplex o full duplex (su doppino);
- velocità di trasmissione: a partire dalla velocità più alta fino a quella più bassa (su doppino e fibra ottica).

Sequenza di negoziazione

- 1 Gb/s full duplex
- 1 Gb/s half duplex
- 100 Mb/s full duplex
- 100 Mb/s half duplex
- 10 Mb/s full duplex
- 10 Mb/s half duplex

5.1.1 Problemi

L'autonegoziazione è possibile solo se la stazione si connette a un altro host o a un bridge: gli hub infatti operano a velocità fissa, quindi non possono negoziare niente. Se durante la procedura l'altra parte non risponde, la stazione in negoziazione assume di essere connessa a un hub ⇒ imposta automaticamente la modalità a half duplex.

Se l'utente configura manualmente la propria scheda di rete a lavorare sempre in modalità full duplex disabilitando la funzione di autonegoziazione, quando si collega a un bridge quest'ultimo, non ricevendo risposta dall'altra parte, assume di essere collegato a un hub e imposta la modalità half duplex ⇒ l'host considera possibile inviare e ricevere nello stesso momento sul canale, mentre il bridge considera ciò una collisione sul canale condiviso ⇒ il bridge rileva molte collisioni che sono dei falsi positivi, e scarta erroneamente molte trame ⇒ ogni trama scartata viene recuperata dai meccanismi di recupero dagli errori del TCP, che però sono molto lenti ⇒ la velocità di accesso alla rete è molto bassa. Valori molto alti dei contatori di collisione su una specifica porta di un bridge sono sintomo di questa problematica.

5.2 Aumento della dimensione massima della trama

La specifica originale di Ethernet definisce:

- dimensione massima della trama: 1518 byte;
- dimensione massima del payload (Maximum Transmission Unit [MTU]): 1500 byte.

In molti casi però sarebbe utile avere una trama più grande del normale:

- intestazioni aggiuntive (sezione 5.2.1)
- payload più grande (sezione 5.2.2)
- meno interrupt alla CPU (sezione 5.2.3)

5.2.1 Trame Baby Giant

Le **trame Baby Giant** sono trame con una dimensione maggiore della dimensione massima di 1518 byte definita dalla specifica originale di Ethernet, a causa dell'inserimento di nuove intestazioni di livello data-link al fine di trasportare informazioni aggiuntive sulla trama:

- il tagging VLAN delle trame (IEEE 802.1Q) aggiunge 4 byte;[1];
- il tag stacking VLAN (IEEE 802.1ad) aggiunge 8 byte;[2]
- MPLS aggiunge 4 byte per ogni etichetta impilata.[3]

Nelle trame Baby Giant la dimensione massima del payload (ad es. pacchetto IP) è invariata ⇒ un router, quando riceve una trama Baby Giant, nel rigenerare il livello data-link può imbustare il payload in una trama Ethernet normale ⇒ non è un problema l'interconnessione di LAN con differenti dimensioni massime delle trame supportate.

Lo standard IEEE 802.3as (2006) propone di estendere la dimensione massima della trama a 2000 byte, mantenendo invariata la dimensione della MTU.

5.2.2 Jumbo Frame

Le **Jumbo Frame** sono trame con una dimensione maggiore della dimensione massima di 1518 byte definita dalla specifica originale di Ethernet:

- Mini Jumbo: trame con MTU di dimensione pari a 2500 byte;
- Jumbo (o "Giant" o "Giant Frame"): trame con MTU di dimensione fino a 9 KB;

a causa dell'imbustamento di payload più grandi al fine di:

- trasportare dati di archiviazione: tipicamente le unità elementari dei dati di archiviazione sono troppo grandi da trasportare in un'unica trama Ethernet:
  - il protocollo NFS per i NAS trasporta blocchi di dati da 8 KB circa;[4]
  - il protocollo FCoE per le SAN e il protocollo FCIP per l'interconnessione di SAN trasportano trame Fibre Channel da 2,5 KB circa;[5]
  - il protocollo iSCSI per le SAN trasporta comandi SCSI da 8 KB circa;[6]

[1] Si rimanda alla sezione 11.3.
[2] Si rimanda alla sezione 11.3.3.
[3] Si rimanda alla sezione *Intestazione MPLS* nel capitolo *MPLS* negli appunti di "Tecnologie e servizi di rete".
[4] Si rimanda alla sezione 14.3.
[5] Si rimanda alle sezioni 14.4.2 e 14.4.4.
[6] Si rimanda alla sezione 14.4.3.

- diminuire l'overhead delle intestazioni in termini di:
  - risparmio di byte: non è molto significativo, soprattutto considerando l'elevata larghezza di banda a disposizione nelle reti odierne;
  - capacità di elaborazione per i meccanismi TCP (numeri di sequenza, timer...): ogni pacchetto TCP scatena un interrupt alla CPU.

Se una LAN che usa Jumbo Frame viene connessa con una LAN che non usa Jumbo Frame, tutte le Jumbo Frame verranno frammentate a livello IP, ma la frammentazione IP non è conveniente dal punto di vista delle prestazioni ⇒ le Jumbo Frame vengono usate in reti indipendenti entro ambiti particolari.

5.2.3 TCP offloading

Le schede di rete con la funzione **TCP offloading** possono automaticamente condensare al volo più payload TCP in un solo pacchetto IP prima di passarlo al sistema operativo (i numeri di sequenza e altri parametri sono gestiti internamente dalla scheda di rete) ⇒ il sistema operativo, anziché dover elaborare più pacchetti piccoli scatenando tanti interrupt, vede un singolo pacchetto IP più grande e può farlo elaborare alla CPU in una volta sola ⇒ ciò diminuisce l'overhead dovuto ai meccanismi del TCP.

5.3 PoE

I bridge dotati della funzionalità **Power over Ethernet** (PoE) sono in grado di distribuire energia elettrica (fino a qualche decina di Watt) sui cavi Ethernet (solo doppini di rame), per connettere dispositivi con moderate esigenze di energia (telefoni VoIP, access point wi-fi, videocamere di sorveglianza, ecc.) evitando cavi aggiuntivi per l'energia elettrica.

Stazioni non PoE possono essere collegate a prese PoE.

5.3.1 Problemi

- consumo di energia: un bridge PoE consuma molta più energia elettrica (ad es. 48 porte a 25 W l'una consumano 1,2 kW) ed è più costoso di un bridge tradizionale;
- continuità del servizio: un guasto del bridge PoE o l'interruzione dell'energia elettrica fanno smettere di funzionare i telefoni, che sono invece un servizio importante in caso di emergenza ⇒ è necessario installare dei gruppi di continuità (UPS, dall'inglese "uninterruptible power supply") ma, invece di fornire energia elettrica solamente ai telefoni tradizionali, devono fornire energia elettrica a tutta l'infrastruttura dati.

Parte II

Albero ricoprente

Capitolo 6

Spanning Tree Protocol

6.1 Il problema dei cicli

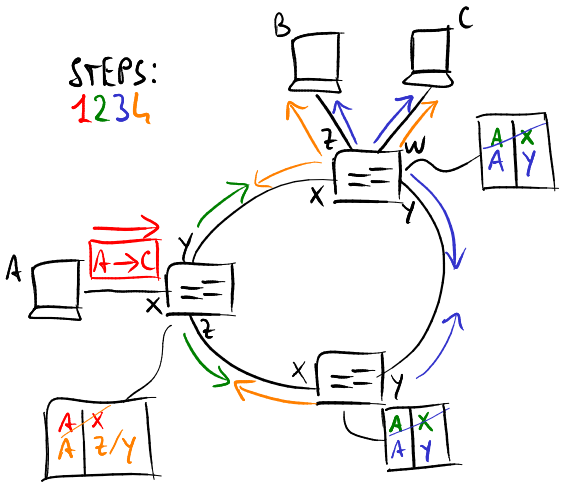


Figura 6.1: Esempio di invio di una trama unicast a una stazione non presente nel filtering database in una rete di livello data-link con un anello in topologia.

Se la rete presenta un anello logico in topologia, alcune trame possono iniziare a girare all'infinito in una moltiplicazione a catena all'interno del ciclo:

- trame broadcast/multicast: vengono sempre propagate su tutte le porte, provocando un **broadcast storm**;
- trame unicast inviate a una stazione ancora non presente nel filtering database o non esistente: vengono mandate in flooding.

Inoltre, i bridge nel ciclo possono avere filtering database inconsistenti, cioè la entry nel filtering database relativa alla stazione mittente cambia la porta ogni volta che arriva una replica della trama su una porta diversa, facendo credere al bridge che la trama arrivi dalla stazione stessa che si è spostata.

6.2 Algoritmo di spanning tree[1]

L'**algoritmo di spanning tree** consente di eliminare gli anelli logici dalla topologia fisica della rete, disattivando dei link[2] per trasformare una topologia magliata (grafo) in un albero chiamato **albero ricoprente**, la cui radice è uno dei bridge chiamato **root bridge**.

Ogni link è caratterizzato da un costo basato sulla velocità del link: dato un root bridge, possono essere costruiti più alberi ricoprenti che connettono tutti i bridge tra loro, ma l'algoritmo di spanning tree sceglie l'albero ricoprente composto dagli archi a costo minore.

Parametri

- **Bridge Identifier**: identifica il bridge in modo univoco e contiene:
  - priorità del bridge: può essere impostata liberamente (valore predefinito = 32768);
  - indirizzo MAC del bridge: è scelto tra gli indirizzi MAC delle sue porte da un algoritmo specifico del produttore e non può essere modificato;
- **Port Identifier**: identifica la porta del bridge e contiene:
  - priorità della porta: può essere impostata liberamente (valore predefinito = 128);
  - numero della porta: in teoria un bridge non può avere più di 256 porte ⇒ in pratica si può anche utilizzare il campo della priorità della porta se necessario;
- **Root Path Cost**: è pari alla somma dei costi dei link, selezionati dall'algoritmo di spanning tree, attraversati per raggiungere il root bridge (il costo di attraversamento di un bridge è nullo).

6.2.1 Criteri

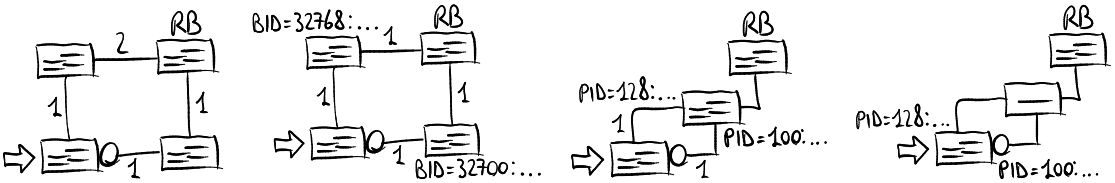
L'albero ricoprente si può determinare con i seguenti criteri.

Root bridge

Un **root bridge** è la radice dell'albero ricoprente: tutte le trame che vanno da uno dei suoi sottoalberi a un altro devono attraversare il root bridge.[3]

Si seleziona come root bridge il bridge con il Bridge Identifier più piccolo: la radice dell'albero ricoprente sarà quindi il bridge con la priorità più bassa, o a parità di priorità quello con l'indirizzo MAC più basso.

Porta root



(a) Percorso a costo minimo. *(b) Bridge ID remoto più piccolo.* *(c) Port ID remoto più piccolo.* *(d) Port ID locale più piccolo.*

Figura 6.2: Criteri per la selezione di una porta root per il bridge indicato dalla freccia.

[1] Questa sezione contiene contenuti CC BY-SA dalla voce Spanning Tree Protocol su Wikipedia in inglese.

[2] In realtà l'algoritmo di spanning tree blocca le porte, non i link (si rimanda alla sezione 6.4.1).

[3] Si prega di fare attenzione al fatto che il traffico che si sposta all'interno dello stesso sottoalbero non attraversa il root bridge.

Una **porta root** è la porta responsabile della connessione al root bridge: invia le trame verso il root bridge e riceve le trame dal root bridge.

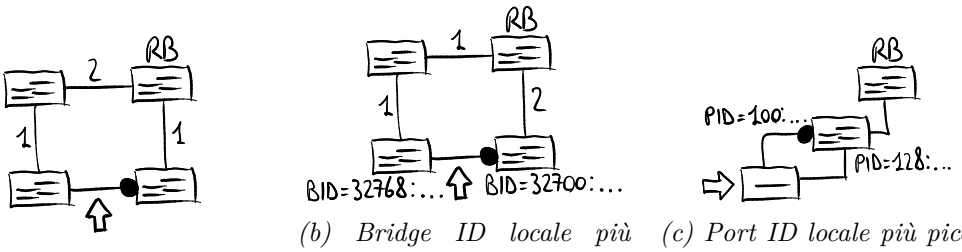
1. Si determina il costo di ogni percorso possibile dal bridge alla radice. Da questi, si sceglie quello con il costo più piccolo (un percorso a costo minimo). La porta connessa a quel percorso è quindi la porta root del bridge.

2. Quando più percorsi a partire da un bridge sono percorsi a costo minimo verso la radice, il bridge utilizza il bridge vicino con il Bridge Identifier più piccolo per inoltrare le trame alla radice. La porta root è così quella connessa al bridge con il Bridge Identifier più basso.

3. Quando due bridge sono connessi con più cavi, più porte su un singolo bridge sono candidate per la porta root. In questo caso, si utilizza il percorso che passa attraverso la porta sul bridge vicino che ha il Port Identifier più piccolo.

4. In una particolare configurazione con un hub dove i Port Identifier remoti sono uguali, si utilizza il percorso che passa attraverso la porta sul bridge stesso che ha il Port Identifier più piccolo.

Porta designata



(a) Percorso a costo minimo. *(b) Bridge ID locale più piccolo.* *(c) Port ID locale più piccolo.*

Figura 6.3: Criteri per la selezione di una porta designata per il link indicato dalla freccia.

Una **porta designata** è la porta responsabile di servire il link: invia le trame alle foglie e riceve le trame dalle foglie.

1. Si determina il costo di ogni percorso possibile da ogni bridge connesso al link alla radice. Da questi, si sceglie quello con il costo più piccolo (un percorso a costo minimo). La porta connessa al link del bridge che conduce a quel percorso è quindi la porta designata del link.

2. Quando più bridge su un link conducono a un percorso a costo minimo verso la radice, il link utilizza il bridge con il Bridge Identifier più piccolo per inoltrare le trame alla radice. La porta che connette quel bridge al link è la porta designata per il link.

3. Quando un bridge è connesso a un link con più cavi, più porte su un singolo bridge sono candidate per la porta designata. In questo caso, si utilizza il percorso che passa attraverso la porta sul bridge stesso che ha il Port Identifier più piccolo.

Porta bloccata

Una **porta bloccata** non invia mai le trame sul link e scarta tutte le trame ricevute (eccetto le Configuration BDPU).

Qualsiasi porta attiva che non sia né una porta root né una porta designata è una porta bloccata.

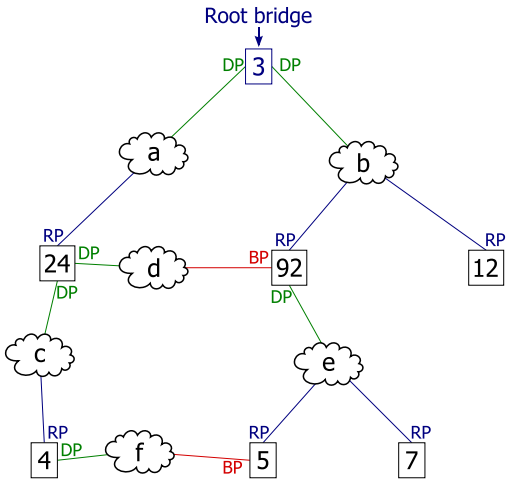


Figura 6.4: Questo schema illustra tutti gli stati delle porte calcolati dall'algoritmo di spanning tree per una rete di esempio.[4]

6.3 Messaggi BPDU

I criteri qui sopra descrivono un modo di determinare quale albero ricoprente sarà calcolato dall'algoritmo, ma le regole così come scritte richiedono la conoscenza dell'intera rete. I bridge devono determinare il root bridge e calcolare i ruoli delle porte (root, designata o bloccata) con le sole informazioni di cui dispongono.

Affinché i bridge possano scambiarsi informazioni sui Bridge Identifier e sui Root Path Cost, lo **Spanning Tree Protocol** (STP), standardizzato come IEEE 802.1D (1990), definisce dei messaggi chiamati **Bridge Protocol Data Unit** (BPDU).

6.3.1 Formato delle BPDU

Le BPDU hanno il seguente formato:

1	7	8	16	24	32
Protocol ID (0)				Version (0)	BPDU Type (0)
TC	000000	TCA	Root Priority		
Root MAC Address					
Root Path Cost					
Bridge Priority					
Bridge MAC Address					
Port Priority				Port Number	Message Age
Max Age					Hello Time
Forward Delay					

Tabella 6.1: Formato della Configuration BPDU (35 byte) nello STP.

16	24	32
Protocol ID (0)	Version (0)	BPDU Type (0x80)

Tabella 6.2: Formato della Topology Change Notification BPDU (4 byte).

[4]Questa immagine è tratta da Wikimedia Commons (Spanning tree protocol at work 5.svg), è stata realizzata da Nancy Griffeth e dall'utente GhosT ed è concessa sotto la licenza Creative Commons Attribuzione 3.0 Unported.

dove i campi sono:

- campo Protocol Identifier (2 byte): specifica il protocollo IEEE 802.1D (valore 0);
- campo Version (1 byte): distingue lo Spanning Tree Protocol (valore 0) dal Rapid Spanning Tree Protocol (valore 2) (si rimanda al capitolo 7);
- campo BPDU Type (1 byte): specifica il tipo di BPDU:
  - **Configuration BPDU** (CBPDU) (valore 0): utilizzata per il calcolo dell'albero ricoprente, cioè per determinare il root bridge e gli stati delle porte (si rimanda alla sezione 6.4.2);
  - **Topology Change Notification BPDU** (TCN BPDU) (valore 0x80): utilizzata per annunciare i cambiamenti nella topologia della rete al fine di aggiornare le entry nei filtering database (si rimanda alla sezione 6.5.2);
- flag Topology Change (TC) (1 bit): impostato dal root bridge per informare tutti i bridge che è avvenuto un cambiamento nella rete;
- flag Topology Change Acknowledgement (TCA) (1 bit): impostato dal root bridge per informare il bridge che ha rilevato il cambiamento che la sua Topology Change Notification BPDU è stata ricevuta;
- campo Root Identifier (8 byte): specifica il Bridge Identifier del root bridge della rete:
  - campo Root Priority (2 byte): contiene la priorità del root bridge;
  - campo Root MAC Address (6 byte): contiene l'indirizzo MAC del root bridge;
- campo Bridge Identifier (8 byte): specifica il Bridge Identifier del bridge che sta propagando la Configuration BPDU:
  - campo Bridge Priority (2 byte): contiene la priorità del bridge;
  - campo Bridge MAC Address (6 byte): contiene l'indirizzo MAC del bridge;
- campo Root Path Cost (4 byte): contiene il costo del percorso per raggiungere il root bridge, visto dal bridge che sta propagando la Configuration BPDU;
- campo Port Identifier (2 byte): specifica il Port Identifier della porta su cui il bridge sta propagando la Configuration BPDU:
  - campo Port Priority (1 byte): contiene la priorità della porta;
  - campo Port Number (1 byte): contiene il numero della porta;
- campo Message Age (2 byte): valore, inizializzato a 0, che ogni volta che la Configuration BPDU attraversa un bridge viene incrementato per il tempo di transito attraverso il bridge;[5]
- campo Max Age (2 byte, valore predefinito = 20 s): se il Message Age raggiunge il valore Max Age, la Configuration BPDU ricevuta non è più valida;[5]
- campo Hello Time (2 byte, valore predefinito = 2 s): specifica ogni quanto tempo il root bridge genera la Configuration BPDU;[5]
- campo Forward Delay (2 byte, valore predefinito = 15 s): specifica il tempo di attesa prima di forzare la transizione di una porta ad un altro stato.[5]

[5] I campi temporali sono espressi in unità di 256^{esimi} di secondo (circa 4 ms).

6.3.2 Generazione e propagazione delle BPDU

Solo il root bridge può generare Configuration BPDU: tutti gli altri bridge si limitano a propagare le Configuration BPDU ricevute su tutte le loro porte designate. Le porte root sono quelle che ricevono le Configuration BPDU migliori, cioè con valore Message Age più basso = Root Path Cost più basso. Le porte bloccate non inviano mai le Configuration BPDU ma rimangono in ascolto delle Configuration BPDU in arrivo.

Invece le Topology Change Notification BPDU possono essere generate da qualsiasi bridge non root, e vengono propagate sempre sulle porte root.

Quando un bridge genera/propaga una trama BPDU, utilizza l'indirizzo MAC univoco della porta stessa come indirizzo sorgente, e l'indirizzo multicast 01:80:C2:00:00:00 dello STP come indirizzo di destinazione:

6 byte	6 byte	2 byte	1 byte	1 byte	1 byte		4 byte
01:80:C2:00:00:00 (multicast)	indirizzo bridge sorgente (unicast)	...	0x42	0x42	0x03	BPDU	...
indirizzo MAC di destinazione	indirizzo MAC sorgente	length	DSAP	SSAP	CTRL	payload	FCS

6.4 Comportamento dinamico

6.4.1 Stati delle porte

Disabled Una porta spenta perché nessun link è connesso alla porta.

Blocking Una porta che provocherebbe un ciclo se fosse attiva. Nessuna trama è inviata o ricevuta su una porta in stato di blocking (le Configuration BPDU vengono comunque ricevute in stato di blocking), ma essa potrebbe andare in stato di forwarding se gli altri link in uso si guastano e l'algoritmo di spanning tree determina che la porta può passare allo stato di forwarding.

Listening Il bridge elabora le Configuration BPDU e attende eventuali nuove informazioni che farebbero ritornare la porta allo stato di blocking. Non popola il filtering database e non inoltra trame.

Learning Mentre la porta non inoltra ancora le trame, il bridge apprende gli indirizzi sorgente dalle trame ricevute e li aggiunge al filtering database. Popola il filtering database, ma non invia le trame.

Forwarding Una porta che sta ricevendo e inviando dei dati. Lo STP continua comunque a monitorare le Configuration BPDU in arrivo, così la porta potrebbe ritornare allo stato di blocking per impedire un ciclo.

stato porta	**ruolo porta**	riceve trame?	riceve e elabora CBPDU?	genera o propaga CBPDU?	aggiorna filtering database?	inoltra trame?	genera o propaga TCN BPDU?
disabled	bloccata	no	no	no	no	no	no
blocking		sì	sì	no	no	no	no
listening	(in transizione)	sì	sì	sì	no	no	no
learning	designata	sì	sì	sì	sì	no	no
	root	sì	sì	no	sì	no	sì
forwarding	designata	sì	sì	sì	sì	sì	no
	root	sì	sì	no	sì	sì	sì

Tabella 6.3: Ruoli e stati delle porte nello STP.

6.4.2 Ingresso di un nuovo bridge

Quando un nuovo bridge viene connesso a una rete di livello data-link, assumendo che abbia un Bridge Identifier più alto di quello del root bridge corrente della rete:

1. all'inizio il bridge, senza sapere (ancora) nulla sul resto della rete, assume di essere il root bridge: imposta tutte le sue porte come designate (stato di listening) e inizia a generare delle Configuration BPDU su di esse, dicendo di essere il root bridge;

2. gli altri bridge ricevono le Configuration BPDU generate dal nuovo bridge e confrontano il Bridge Identifier del nuovo bridge con quello del root bridge corrente della rete, quindi le scartano;

3. periodicamente il root bridge della rete genera delle Configuration BPDU, che gli altri bridge ricevono dalle loro porte root e propagano attraverso le loro porte designate;

4. quando il nuovo bridge riceve una Configuration BPDU dal root bridge della rete, apprende di non essere il root bridge perché esiste un altro bridge avente un Bridge Identifier più basso del suo, quindi smette di generare Configuration BPDU e imposta la porta da cui ha ricevuto la Configuration BPDU del root bridge come porta root;

5. anche il nuovo bridge inizia a propagare le Configuration BPDU, stavolta relative al root bridge della rete, su tutte le sue altre porte (designate), mentre continua a ricevere le Configuration BPDU propagate dagli altri bridge;

6. quando il nuovo bridge riceve su una porta designata una Configuration BPDU "migliore", in base ai criteri per la selezione della porta designata, rispetto alla Configuration BPDU che sta propagando su tale porta, quest'ultima smette di propagare le Configuration BPDU e diventa bloccata (stato di blocking);

7. dopo un tempo pari al Forward Delay, le porte rimaste designate e la porta root passano dallo stato di listening a quello di learning: il bridge inizia a popolare il filtering database, per evitare che il bridge cominci subito a mandare le trame in flooding sovraccaricando la rete;

8. dopo un tempo pari al Forward Delay, le porte designate e la porta root passano dallo stato di learning a quello di forwarding: il bridge può propagare anche le trame normali su tali porte.

6.5 Cambiamenti nella topologia della rete

6.5.1 Ricalcolo dell'albero ricoprente

Quando si verifica un cambiamento di topologia, lo STP è in grado di rilevare il cambiamento di topologia, grazie alla generazione periodica delle Configuration BPDU da parte del root bridge, e di continuare a garantire l'assenza di anelli in topologia, ricalcolando se necessario l'albero ricoprente, vale a dire il root bridge e gli stati delle porte.

Guasto di un link

Quando un link (appartenente all'albero ricoprente corrente) si guasta:

1. le Configuration BPDU che il root bridge genera non riescono più a raggiungere l'altra porzione della rete: in particolare, la porta designata del link guasto non invia più Configuration BPDU;

2. l'ultima Configuration BPDU ascoltata dalla porta bloccata al di là del link "invecchia" nel bridge stesso, cioè il suo Message Age viene incrementato con il passare del tempo;

3. quando il Message Age raggiunge il valore di Max Age, l'ultima Configuration BPDU ascoltata scade e il bridge riparte da capo rieleggendosi root bridge: reimposta tutte le porte come designate, e inizia a generare Configuration BPDU dicendo di essere il root bridge;

4. lo STP continua in un modo analogo al caso precedentemente trattato relativo all'ingresso di un nuovo bridge:

   - se esiste un link non appartenente all'albero ricoprente che collega le due porzioni di rete, la porta bloccata connessa a quel link alla fine diventerà porta root in stato di forwarding, e il link entrerà a far parte dell'albero ricoprente;
   - altrimenti, se le due porzioni di rete non possono più essere connesse tra loro, in ogni porzione di rete verrà eletto un root bridge.

Inserimento di un nuovo link

Quando viene inserito un nuovo link, le porte a cui il nuovo link è connesso diventano designate in stato di listening, e iniziano a propagare le Configuration BPDU generate dal root bridge della rete ⇒ arrivano delle nuove Configuration BPDU attraverso il nuovo link:

- se il link ha un costo sufficientemente basso, il bridge connesso al link inizia a ricevere da una porta non root delle Configuration BPDU aventi un Root Path Cost minore rispetto a quello delle Configuration BPDU ricevute dalla porta root ⇒ la porta root viene aggiornata in modo che il root bridge sia raggiungibile attraverso il percorso migliore (in base ai criteri per la selezione della porta root), così come le porte designate e bloccate sono eventualmente aggiornate di conseguenza;

- se il link ha un costo troppo alto, le Configuration BPDU che lo attraversano hanno un Root Path Cost troppo alto ⇒ una delle due porte connesse al nuovo link diventa bloccata e l'altra rimane designata (in base ai criteri per la selezione della porta designata).

6.5.2 Annuncio di cambiamenti di topologia

Quando in seguito a un cambiamento di topologia lo STP modifica l'albero ricoprente cambiando gli stati delle porte, esso non modifica le entry nei filtering database dei bridge per riflettere i cambiamenti ⇒ le entry possono non essere aggiornate: ad esempio, le trame verso una certa destinazione potrebbero continuare ad essere mandate su una porta diventata bloccata, fino a quando la entry relativa a quella destinazione non scade perché il suo ageing time è andato a 0 (nel caso peggiore: 5 minuti!).

Lo STP prevede un meccanismo per velocizzare la convergenza della rete per quanto riguarda il filtering database quando è rilevato un cambiamento di topologia:

1. il bridge che ha rilevato il cambiamento di topologia genera una Topology Change Notification BPDU attraverso la sua porta root verso il root bridge per annunciare il cambiamento di topologia;[6]

2. i bridge attraversati inoltrano immediatamente la Topology Change Notification BPDU attraverso le loro porte root;

3. il root bridge genera in risposta una Configuration BPDU con i flag Topology Change e Topology Change Acknowledgement impostati a 1, che dopo essere stata inoltrata indietro dai bridge attraversati verrà ricevuta dal bridge che ha rilevato il cambiamento di topologia;[7]

[6] Il bridge continua a generare la Topology Change Notification BPDU ogni Hello Time, finché non riceve l'acknowledge.

[7] Il root bridge continua a generare in risposta la Configuration BPDU di acknowledgement per Max Age + Forward Delay.

4. il root bridge genera su tutte le sue porte designate una Configuration BPDU con il flag Topology Change impostato;

5. ogni bridge, quando riceve la Configuration BPDU:

   (a) scarta tutte le entry nel suo filtering database aventi ageing time inferiori al Forward Delay;

   (b) propaga a sua volta la Configuration BPDU su tutte le sue porte designate (mantenendo il flag Topology Change impostato);

6. la rete funziona temporaneamente in una condizione sub-ottimale perché vengono mandate più trame in flooding, fino a quando i bridge non popolano di nuovo i loro filtering database con i nuovi percorsi tramite gli algoritmi di apprendimento[8].

6.6 Problemi

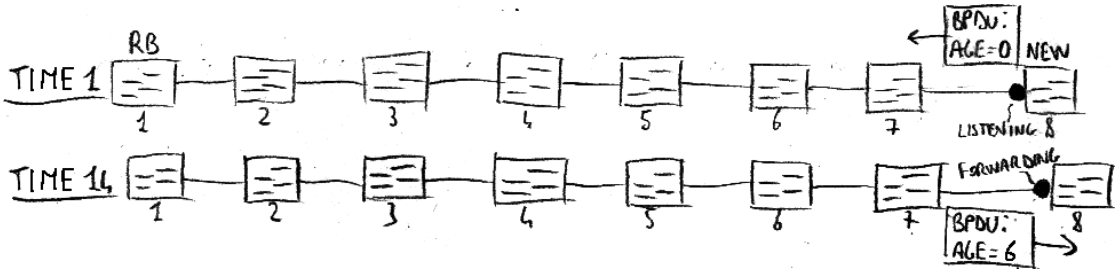
6.6.1 Prestazioni

La filosofia dello STP è "nega sempre, consenti solo quando sicuro": quando avviene un cambiamento di topologia, le trame non vengono inoltrate finché non si è sicuri che il transitorio è esaurito, cioè non ci sono cicli e la rete è in uno stato coerente, anche introducendo dei tempi di attesa lunghi a scapito della rapidità di convergenza e della raggiungibilità di alcune stazioni.

Supponendo di rispettare le tempistiche consigliate dallo standard, cioè:

- siano adottati i valori temporali consigliati dallo standard: Max Age = 20 s, Hello Time = 2 s, Forward Delay = 15 s;
- il tempo di transito attraverso ogni bridge da parte di una BPDU non superi il TransitDelay = HelloTime ÷ 2 = 1 s;

non si possono collegare più di 7 bridge in cascata tra due sistemi terminali affinché una Configuration BPDU possa attraversare l'intera rete due volte entro il Forward Delay: se venisse messo un ottavo bridge in cascata, infatti, nel caso peggiore le porte del nuovo bridge, autoelettosi root bridge, passerebbero dallo stato di listening a quello di forwarding[9] prima che la Configuration BPDU proveniente dal root bridge all'altra estremità della rete giunga in tempo al nuovo bridge:[10]



Con l'introduzione dello stato di learning, dopo il guasto di un link la rete impiega approssimativamente 50 secondi per convergere a uno stato coerente:

- 20 s (Max Age): richiesti affinché l'ultima Configuration BPDU ascoltata scada e il guasto sia rilevato;

[8] Si veda la sezione 3.2.2.

[9] Quando questo vincolo fu stabilito, lo stato di learning non era ancora stato introdotto e le porte passavano direttamente dallo stato di listening a quello di forwarding.

[10] Esattamente il valore minimo per il Forward Delay sarebbe pari a 14 s, ma è stata prevista una tolleranza di 1 s.

- 15 s (Forward Delay): richiesti per la transizione della porta dallo stato di listening a quello di learning;
- 15 s (Forward Delay): richiesti per la transizione della porta dallo stato di learning a quello di forwarding.

Inoltre, il raggiungimento di uno stato coerente nella rete non implica necessariamente la fine del disservizio sperimento dall'utente: infatti il guasto si può ripercuotere anche a livello applicativo, molto sensibile alle perdite di connettività oltre a una certa soglia:

- i sistemi per la gestione delle basi di dati potrebbero iniziare delle lunghe procedure di recupero dai guasti;
- le applicazioni multimediali di rete che generano del traffico anelastico (come le applicazioni VoIP) soffrono molto delle variazioni di ritardo.

Si potrebbe provare a personalizzare i valori relativi ai parametri temporali per aumentare la rapidità di convergenza ed ampliare il diametro di bridge massimo, ma quest'operazione è sconsigliata:

- se non si fa attenzione si rischia di ridurre la reattività della rete ai cambiamenti di topologia e di intaccare le funzionalità della rete;
- a prima vista sembra sufficiente operare solo sul root bridge poiché questi valori vengono tutti propagati dal root bridge all'intera rete, ma in realtà se il root bridge cambia il nuovo root bridge deve annunciare gli stessi valori ⇒ è necessario di fatto aggiornare questi parametri su tutti i bridge.

Spesso lo STP viene disabilitato sulle porte edge, cioè le porte collegate direttamente agli host finali, per alleviare i disservizi sperimentati dall'utente:

- a causa dei ritardi di transizione delle porte, un PC che si connette alla rete rimarrebbe inizialmente isolato per un tempo pari a due Forward Delay;
- la connessione di un PC rappresenta un cambiamento di topologia ⇒ la pulizia delle entry vecchie scatenata dall'annuncio del cambiamento di topologia aumenterebbe considerevolmente il numero di trame inviate in flooding nella rete.

Alle porte edge tuttavia vanno collegati esclusivamente gli host, altrimenti si potrebbero creare dei cicli nella rete ⇒ alcuni fornitori non permettono ciò: ad esempio, il meccanismo PortFast proprietario di Cisco raggiunge lo stesso obiettivo senza disabilitare lo STP sulle porte edge, essendo in grado di farle passare immediatamente allo stato di forwarding e di rilevare eventuali cicli su di esse (ovvero due porte edge collegate tra loro da un filo diretto).

6.6.2 Scalabilità

Dato un root bridge, possono essere costruiti più alberi ricoprenti che connettono tutti i bridge tra loro, ma l'algoritmo di spanning tree sceglie l'albero ricoprente composto dagli archi a costo minore. In questo modo, i percorsi sono ottimizzati solo rispetto alla radice dell'albero:

- i link disabilitati sono completamente inutilizzati, ma qualcuno comunque deve pagare per tenerli attivi come link secondari per la tolleranza ai guasti;
- non è possibile il bilanciamento del carico per distribuire il traffico su più link paralleli ⇒ i link appartenenti all'albero ricoprente selezionato devono sopportare anche il carico del traffico che, se non ci fosse lo STP, prenderebbe un percorso più breve passando per i link disabilitati:

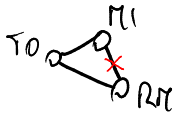


Figura 6.5: Lo STP non è adatto a operare su scala geografica.

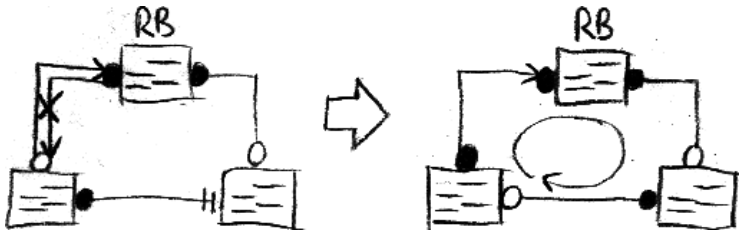
Una rete IP è invece in grado di organizzare meglio il traffico: l'albero ricoprente non è unico nell'intera rete, ma ogni sorgente può calcolare il proprio albero e inviare il traffico sul percorso più breve garantendo un maggiore bilanciamento del carico dei link.

Le Virtual LAN (VLAN) risolvono questo problema instaurando degli alberi ricoprenti multipli (si rimanda al capitolo 11).

6.6.3 Link unidirezionali

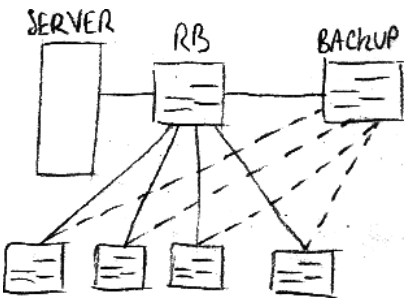
Lo STP ipotizza che ciascun link sia bidirezionale: se un link si guasta, non è più possibile inviare trame in nessuna delle due direzioni. In realtà i cavi in fibra ottica sono unidirezionali ⇒ per collegare due nodi sono necessari due cavi in fibra ottica, uno per la comunicazione in una direzione e l'altro per la comunicazione nella direzione opposta, e un guasto su uno dei due cavi interrompe solo il traffico in una direzione.

Se uno dei due link unidirezionali si guasta, potrebbe nascere un ciclo sull'altro link nonostante lo STP: le Configuration BPDU sono propagate in modo unidirezionale dalla radice alle foglie ⇒ se il percorso di propagazione diretto si rompe, il bridge all'altra estremità non riceve più le Configuration BPDU del root bridge da quel link, quindi sposta la porta root su un altro link e, assumendo che non ci sia nessuno su quel link, imposta la porta come designata creando un ciclo:

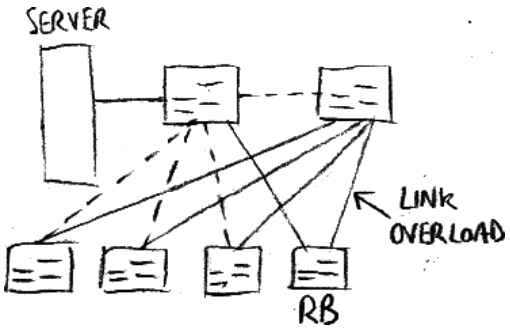


Unidirectional Link Detection (UDLD) è un protocollo proprietario di Cisco in grado di rilevare la presenza di un guasto su un link unidirezionale grazie a una sorta di "ping", e di disabilitare la porta (stato "error disabled") invece di eleggerla come designata.

6.6.4 Posizionamento del root bridge



(a) Configurazione ottimale.



(b) Configurazione da evitare.

La posizione del root bridge ha un pesante impatto nella rete:

- il traffico da una parte all'altra della rete deve passare per il root bridge ⇒ le prestazioni, in termini di throughput aggregato e larghezza di banda delle porte, del bridge selezionato come root bridge devono essere sufficienti per sopportare una grande quantità di traffico;
- è preferibile una topologia a stella, dove il root bridge è il centro stella ⇒ ciascun link collega un solo bridge al root bridge:
  - bilanciamento del carico dei link più equilibrato: il link non deve sostenere il traffico proveniente da altri bridge;
  - maggiore tolleranza ai guasti: un guasto del link interessa solo la connettività di un bridge;
- i server e i datacenter vanno posizionati vicino al root bridge al fine di ridurre la latenza della comunicazione dati;

⇒ è necessario personalizzare a un valore molto basso la priorità del bridge che deve essere il root bridge, in modo da non rischiare che un altro bridge sia eletto root bridge.

È importante anche la posizione del bridge di backup, pensato per entrare in gioco in caso di guasto di un link primario o del root bridge primario:

- guasto di un link primario: la configurazione ottimale è una topologia a stella ridondante composta da link secondari, in cui ogni bridge è collegato al bridge di backup da un link secondario;
- guasto del root bridge primario: è necessario personalizzare anche la priorità del root bridge di backup a un valore appena più alto della priorità del root bridge primario, in modo da forzare quel bridge a essere eletto root bridge in caso di guasto.

6.6.5 Sicurezza

Lo STP non ha dei meccanismi di sicurezza incorporati contro attacchi dall'esterno.

Elezione del bridge dell'utente come root bridge Un utente può collegare alla rete un bridge con una priorità molto bassa forzandolo a diventare il nuovo root bridge e cambiando l'albero ricoprente della rete. La funzione **BPDU Guard** proprietaria di Cisco permette alle porte edge di propagare le sole Configuration BPDU provenienti dall'interno della rete, rifiutando quelle ricevute dall'esterno (la porta va in stato "error disabled").

Rate limit su broadcast storm Quasi tutti i bridge professionali hanno qualche forma di controllo dei broadcast storm in grado di limitare la quantità di traffico broadcast sulle porte scartando il traffico in eccesso oltre a una certa soglia, ma questi misuratori di traffico non riescono a distinguere tra le trame di un broadcast storm e trame broadcast inviate dalle stazioni ⇒ rischiano di filtrare del traffico broadcast legittimo, e un broadcast storm è più difficile da individuare.

Collegamento di bridge senza STP Un singolo bridge privo di STP o con STP disabilitato può iniziare a pompare traffico broadcast nella rete se collegato in modo che si formi un ciclo fuori dal controllo dello STP: sono esempi il collegamento di una porta del bridge direttamente a un'altra dello stesso bridge, o il collegamento del bridge dell'utente a due bridge interni della rete con due link ridondanti.

Domini di STP multipli Talvolta si vuole collegare due diversi domini di STP, ognuno con il proprio albero ricoprente, a uno stesso canale condiviso (ad es. due provider con domini di STP diversi nello stesso datacenter). La funzione **BPDU Filter** proprietaria di Cisco disabilita l'invio e la ricezione di Configuration BPDU sulle porte alle periferie dei domini, per tenere divisi gli alberi ricoprenti.

Capitolo 7

Rapid Spanning Tree Protocol

Il **Rapid Spanning Tree Protocol** (RSTP), standardizzato come IEEE 802.1w (2001), è caratterizzato da una maggiore rapidità di convergenza rispetto allo STP in termini di:

- ricalcolo dell'albero ricoprente (sezione 7.3.1)
- aggiornamento dei filtering database (sezione 7.3.2)

7.1 Ruoli e stati delle porte

Il RSTP definisce nuovi ruoli e stati delle porte:

- stato **discarding**: la porta non inoltra trame e scarta quelle ricevute (eccetto le Configuration BPDU), unificando gli stati di disabled, blocking e listening;
- ruolo **alternate**: la porta, in stato di discarding, è collegata allo stesso link di una porta designata di un altro bridge, rappresentando un rimpiazzamento rapido per la porta root;
- ruolo **backup**: la porta, in stato di discarding, è collegata allo stesso link di una porta designata dello stesso bridge, rappresentando un rimpiazzamento rapido per la porta designata;
- ruolo **edge**: alla porta si possono collegare solo host, mirando a ridurre, rispetto allo STP classico, i disservizi sperimentati dagli utenti nel collegare i propri host alla rete.

stato porta	**ruolo porta**	riceve trame?	riceve e elabora CBPDU?	genera e propaga CBPDU?	aggiorna filtering database?	inoltra trame?
discarding	alternate	sì	sì	no	no	no
	backup	sì	sì	no	no	no
	designata[a]	sì	sì	sì	no	no
learning	designata	sì	sì	sì	sì	no
	root	sì	sì	no	sì	no
forwarding	designata	sì	sì	sì	sì	sì
	root	sì	sì	no	sì	sì
	edge	sì	sì	no	sì	sì

Tabella 7.1: Ruoli e stati delle porte nel RSTP.

[a] Una porta designata è in stato di discarding durante la sequenza proposal/agreement (si veda la sezione 7.3.1).

7.2 Formato della Configuration BPDU

La Configuration BPDU ha il seguente formato:

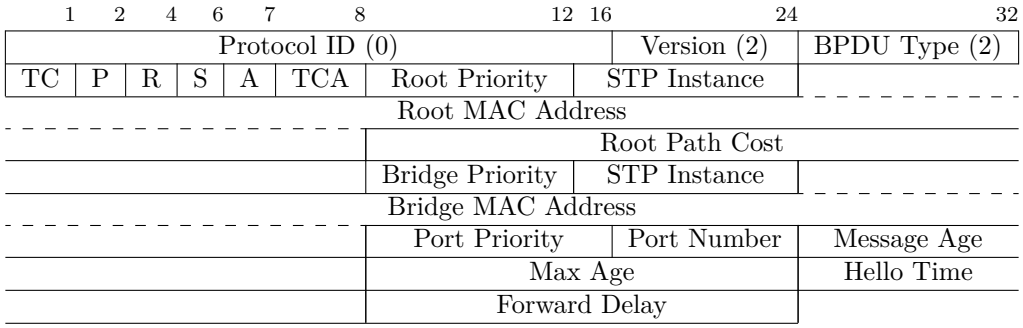


Tabella 7.2: Formato della Configuration BPDU (35 byte) nel RSTP.

dove ci sono alcuni cambiamenti rispetto alle BPDU dello STP classico[1]:

- campo Version (1 byte): identifica il RSTP come numero di versione 2 (nello STP era 0);
- campo BPDU Type (1 byte): identifica la Configuration BPDU sempre come tipo 2 (nello STP era 0), dato che non esistono più le Topology Change Notification BPDU;[2]
- 6 nuovi flag: controllano il meccanismo proposal/agreement (si rimanda alla sezione 7.3.1):
  - flag Proposal (P) e Agreement (A) (1 bit ciascuno): specificano se il ruolo della porta è stato proposto da un bridge (P = 1) o accettato dall'altro bridge (A = 1);
  - 2 flag nel campo Role (R) (2 bit): codificano il ruolo della porta proposto o accettato (00 = sconosciuto, 01 = alternate/backup, 10 = root, 11 = designata);
  - 2 flag nel campo State (S) (2 bit): specificano se la porta per cui è stato proposto o accettato il ruolo si trova nello stato di learning (10) o di forwarding (01);
- campi Root Identifier e Bridge Identifier (8 byte ciascuno): il RSTP include le specifiche tecniche di **IEEE 802.1t** (2001) che cambiano il formato del Bridge Identifier:
  - campo Bridge Priority (4 bit, valore predefinito = 8);
  - campo STP Instance (12 bit, valore predefinito = 0): utilizzato nelle Virtual LAN per abilitare più istanze di protocollo entro la stessa rete fisica (si rimanda alla sezione 11.4);
  - campo Bridge MAC Address (6 byte): invariato rispetto a IEEE 802.1D-1998;
- campo Root Path Cost (4 byte): il RSTP include le specifiche tecniche di IEEE 802.1t (2001) che cambiano i valori consigliati per il Port Path Cost includendo nuove velocità delle porte (fino a 10 Tb/s);
- campi Max Age e Forward Delay (2 byte ciascuno): sono completamente inutilizzati nel RSTP, ma sono stati mantenuti per ragioni di compatibilità.

[1] Si veda la sezione 6.3.1.

[2] D'ora in poi si farà riferimento alle Configuration BPDU semplicemente con BPDU.

7.3 Cambiamenti nella topologia della rete

7.3.1 Ricalcolo dell'albero ricoprente

Il RSTP è caratterizzato da una maggiore rapidità di convergenza della topologia rispetto allo STP classico: infatti si passa da 50 secondi a meno di 1 secondo (ordine dei 10 ms circa) se, com'era ormai la norma quando il RSTP fu standardizzato, in presenza di soli link punto-punto full duplex (quindi senza hub).

Rilevamento del guasto di un link

Quando si verifica il guasto di un link, il suo rilevamento da parte del RSTP è più rapido rispetto allo STP classico grazie a una gestione più efficiente delle BPDU.

I bridge non root non si limitano a propagare le BPDU generate dal root bridge: ogni bridge genera a ogni Hello Time (predefinito: 2 s) una BPDU, con il root bridge corrente come Root Identifier, anche se non ha ricevuto la BPDU dal root bridge. Se non sono state ricevute BPDU da 3 periodi di Hello Time, la BPDU corrente è dichiarata obsoleta e si assume che un guasto si è verificato sul link a cui la porta root è connessa.

Questo invecchiamento delle informazioni più rapido è inutile sulle reti moderne:

- nelle reti più vecchie con gli hub, un bridge non può rilevare a livello fisico un guasto tra l'hub e un altro bridge ⇒ l'unico modo per rilevarlo è accorgersi che i messaggi BPDU "keep-alive" hanno smesso di essere ricevuti;
- nelle reti più recenti che sono commutate pure, un bridge può rilevare immediatamente il guasto di un link a livello fisico senza aspettare per 3 periodi di Hello Time.

Una volta che un bridge ha rilevato il guasto di un link, inizia a generare le proprie BPDU ⇒ ogni bridge vicino, non appena riceve sulla sua porta root una BPDU dal bridge che sta pretendendo di essere il root bridge, invece di scartarla perché è peggiore di quella corrente, accetta la nuova BPDU dimenticando quella memorizzata in precedenza, perché significa che qualcosa è andato storto sul suo percorso verso il root bridge. A questo punto:

- se il suo Bridge Identifier è peggiore di quello nella BPDU, il bridge inizia a generare delle BPDU sulle sue porte designate con il nuovo Root Identifier;
- se il suo Bridge Identifier è migliore di quello nella BPDU, il bridge inizia a generare le proprie BPDU pretendendo di essere il root bridge.

Recupero dal guasto di un link

Una volta rilevato un guasto, alcune porte possono passare direttamente allo stato di forwarding senza transitare per lo stato di learning.

Porte alternate In caso di guasto della porta root, la porta alternate fornisce un percorso alternativo tra il bridge e il root bridge:

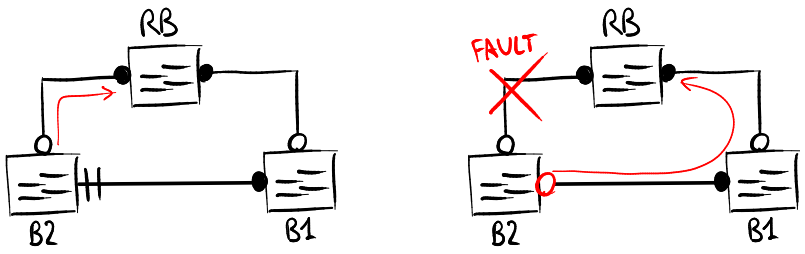


Figura 7.1: La porta alternate rappresenta un rimpiazzamento rapido per la porta root.

Porte di backup In caso di guasto della porta designata, la porta di backup fornisce un percorso alternativo tra il bridge e il link:

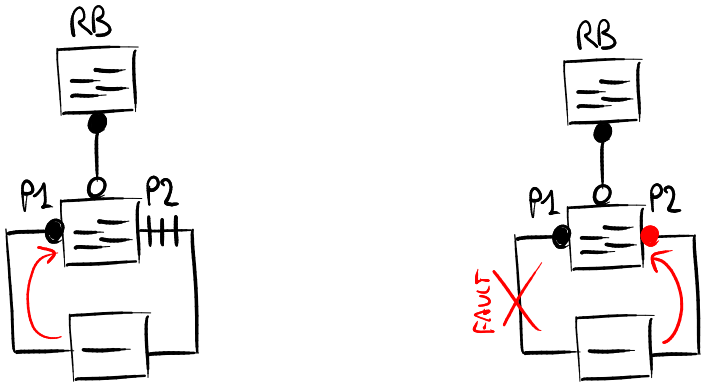


Figura 7.2: La porta di backup rappresenta un rimpiazzamento rapido per la porta designata.

Inserimento di un nuovo link

La **sequenza proposal/agreement** è un algoritmo per la sincronizzazione rapida sul ruolo delle porte tra due bridge.

Quando si inserisce un nuovo link tra due bridge:

1. ciascuno dei due bridge mette in stato di discarding la sua porta connessa al nuovo link, così come tutte le sue eventuali altre porte root e designate connesse su altri link, per prevenire il formarsi di eventuali cicli durante il transitorio;
2. ciascuno dei due bridge propone la sua porta come designata del link inviando una BPDU sul nuovo link con il flag Proposal impostato;
3. il bridge peggiore accetta la proposta dell'altro bridge inviando in risposta una BPDU con il flag Agreement impostato, e mette la sua porta nel ruolo appropriato (root, alternate o backup) secondo i criteri dell'algoritmo di spanning tree;
4. il bridge migliore riceve la BPDU di accettazione e mette la sua porta come designata del link;
5. ciascuno dei due bridge ripete la sequenza per le altre porte che all'inizio aveva messo in stato di discarding.

La cooperazione tra i due bridge tramite l'invio di BPDU è più rapida rispetto al meccanismo basato su timer dello STP classico, e più efficiente in quanto non blocca tutta la rete per un certo tempo ma di volta in volta solo l'intorno di un bridge. Il nuovo link dev'essere full duplex affinché le BPDU possano essere scambiate in entrambi i sensi: la BPDU di proposta in una direzione e la BPDU di accettazione nell'altra.

7.3.2 Aggiornamento dei filtering database

Rilevamento di cambiamenti di topologia

Il RSTP mira ad essere meno invasivo rispetto allo STP classico per quanto riguarda l'aggiornamento dei filtering database in seguito a cambiamenti di topologia: infatti evita di pulire i filtering database dalle vecchie entry, con il conseguente aumento considerevole del traffico mandato in flooding, quando non è necessario.

Passaggio allo stato di discarding Il passaggio di una porta allo stato di discarding non scatena l'aggiornamento dei filtering database:

- se il link rimosso non apparteneva a un ciclo, cioè non esistono dei percorsi alternativi, allora le stazioni nell'altro segmento di rete non sono più raggiungibili e le entry ad esse associate non sono più valide, ma ciò non è considerato un problema da risolvere immediatamente: se una trama viene inviata a una di quelle stazioni, essa arriverà al bridge che era collegato al link rimosso e verrà scartata, fino a quando la entry non scadrà naturalmente e verrà cancellata dal bridge senza dover toccare le altre entry;
- se il link rimosso apparteneva a un ciclo, cioè esiste un percorso alternativo attraverso una porta in stato di discarding, allora sarà quest'ultima porta a scatenare l'aggiornamento dei filtering database nel passare allo stato di forwarding secondo i meccanismi del RSTP.

Passaggio allo stato di forwarding Solo il passaggio di una porta non edge allo stato di forwarding scatena l'aggiornamento dei filtering database:

- se il nuovo link non crea un ciclo, allora non sarebbe necessario scatenare l'aggiornamento dei filtering database perché nessuna stazione diventa irraggiungibile, ma si ricorda che un bridge non ha la conoscenza della topologia globale della rete;
- se il nuovo link crea un ciclo, allora il passaggio allo stato di forwarding della porta comporta il passaggio allo stato di discarding di un'altra porta lungo il ciclo secondo i meccanismi del RSTP ⇒ le stazioni che si raggiungevano attraverso quella porta sono ora raggiungibili per un altro percorso, e occorre perciò aggiornare le entry ad esse associate.

Annuncio di cambiamenti di topologia

Quando un bridge rileva un cambiamento di topologia che richiede l'aggiornamento dei filtering database:

1. il bridge che ha rilevato il cambiamento di topologia genera su tutte le sue porte root e designate una BPDU con il flag Topology Change impostato;[3]
2. ogni bridge, quando riceve la BPDU:
   (a) scarta tutte le entry nel suo filtering database associate a tutte le sue porte root e designate, tranne quella da cui ha ricevuto la BPDU;
   (b) propaga la BPDU su tutte le sue porte root e designate, tranne quella da cui ha ricevuto la BPDU.[3]

7.3.3 Comportamento delle porte edge

Quando un host si connette a una porta edge, la porta diventa subito designata e passa allo stato di forwarding senza transitare per lo stato di learning ⇒ non è più necessario aspettare 30 secondi (2 volte il Forward Delay) prima di avere la porta pienamente operativa.

Inoltre, le porte edge non scatenano mai l'aggiornamento dei filtering database né al passaggio allo stato di forwarding (connessione di host) né al passaggio allo stato di discarding (disconnessione di host) ⇒ l'utente non sperimenta più un rallentamento della rete a causa dell'aumento di traffico inviato in flooding, e la prima trama broadcast che l'host invierà aggiornerà i filtering database secondo i normali algoritmi di apprendimento dei bridge.

Una porta edge si tiene comunque in ascolto di BPDU in arrivo da eventuali bridge collegati erroneamente ad essa, in modo da essere pronta ad uscire immediatamente dal ruolo di edge e assumere uno degli altri ruoli per proteggere la rete da possibili cicli.

[3] Il bridge continua a generare/propagare la BPDU finché non scade il timer TC While dopo un tempo pari a due volte l'Hello Time.

7.4 Problemi

7.4.1 Coesistenza di STP e RSTP

Se viene introdotto nella rete un bridge che non supporta il RSTP, alla ricezione di Configuration BPDU con Type pari a 0 sono in grado di passare automaticamente in modalità STP, ma ciò ha degli effetti collaterali:

- a causa di un singolo bridge che non supporta il RSTP, tutta la rete va in modalità STP e vengono così persi i tempi di convergenza rapidi;
- se il singolo bridge che non supporta il RSTP si guasta o viene scollegato dalla rete, gli altri bridge continuano a operare in modalità STP, ed è necessario praticare esplicita configurazione manuale su ogni singolo bridge.

Un bridge può essere configurato in modo da operare in modalità RSTP su alcune porte, e in modalità STP su altre porte ⇒ la rete è suddivisa in due porzioni che operano con versioni diverse del protocollo di spanning tree. Ciò tuttavia può portare all'instabilità della rete a causa di cicli transitori dovuti al fatto che la porzione RSTP attiva l'inoltro delle trame dati prima della porzione STP.

Per una coesistenza senza problemi di bridge RSTP e non RSTP nella stessa rete, occorre utilizzare il **Multiple Spanning Tree Protocol**, standardizzato come IEEE 802.1s (2002): le porzioni della rete che funzionano con RSTP e quelle che funzionano con STP sono separate in domini diversi.

7.4.2 Affidabilità del livello fisico

Il RSTP funziona perfettamente quando i link a livello fisico sono affidabili. Se invece un link si attiva e si disattiva frequentemente a causa di un connettore sporco (le fibre ottiche sono piuttosto sensibili), il RSTP riconfigura la rete a ogni cambiamento di stato del link ⇒ la rete rimarrà in uno stato transitorio instabile per la maggior parte del tempo, a causa della troppo rapida reattività del RSTP.

Il meccanismo di "antiflapping" proprietario di Cisco mette la porta in stato "error disabled" quando rileva il flapping di un link.

Parte III

Standard aggiuntivi per le LAN

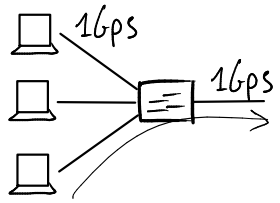
Capitolo 8

Qualità del servizio nelle LAN IEEE 802

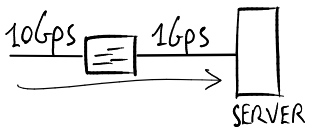
La **qualità del servizio** nell'inoltro del traffico è richiesta quando c'è una limitata quantità di risorse tale che il traffico offerto eccede la capacità di smaltimento dei dati creando delle congestioni.

Di solito le LAN sono sovrabbondanti, in quanto è molto più economico espandere la rete che imporre la qualità del servizio ⇒ nel caso peggiore, l'occupazione del canale è pari al 30-40% della banda disponibile ⇒ apparentemente non c'è alcun bisogno della qualità del servizio perché non ci sono congestioni.

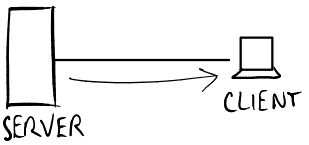
In alcuni scenari possibili si potrebbero avere dei problemi:



(a) Backbone non ben dimensionato.



(b) Trasferimento dati da diversi client a un singolo server.



(c) Trasferimento dati da un server veloce a un client lento.

(a) un bridge dotato di buffer troppo piccoli nel backbone può portare a **micro-congestioni** sugli uplink, le quali non sono persistenti ma sono tante e di breve durata (micro) perché il traffico dei client è estremamente bursty;

(b) un bridge dotato di buffer troppo piccoli come singolo punto di accesso a un server può portare a **congestioni persistenti** dovute alla concorrenza di tanti client collegati al server allo stesso tempo;

(c) un client lento (in termini di velocità del link, capacità della CPU, ecc.) può portare a **congestioni temporanee** sul client stesso perché non riesce a smaltire il traffico proveniente da un server veloce.

La qualità del servizio è potenzialmente una bella funzionalità, ma presenta delle controindicazioni che ne rendono non così forte la necessità: la qualità del servizio è solo uno dei problemi da risolvere per rendere la rete efficiente, e i miglioramenti che essa porta spesso non sono percepiti dall'utente finale.

8.1 IEEE 802.1p

Lo standard **IEEE 802.1p** definisce 8 classi di servizio, chiamate **livelli di priorità**, e a ognuna di esse è assegnata una diversa coda (logica).

Una trama può essere contrassegnato con una specifica classe di servizio nel campo "Priority Code Point" (PCP) del tag VLAN (si rimanda alla sezione 11.3).[1] Lo standard offre anche la possibilità di selezionare l'algoritmo di scheduling a priorità desiderato: round robin, weighted round robin, weighted fair queuing.

Sarebbe meglio lasciare che la sorgente, a livello applicazione, effettui la marcatura perché solo la sorgente conosce esattamente il tipo di traffico (traffico voce o traffico dati), ma quasi tutti gli utenti dichiarerebbero tutti i pacchetti come ad alta priorità perché non sarebbero onesti ⇒ la marcatura va eseguita dai bridge di accesso che sono sotto il controllo del provider. Tuttavia il riconoscimento del tipo di traffico è molto difficile per i bridge e li rende molto costosi, perché richiede di salire al livello applicazione e può non funzionare con il traffico criptato ⇒ si può semplificare la distinzione per i bridge in due modi:

- marcatura per porta: il PC è connesso a una porta e il telefono a un'altra porta, così il bridge può marcare il traffico in base alla porta di ingresso;
- marcatura per dispositivo edge: il PC è connesso al telefono e il telefono al bridge ⇒ tutto il traffico del PC passa per il telefono, che semplicemente lo marca come traffico dati, mentre marca il suo traffico come traffico voce.

Lo standard suggerisce a quale tipo di traffico è destinato ogni livello di priorità (ad es. 6 = traffico voce), ma lascia la libertà di cambiare queste associazioni ⇒ possono sorgere dei problemi di interoperabilità tra fornitori diversi.

8.2 IEEE 802.3x

Lo standard **802.3x** implementa un **controllo di flusso** a livello Ethernet, in aggiunta al controllo di flusso esistente a livello TCP: dato un link, se il nodo (bridge o host) a valle ha i buffer pieni può inviare al nodo a monte all'altra estremità del link un **pacchetto PAUSE** chiedendo ad esso di interrompere la trasmissione dei dati su quel link per una certa quantità di tempo, detta **tempo di pausa** che è espresso in "quanti di pausa" (1 quanto = tempo per trasmettere 512 bit). Il nodo a monte quindi memorizza i pacchetti che arrivano durante il tempo di pausa nel suo buffer di uscita, e li invierà quando il buffer di ingresso del nodo a valle sarà pronto a ricevere altri pacchetti ⇒ i pacchetti non vanno più persi per colpa della congestione dei buffer.

Esistono due **modalità di controllo di flusso**:

- modalità asimmetrica: solo un nodo invia il pacchetto PAUSE, l'altro si limita a ricevere il pacchetto e a interrompere la trasmissione;
- modalità simmetrica: entrambi i nodi alle estremità del link possono trasmettere e ricevere pacchetti PAUSE.

Su ogni nodo si può configurare la modalità di controllo di flusso, ma la fase di autonegoziazione deve determinare l'effettiva configurazione in modo che la modalità scelta sia coerente su entrambi i nodi alle estremità del link.

L'invio di pacchetti PAUSE può essere problematico nel backbone: un bridge con i buffer pieni è in grado di far interrompere il traffico solo sul link a cui è direttamente collegato ma,

[1] Esistono due campi di marcatura per la qualità del servizio, uno a livello data-link e l'altro a livello rete:

- il campo "Priority Code Point" (PCP), usato dallo standard IEEE 802.1p, sta nell'intestazione della trama Ethernet;
- il campo "Differentiated Service Code Point" (DSCP), usato dall'architettura Differentiated Services (DiffServ), sta nell'intestazione del pacchetto IP, in particolare nel campo "Type of Service" dell'intestazione IPv4 e nel campo "Priority" dell'intestazione IPv6.

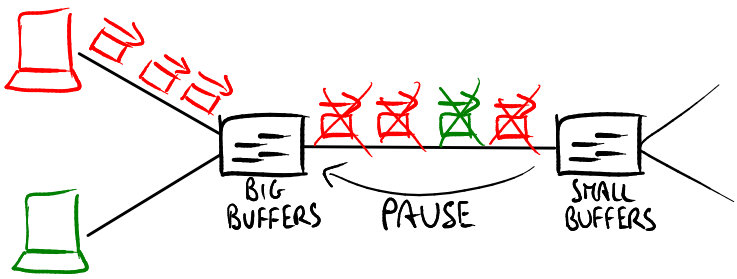


Figura 8.2: L'invio di pacchetti PAUSE può essere problematico nel backbone.

se i bridge intermedi nel percorso di upstream non sentono la necessità di mandare a loro volta dei pacchetti PAUSE perché dotati di buffer più grandi, non è in grado di "zittire" l'host che sta inviando troppi pacchetti ⇒ finché il bridge di accesso non invierà a sua volta un pacchetto PAUSE all'host interessato, la rete appare bloccata anche a tutti gli altri host che non sono responsabili del problema ⇒ i pacchetti PAUSE inviati da bridge non di accesso non hanno la capacità di selezionare il traffico in eccesso per rallentare l'host responsabile, ma hanno effetto sul traffico di tutti gli host.

Per questo motivo, è consigliabile disabilitare il controllo di flusso nel backbone e utilizzare i pacchetti PAUSE solo tra i bridge di accesso e gli host. Spesso viene scelta la modalità di controllo di flusso asimmetrica, dove solo gli host possono mandare i pacchetti PAUSE: in genere i buffer dei bridge di accesso sono sufficientemente grandi, e diversi bridge commerciali accettano pacchetti PAUSE dagli host, bloccando la trasmissione dati sulla porta interessata, ma non possono inviarli.

Tuttavia l'invio dei pacchetti PAUSE può essere problematico anche per gli host, in quanto può scatenare un **livelock** nel kernel del sistema operativo: la CPU dell'host lento è così impegnata ad elaborare i pacchetti in arrivo sull'interfaccia NIC che non riesce a trovare un momento per mandare un pacchetto PAUSE ⇒ i pacchetti si accumulano in RAM portandola alla saturazione.

Capitolo 9

Link aggregation – IEEE 802.3ad

Il **link aggregation**, standardizzato come IEEE 802.3ad, è normalmente utilizzato tra i bridge nel backbone o tra un bridge e un server per aggregare più link fisici (solitamente 2-4) in un singolo canale logico al fine di:

- incrementare la capacità di banda del link: il traffico viene distribuito tra i link nell'aggregato;
- migliorare la resilienza, ovvero la tolleranza ai guasti: in caso di guasto di uno dei link nell'aggregato:
  - la riduzione della larghezza di banda del canale logico è lieve;
  - non è necessario attendere i tempi di convergenza dello STP, che vede il canale logico come un singolo link di capacità maggiore ⇒ cambia solo il costo del link.

Tutti i link fisici aggregati nello stesso gruppo devono:

- essere punto-punto tra gli stessi due nodi;
- essere full duplex;
- avere la stessa velocità.

9.1 LACP

Il **Link Aggregation Control Protocol** (LACP) serve per la configurazione automatica degli aggregati:

1. prima di tutto le porte da aggregare devono essere impostate manualmente sui bridge dall'amministratore di rete;
2. prima di attivare le porte aggregate, il LACP è in grado automaticamente di riconoscere il numero di link disponibili nel canale, e di verificare se la connessione con l'altra parte è corretta (in particolare se tutti i link stanno tra gli stessi bridge);
3. vengono scambiati periodicamente dei messaggi chiamati **LACPDU** per rilevare eventuali guasti dei link ⇒ la convergenza è rapida (solitamente meno di 1 s) in caso di guasto.

Ogni aggregato è identificato da un **Link Aggregation Group Identifier** (LAG ID).

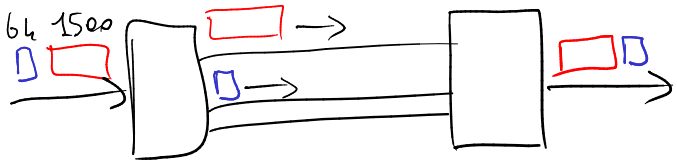
9.2 Distribuzione delle trame sulle porte aggregate

Quando arriva una trama, su quale dei link all'interno di un aggregato va inviata? Lo standard, pur suggerendo dei possibili criteri di distribuzione delle trame sulle porte, non definisce un algoritmo per distribuire le trame ⇒ bridge di fornitori diversi possono utilizzare algoritmi diversi di distribuzione delle trame.

9.2.1 Round robin

La soluzione più semplice consiste nell'inoltrare la trama in arrivo sulla porta libera successiva a quella su cui è stata inoltrata la trama precedente.

Possono nascere dei problemi di **riordinamento**: una trama più piccola che arriva al bridge subito dopo una trama più grande può terminare di essere ricevuta dall'altro bridge prima della trama più grande ⇒ l'ordine delle trame in uscita dall'altro bridge non è corretto perché la trama più piccola ha "superato" quella più grande:



9.2.2 In base alle conversazioni

Il problema del riordinamento delle trame si può risolvere inviando sullo stesso link le trame appartenenti alla stessa conversazione. La soluzione più comune per individuare le trame appartenenti alla stessa conversazione si basa sulle coppie indirizzo MAC sorgente e indirizzo MAC di destinazione.

L'individuazione delle conversazioni in base agli indirizzi MAC però in alcuni casi non è efficace in termini di bilanciamento del carico dei link:

- due soli host comunicano attraverso l'aggregato ⇒ la conversazione è unica e può sfruttare la capacità di un solo link fisico;
- l'aggregato collega due router ⇒ le conversazioni non sono più riconoscibili perché i router cambiano gli indirizzi MAC delle trame.

9.3 Configurazioni particolari

Se due nodi sono collegati da più aggregati, solo un aggregato sarà attivo a causa dello STP: lo STP disabiliterà l'altro aggregato perché vede ogni aggregato come un singolo link di costo pari alla somma dei costi dei link nell'aggregato.

Tramite un'appropriata impostazione delle priorità dei link, è possibile una configurazione con N link aggregati di cui solo $M < N$ attivi ⇒ gli altri $N - M$ link sono **stand-by link**: in caso di guasto di un link attivo, uno stand-by link si attiva evitando di ridurre la larghezza di banda disponibile nel canale logico.

La funzione **Virtual Switching System** proprietaria di Cisco permette di superare il vincolo di avere due soli nodi alle estremità degli aggregati: un bridge può essere collegato a due bridge che lo STP vede come un singolo bridge logico, così il traffico può essere distribuito su entrambi i link aggregati.

Capitolo 10

IGMP snooping

I bridge inoltrano le trame in modi diversi a seconda del tipo di indirizzo MAC di destinazione:

- trame unicast: sono mandate solo sulla porta verso la singola destinazione, grazie al filtering database;
- trame broadcast: sono mandate sempre in flooding su tutte le porte, poiché le destinazioni sono tutti gli host nella rete;
- trame multicast: sono mandate sempre in flooding su tutte le porte, anche se le destinazioni sono solo alcuni degli host nella rete.

Se il bridge conoscesse a quali gruppi multicast appartengono le stazioni collegate alle sue porte, il bridge potrebbe inoltrare le trame dirette a un certo gruppo multicast solo sulle porte su cui sono collegati host registrati a quel gruppo multicast, al fine di ridurre il traffico mandato in flooding.

10.1 GMRP

Il **GARP Multicast Registration Protocol** (GMRP) permette a una stazione di comunicare il proprio gruppo multicast di appartenenza al bridge.

Il GMRP tuttavia è scarsamente utilizzato, poiché all'aggiunta di un nuovo protocollo di rete si preferisce sfruttare una tecnologia già esistente e comunemente utilizzata, cioè l'IGMP.

10.2 IGMP snooping

10.2.1 IGMP

L'**Internet Group Management Protocol** (IGMP) consente a una stazione di comunicare il proprio gruppo multicast di appartenenza ai router sulla rete IP:

1. messaggio **Host Membership Query**: il router invia a tutti gli host un messaggio IGMP chiedendo se qualcuno di essi è interessato a registrarsi ad un certo gruppo multicast;
2. messaggio **Host Membership Report**: l'host invia in risposta un messaggio IGMP accettando la richiesta di registrazione al gruppo multicast.
   Il messaggio Host Membership Report arriva, oltre che al router, anche a tutte le altre stazioni sulla LAN ⇒ ogni altra stazione interessata al gruppo multicast, sapendo che almeno una stazione sulla LAN si è registrata a quel gruppo, può evitare l'invio di un messaggio Host Membership Report al router, in quanto il traffico relativo a quel gruppo multicast esce dall'interfaccia del router e si propaga a tutta la LAN.

Ogni messaggio IGMP ha:

- indirizzo IP di destinazione: è l'indirizzo IP del gruppo multicast di cui si sta effettuando la query o il report, che inizia sempre con i bit "1110";
- indirizzo MAC di destinazione: è derivato dall'indirizzo IP multicast:

24	25	48
01:00:5E	0	ultimi 23 bit dell'indirizzo IP multicast

Gli indirizzi IP multicast del tipo "224.0.0.x" sono indirizzi "well-known" che non richiedono l'IGMP (ad es. i pacchetti multicast inviati dai protocolli di instradamento del livello rete).

10.2.2 Come l'IGMP viene sfruttato

La funzione **IGMP snooping** permette a un bridge di apprendere a quali gruppi multicast sono registrate le stazioni collegate alle sue porte, osservando i messaggi IGMP che passano per il bridge stesso:

1. messaggio Host Membership Query: il bridge registra la porta da cui è arrivato come porta verso il router, e lo manda in flooding su tutte le altre porte;
2. messaggio Host Membership Report: il bridge registra la porta da cui è arrivato come porta verso una stazione interessata, e lo manda solo sulla porta verso il router (cioè quella da cui è arrivato il messaggio Host Membership Query).
   Il bridge non lo manda sulle altre porte, perché altrimenti gli host ricevendolo disabiliterebbero l'invio di messaggi Host Membership Report, impedendo al bridge di sapere quali host sono interessati a quel gruppo multicast;
3. trama inviata in multicast: il bridge ne analizza l'indirizzo MAC di destinazione per individuare il gruppo multicast[1]:
   - se è un indirizzo multicast "well-known", la inoltra su tutte le altre porte in flooding;
   - se è un indirizzo multicast dinamico, la invia solo sulle porte collegate alle stazioni registrate a quel gruppo multicast.

Svantaggio L'IGMP snooping è una violazione del modello OSI: ai bridge è richiesto di riconoscere se la trama di livello data-link incapsula un pacchetto IP che a sua volta incapsula un messaggio IGMP ⇒ i bridge non operano più in modo indipendente dal livello rete: i bridge che non supportano il protocollo IPv6 potrebbero scartare i pacchetti multicast, molto utilizzati in IPv6 (ad es. nel processo di autoconfigurazione), perché non li riconoscono.

[1] In realtà un singolo indirizzo MAC multicast corrisponde a 2^5 indirizzi IP = 2^5 gruppi multicast.

Parte IV

Configurazione e progettazione avanzate delle LAN

Capitolo 11

Virtual LAN

Le **Virtual LAN** (VLAN) permettono di condividere una singola infrastruttura fisica (stessi apparati, stesso cablaggio) tra più LAN logiche: attraverso alcune porte di un bridge passa solo il traffico di una certa LAN, attraverso altre porte passa solo il traffico di un'altra LAN, e così via ⇒ ogni bridge ha un filtering database per ogni VLAN.[1]

Una rete di livello data-link costituita da più VLAN è più vantaggiosa rispetto a:

- una rete di livello rete, grazie al supporto della mobilità: gli host possono continuare a essere raggiungibili allo stesso indirizzo (l'indirizzo MAC) quando si spostano;
- una singola LAN fisica, grazie a:
  - maggiore scalabilità: il traffico broadcast è confinato in domini di broadcast più ridotti;
  - maggiore sicurezza: un utente appartenente a una VLAN non può effettuare un MAC flooding attack su altre VLAN;
  - migliore policing: l'amministratore di rete può configurare politiche diverse in base alla VLAN;
- più LAN completamente separate dal punto di vista fisico, grazie al maggiore risparmio di risorse e di costi: i bridge non sono duplicati per ogni LAN ma sono condivisi tra tutte le VLAN, così come i cavi tra i bridge possono trasportare il traffico di qualsiasi VLAN.

Un esempio di applicazione delle VLAN è la rete di un'università: una VLAN è riservata agli studenti, un'altra VLAN è riservata ai docenti con politiche meno restrittive, e così via.

11.1 Interconnessione di VLAN

I dati non possono attraversare a livello data-link i confini delle VLAN: una stazione in una VLAN non può mandare una trama a un'altra stazione in una VLAN differente, poiché le VLAN hanno domini di broadcast differenti. Una soluzione possibile potrebbe consistere nel collegare la porta di una VLAN alla porta di un'altra VLAN, ma in questo modo si creerebbe un unico dominio di broadcast ⇒ le due VLAN apparterrebbero di fatto alla stessa LAN.

Pertanto una stazione in una VLAN può mandare dati a un'altra stazione in una VLAN differente solo a livello rete ⇒ serve un router che colleghi la porta di una VLAN con la porta di un'altra VLAN: una stazione in una VLAN invia un pacchetto IP[2] verso il router, quindi quest'ultimo rigenera l'intestazione di livello data-link del pacchetto (in particolare cambia gli indirizzi MAC) e invia il pacchetto alla stazione nell'altra VLAN. Questa soluzione però occupa due interfacce di un router e due porte di uno stesso bridge, e richiede due fili che collegano questi

[1] Nell'implementazione reale, il filtering database è unico e solitamente realizzato con una singola TCAM nell'intero apparato di rete.

[2] Per semplicità qui si considera il protocollo IP come protocollo di livello rete.

stessi due apparati di rete ⇒ un **router a braccio singolo** permette di interconnettere due VLAN tramite un singolo filo, occupando una singola porta del bridge e una singola interfaccia del router: attraverso l'unico filo e attraverso la porta del bridge può passare il traffico di entrambe le VLAN.

Il traffico broadcast di livello data-link non può ancora attraversare i confini delle VLAN, poiché il router non lo propaga sulle altre interfacce spezzando il dominio di broadcast ⇒ una stazione in una VLAN che vuole contattare una stazione in un'altra VLAN non può scoprirne l'indirizzo MAC tramite il protocollo ARP, ma deve inviare un pacchetto al suo indirizzo IP, il quale ha un diverso prefisso di rete poiché le due VLAN devono avere spazi di indirizzamento diversi.

11.2 Assegnazione di host alle VLAN

Ogni bridge mette a disposizione delle porte, dette **porte access**, a cui gli host si possono collegare tramite dei **link access**. Nei link access passano **trame non contrassegnate**, ossia prive del tag VLAN (si rimanda alla sezione 11.3); le porte access contrassegnano le trame in base alle VLAN di appartenenza.

Quando un host si collega a una porta access, si può riconoscerne la VLAN di appartenenza in quattro modi:

- assegnazione basata sulle porte (sezione 11.2.1)
- assegnazione trasparente (sezione 11.2.2)
- assegnazione per utente (sezione 11.2.3)
- assegnazione cooperativa o anarchica (sezione 11.2.4)

11.2.1 Assegnazione basata sulle porte

Ogni porta access è associata a una singola VLAN ⇒ un host può accedere a una VLAN collegandosi alla relativa porta sul bridge.

Vantaggi

- configurazione: non è necessario configurare le VLAN sugli host ⇒ compatibilità massima con i dispositivi.

Svantaggi

- sicurezza: l'utente può collegarsi a qualsiasi VLAN ⇒ non è possibile stabilire delle politiche diverse in base alle VLAN;
- mobilità a livello rete: sebbene l'utente possa collegarsi a qualsiasi VLAN, non può comunque mantenere lo stesso indirizzo IP da una VLAN all'altra.

11.2.2 Assegnazione trasparente

Ogni host è associato a una certa VLAN in base al suo indirizzo MAC.

Svantaggi

- configurazione: un nuovo utente deve contattare l'amministratore di rete per registrare l'indirizzo MAC del suo dispositivo ⇒ può non essere semplice per un utente trovare l'indirizzo MAC del proprio dispositivo;

- costo del database: è necessario un server, insieme a del personale per la sua gestione, che memorizzi il database contenente le associazioni tra gli indirizzi MAC e le VLAN;
- manutenzione del database: occorre cancellare le entry corrispondenti a indirizzi MAC non più in uso, ma l'utente spesso quando dismette un dispositivo si dimentica di ricontattare l'amministratore di rete per chiederne la cancellazione dell'indirizzo MAC ⇒ nel tempo il database continua a crescere;
- sicurezza: l'utente può configurare un indirizzo MAC fasullo e accedere a un'altra VLAN fingendosi un altro utente.

11.2.3 Assegnazione per utente

Ogni utente possiede un account, e ogni account utente è associato a una certa VLAN. Quando si collega alla porta di un bridge, l'utente si autentica inserendo le proprie credenziali di accesso tramite il protocollo standard 802.1x, e il bridge è in grado di contattare un server RADIUS per verificare le credenziali e di assegnare all'utente la VLAN appropriata in caso di successo.

Svantaggi

- compatibilità: l'autenticazione viene svolta a livello data-link direttamente dalla scheda di rete ⇒ ogni dispositivo deve disporre di una scheda di rete compatibile con lo standard 802.1x;
- configurazione: l'utente deve impostare molti parametri di configurazione (ad es. il tipo di autenticazione) sul proprio dispositivo prima di poter accedere alla rete.

11.2.4 Assegnazione cooperativa

Ogni utente si associa da sé alla VLAN che desidera: è il sistema operativo sull'host a contrassegnare le trame in uscita, così esse arrivano tramite un link trunk alla porta del bridge già contrassegnate.

Svantaggi

- configurazione: l'utente deve configurare manualmente il proprio dispositivo prima di poter accedere alla rete;
- sicurezza: l'utente può collegarsi a qualsiasi VLAN ⇒ non è possibile stabilire delle politiche diverse in base alle VLAN.

11.3 Tagging delle trame

I **link trunk** sono i link che possono trasportare il traffico di VLAN diverse:

- link trunk tra bridge (sezione 11.3.1)
- link trunk tra un bridge e un server (sezione 11.3.2)
- link trunk tra un bridge e un router a braccio singolo (sezione 11.3.2)

Nei link trunk passano **trame contrassegnate**, ossia dotate del tag VLAN standardizzato come **IEEE 802.1Q** (1998):

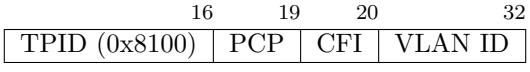


Tabella 11.1: Formato del tag VLAN (4 byte).

dove i campi sono:

- campo Tag Protocol Identifier (TPID) (2 byte): identifica una trama contrassegnata (valore 0x8100);
- campo Priority Code Point (PCP) (3 bit): specifica la priorità utente per la qualità del servizio[3];
- flag Canonical Format Indicator (CFI) (1 bit): specifica se l'indirizzo MAC è in formato canonico (valore 0, ad es. Ethernet) oppure no (valore 1, ad es. token ring);
- campo VLAN Identifier (VID) (12 bit): identifica la VLAN della trama:
  - valore 0: la trama non appartiene ad alcuna VLAN ⇒ utilizzato nel caso l'utente voglia solamente impostare la priorità per il traffico;
  - valore 1: la trama appartiene alla VLAN predefinita;
  - valori da 2 a 4094: la trama appartiene alla VLAN identificata da questo valore;
  - valore 4095: riservato.

IEEE 802.1Q in realtà non incapsula la trama originale; invece, aggiunge il tag tra i campi indirizzo MAC sorgente e EtherType/Length della trama originale, lasciando la dimensione minima della trama invariata a 64 byte ed estendendo la dimensione massima della trama da 1518 byte a 1522 byte ⇒ sui link trunk non possono esserci degli hub perché non supportano le trame più lunghe di 1518 byte:

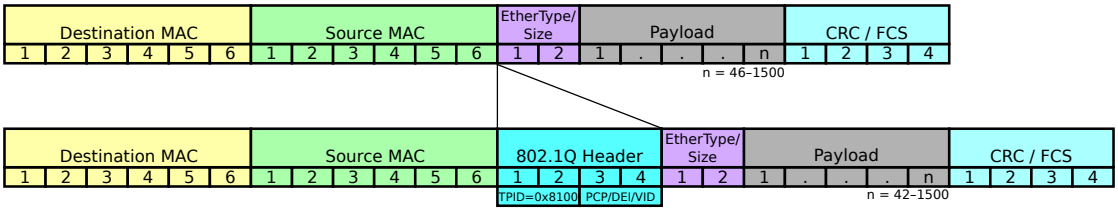


Figura 11.1: Inserimento del tag VLAN in una trama Ethernet.[4]

11.3.1 Nel backbone

Il trasporto di una trama da una stazione all'altra attraverso i link trunk avviene nel seguente modo:[5]

1. l'host sorgente invia verso la porta access una trama non contrassegnata;
2. quando una trama arriva alla porta access, il bridge contrassegna la trama con il tag corrispondente alla VLAN associata alla porta;
3. il bridge invia la trama contrassegnata su un link trunk;
4. ogni bridge che riceve la trama guarda il filtering database relativo alla VLAN specificata dal tag:
   - se la destinazione è "remota", il bridge propaga la trama su un link trunk lasciandone il tag VLAN invariato;
   - se la destinazione è "locale", ossia è raggiungibile attraverso una delle porte access associate alla VLAN della trama, il bridge rimuove il tag VLAN dalla trama e invia la trama non contrassegnata sul link access verso l'host di destinazione.

[3] Si veda la sezione 8.1.

[4] Questa immagine è derivata da un'immagine su Wikimedia Commons (Ethernet 802.1Q Insert.svg), realizzata dall'utente Arkrishna e da Bill Stafford, ed è concessa sotto la licenza Creative Commons Attribuzione 3.0 Unported.

[5] Si assume di adottare l'assegnazione basata sulle porte (si veda la sezione 11.2.1).

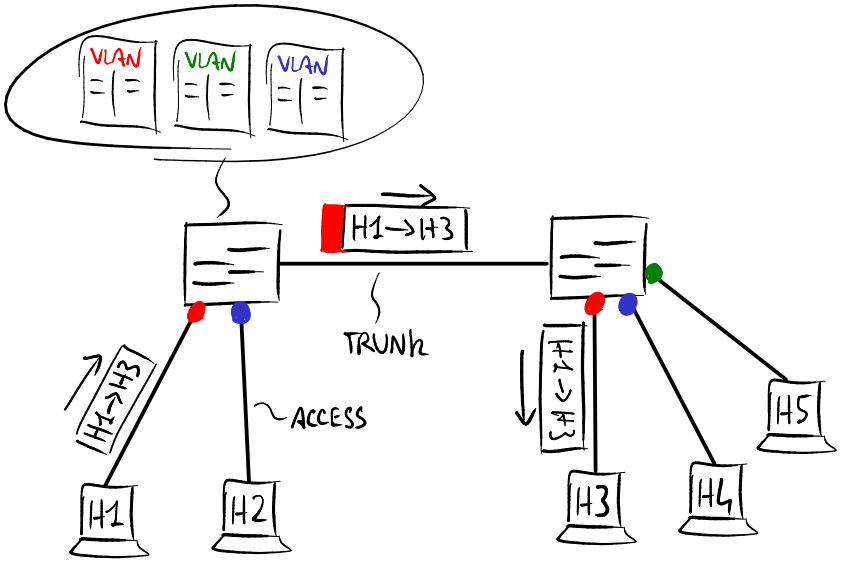


Figura 11.2: Esempio di trasporto VLAN di una trama attraverso un link trunk nel backbone.

11.3.2 Interfacce di rete virtuali

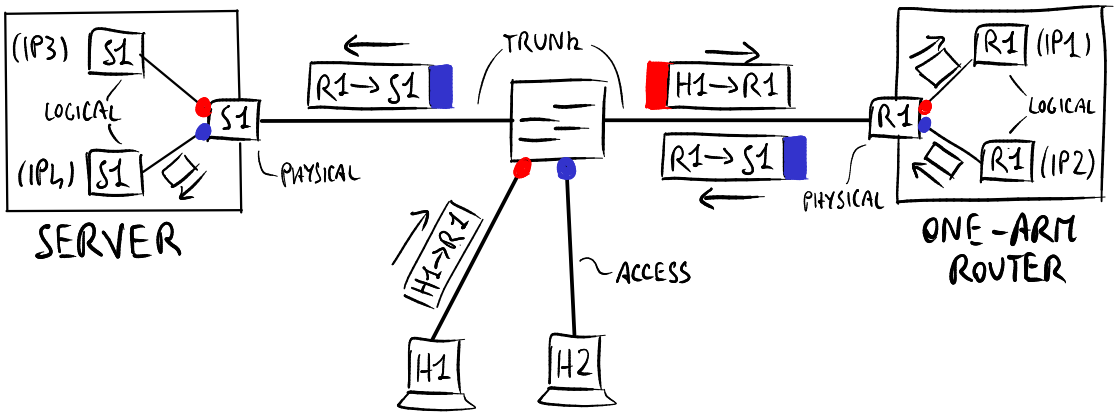


Figura 11.3: Esempio di trasporto VLAN di una trama attraverso le interfacce di rete virtuali di un router a braccio singolo e di un server.

Un server tipicamente ha bisogno di essere contattato allo stesso tempo da più host situati in VLAN diverse ⇒ siccome a ogni interfaccia di rete è possibile associare solo una VLAN, il server richiederebbe di avere una interfaccia di rete per ogni VLAN, ciascuna collegata al bridge con un proprio link fisico. Analogo problema vale per un router a braccio singolo, che ha bisogno di ricevere e mandare traffico da/a più VLAN diverse per permettere la loro interconnessione.

Le **interfacce di rete virtuali** permettono di avere allo stesso tempo più interfacce di rete logiche virtualizzate sulla stessa scheda di rete fisica, la cui singola interfaccia fisica è collegata con il bridge tramite un unico link fisico trunk: il sistema operativo vede più interfacce di rete installate nel sistema, e il driver della scheda di rete in base al tag VLAN espone al sistema operativo ogni trama come se fosse arrivata da una delle interfacce di rete virtuali.

Le interfacce di rete virtuali hanno indirizzi IP diversi, perché ogni VLAN ha il suo spazio di indirizzamento, ma hanno lo stesso indirizzo MAC, uguale a quello della scheda di rete fisica; ciò non costituisce tuttavia un problema in quanto è sufficiente che l'indirizzo MAC sia univoco all'interno del dominio di broadcast (quindi all'interno della VLAN).

11.3.3 Tag stacking

Il **tag stacking** (anche noto come "provider bridging" o "Stacked VLAN" o "QinQ"), standardizzato come IEEE 802.1ad (2005), permette di inserire più tag VLAN nella pila di una trama contrassegnata, dal tag più esterno a quello più interno:

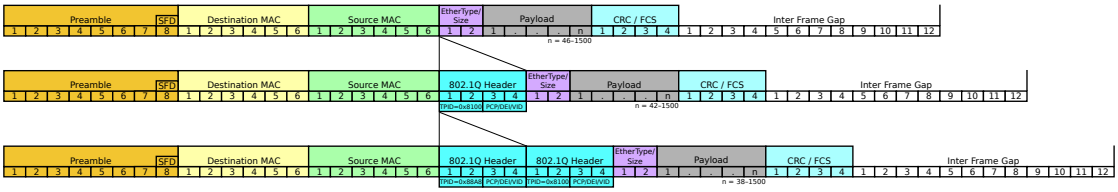


Figura 11.4: Inserimento di due tag VLAN in una trama Ethernet.[6]

Il tag stacking è utile per trasportare il traffico di più clienti che utilizzano le VLAN su una rete del provider condivisa: due clienti diversi potrebbero decidere di utilizzare lo stesso VLAN Identifier nelle loro reti aziendali ⇒ i bridge ai margini della rete del provider aggiungono alle trame in ingresso e rimuovono dalle trame in uscita dei tag esterni che distinguono le VLAN che hanno lo stesso VLAN Identifier ma sono di clienti diversi.

Vantaggi

- flessibilità: il tag stacking è più flessibile e meno distruttivo rispetto alla definizione di un altro formato di tagging con un VLAN Identifier più grande;
- semplicità: il tag stacking è più semplice rispetto al **tunneling Ethernet**:
  - tunneling Ethernet: i bridge ai margini devono incapsulare la trama in una nuova intestazione Ethernet ⇒ operazione complessa;
  - tag stacking: i bridge ai margini si limitano a effettuare delle più rapide operazioni di push e pop nella pila dei tag;
- scalabilità delle VLAN: il tag stacking è più scalabile rispetto alla **traslazione di VLAN**:
  - traslazione di VLAN: i bridge ai margini modificano il VLAN Identifier di ogni trama in modo tale che ciascun VLAN Identifier sia univoco all'interno della rete del provider ⇒ problema di scalabilità: sono a disposizione solo un massimo di 4094 VLAN;
  - tag stacking: i bridge ai margini usano un VLAN Identifier esterno per ogni cliente, indipendentemente dal numero di VLAN Identifier interni che ogni cliente utilizza ⇒ la rete del provider può servire fino a 4094 clienti, ciascuno con 4094 VLAN.

Svantaggi

- scalabilità degli indirizzi MAC: il tag stacking è meno scalabile rispetto al tunneling Ethernet:
  - tag stacking: il filtering database di ogni bridge nella rete del provider deve apprendere tutti gli indirizzi MAC delle interfacce di rete situate nelle VLAN di tutti i clienti ⇒ problema di scalabilità: i filtering database dei bridge sono memorizzati in memorie TCAM di dimensione limitata;
  - tunneling Ethernet: il filtering database di ogni bridge nella rete del provider vede solo gli indirizzi MAC dei bridge ai margini della rete;
- sicurezza: un broadcast storm sulla VLAN di un cliente può avere ripercussioni sul traffico di altri clienti (si rimanda alla sezione 11.5.1).

[6] Questa immagine è derivata da un'immagine su Wikimedia Commons (TCPIP 802.1ad DoubleTag.svg), realizzata dall'utente Arkrishna e da Luca Ghio, ed è concessa sotto la licenza Creative Commons Attribuzione - Condividi allo stesso modo 4.0 Internazionale.

11.4 PVST

Lo STP e il RSTP standard non supportano le VLAN: l'albero ricoprente è unico nell'intera rete e l'algoritmo di spanning tree opera indipendentemente dalle VLAN. Molti fornitori offrono funzionalità proprietarie per il supporto alle VLAN: per esempio Cisco offre il **Per-VLAN Spanning Tree** (PVST) e il Per-VLAN Spanning Tree Plus (PVST+), basati sullo STP, e il Rapid Per-VLAN Spanning Tree Plus (Rapid-PVST+), basato sul RSTP.

Il PVST consente più alberi ricoprenti nella rete, uno per ogni VLAN; ciascun albero è determinato tramite la configurazione per VLAN dei parametri del protocollo di spanning tree. In particolare, è necessario personalizzare la priorità di ogni bridge in base alla VLAN al fine di differenziare il root bridge fra VLAN diverse, altrimenti risulterebbe lo stesso albero per tutte le VLAN, identificando la VLAN a cui si riferisce il valore di priorità con il campo STP Instance (12 bit), introdotto nel Bridge Identifier da IEEE 802.1t (2001):

4	16	64
Bridge Priority	STP Instance	Bridge MAC Address

Tabella 11.2: Formato del Bridge Identifier stabilito da IEEE 802.1t.

Svantaggi

- ottimizzazione del traffico: l'ottimizzazione operata dal PVST sul carico di traffico non è così significativa, anche tenendo conto della elevata larghezza di banda dei link nelle reti moderne:
  - il PVST ottimizza il carico di traffico nell'intera rete: se gli alberi ricoprenti sono ben bilanciati, tutti i link sono utilizzati ⇒ non ci sono più link attivi ma completamente inutilizzati;
  - il PVST non ottimizza il carico di traffico all'interno di una VLAN: il traffico della VLAN è ancora legato a uno specifico albero ricoprente ⇒ non è possibile scegliere il percorso più breve verso la destinazione come avviene nelle reti IP;
- carico CPU: l'esecuzione di più istanze del protocollo di spanning tree allo stesso tempo aumenta il carico sulle CPU dei bridge;
- interoperabilità: la coesistenza nella stessa rete di bridge dotati del supporto a PVST e di bridge privi di esso può portare a dei broadcast storm;
- complessità: l'amministratore di rete deve gestire più alberi ricoprenti sulla stessa rete ⇒ la risoluzione dei problemi è più complicata: è difficile capire il percorso del traffico, poiché le trame attraversano link diversi a seconda della VLAN a cui appartengono.

11.5 Problemi

11.5.1 Ottimizzazione del traffico broadcast

Il traffico broadcast viene inviato su tutti i link trunk, oltre che sui link access associati alla VLAN a cui la trama broadcast appartiene:

- un broadcast storm su un link, causato dal traffico di una VLAN, può influenzare le altre VLAN saturando i link trunk ⇒ anche se le trame non possono passare da una VLAN all'altra a livello data-link, l'**isolamento della rete** non è completo neanche con le VLAN dovuto al fatto che i link sono condivisi;

- una trama broadcast appartenente a una certa VLAN può raggiungere un bridge all'estremità della rete su cui non è presente alcuna porta access appartenente a quella VLAN ⇒ il filtering database di quel bridge inserirà tramite i meccanismi di apprendimento un'inutile entry contenente l'indirizzo MAC sorgente.

Al fine di ridurre il traffico broadcast sui link trunk ed evitare entry inutili nei filtering database, ogni bridge ha bisogno di sapere di quali VLAN propagare il traffico broadcast su ogni porta trunk:

- protocollo GVRP: è un protocollo standard molto complesso che permette ai bridge di scambiarsi informazioni sulle VLAN nella topologia della rete;
- meccanismi proprietari: vengono impiegati al posto del protocollo GVRP perché sono più semplici, anche se introducono problemi di interoperabilità;
- configurazione manuale: l'amministratore di rete configura esplicitamente su ogni bridge le VLAN di cui propagare il traffico broadcast ⇒ le VLAN sono configurate staticamente e non possono cambiare in caso di riconvergenza dell'albero ricoprente in seguito al guasto di un link.

11.5.2 Interoperabilità

Le VLAN non sono una tecnologia plug and play (com'era lo STP), e gli utenti domestici non sono abbastanza esperti per configurarle ⇒ i bridge di fascia bassa tipicamente sono privi del supporto alle VLAN, e possono scartare le trame contrassegnate perché troppo grosse.

Un altro motivo di incompatibilità tra gli apparati di rete di diversi fornitori è il tagging sulle porte trunk: alcuni bridge contrassegnano il traffico appartenente a tutte le VLAN, altri lasciano il traffico appartenente alla VLAN 1 non contrassegnato.

Capitolo 12

Ridondanza e bilanciamento del carico a livello 3 nelle LAN

Ai confini della LAN aziendale con il livello rete, il router che fornisce la connettività con l'esterno (tipicamente Internet) costituisce, per gli host che lo hanno come loro default gateway, un singolo punto di guasto, a meno che il router non sia ridondato opportunamente.

La semplice duplicazione del router non è sufficiente: gli host non sono in grado di passare automaticamente all'altro router in caso di guasto del loro default gateway, perché essi non sono capaci ad apprendere la topologia di rete tramite i protocolli di instradamento del livello rete.

Sono stati pertanto definiti alcuni protocolli per la gestione automatica di router ridondanti:

- HSRP: protocollo proprietario di Cisco specifico per la **ridondanza del default gateway** e con parziale supporto al bilanciamento del carico (sezione 12.1);
- VRRP: protocollo standard molto simile a HSRP ma libero da brevetti;
- GLBP: protocollo proprietario di Cisco che migliora il **bilanciamento del carico** rispetto a HSRP (sezione 12.2).

12.1 HSRP

L'**Hot Standby Routing Protocol** (HSRP) garantisce automaticamente che ogni host mantenga la connettività con l'esterno della LAN attraverso il proprio default gateway anche in caso di guasto di uno dei router ridondanti.

12.1.1 Configurazione della rete

Alle interfacce appartenenti alla LAN aziendale di tutti i router ridondanti viene assegnato uno stesso **indirizzo IP virtuale** e uno stesso **indirizzo MAC virtuale**, in aggiunta ai loro indirizzi IP e MAC reali. I router possono essere:

- **active**: è il router che ha il diritto di servire la LAN, cioè di rispondere all'indirizzo IP virtuale e all'indirizzo MAC virtuale;
- **stand-by**: è il router che ha il diritto di sostituire il router active in caso di guasto di quest'ultimo;
- **listen**: sono gli altri router né active né stand-by; uno di essi diventa il router stand-by in caso di guasto del router active.

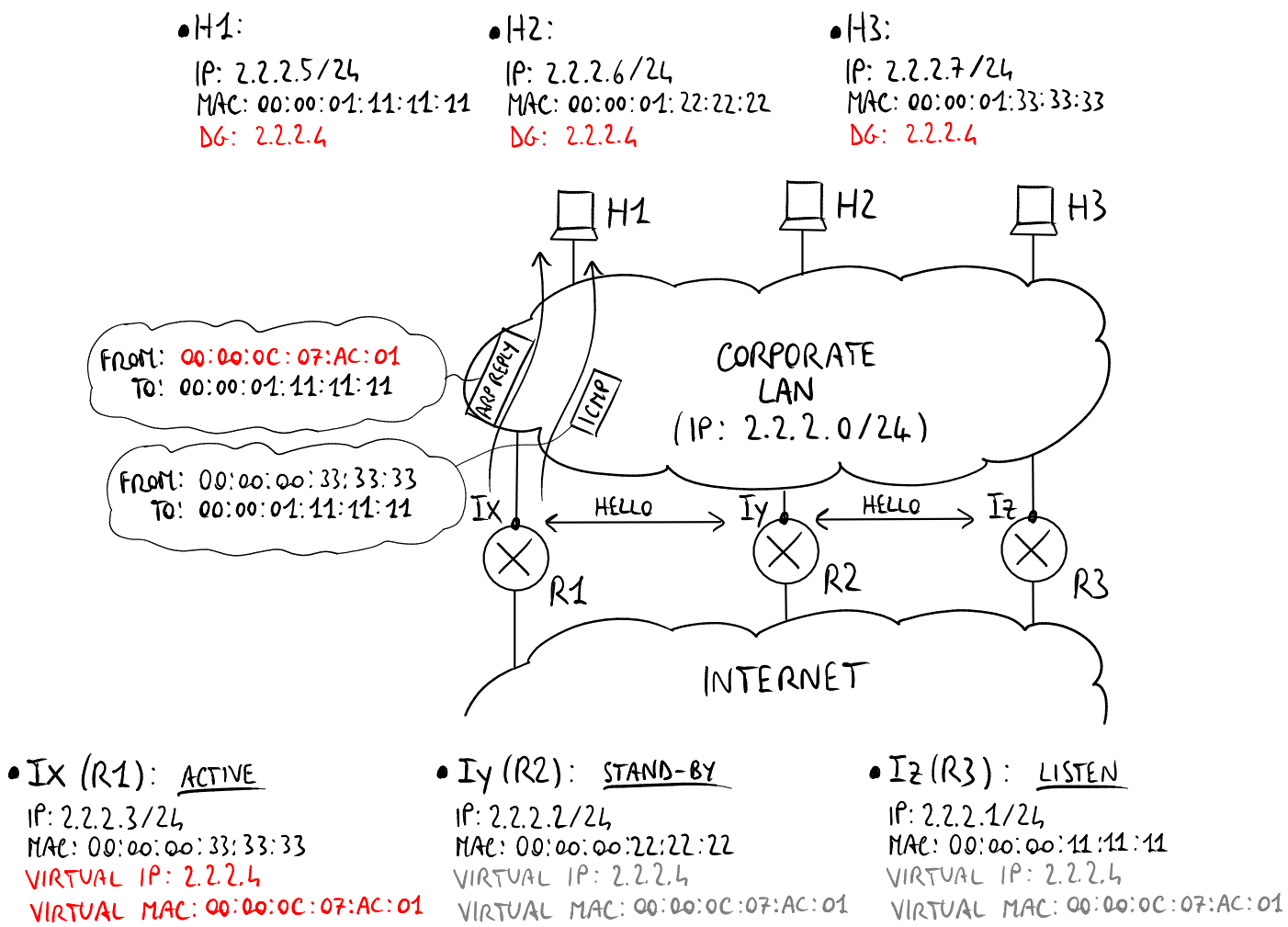


Figura 12.1: Esempio di rete aziendale con tre router ridondanti grazie all'HSRP.

L'indirizzo IP virtuale va impostato esplicitamente dall'amministratore di rete durante la configurazione di HSRP, mentre l'indirizzo MAC virtuale ha il prefisso well-known di Cisco "00:00:0C:07:AC":

24	40	48
OUI (00:00:0C)	stringa HSRP (07:AC)	ID di gruppo

Tabella 12.1: Formato dell'indirizzo MAC virtuale HSRP.

dove i campi sono:

- campo Organizationally Unique Identifier (OUI) (3 byte): la stringa di bit "00:00:0C" è l'OUI assegnato a Cisco affinché gli indirizzi MAC delle schede di rete vendute da Cisco siano globalmente univoci;
- campo stringa HSRP (2 byte): la stringa di bit "07:AC" identifica un indirizzo MAC virtuale HSRP, e non può comparire in alcun indirizzo MAC fisico ⇒ l'indirizzo MAC virtuale è garantito essere univoco all'interno della LAN: non è possibile che un host abbia un indirizzo MAC uguale all'indirizzo MAC virtuale HSRP;
- campo ID di gruppo (1 byte): identifica il gruppo a cui fa riferimento l'istanza di HSRP corrente (si rimanda alla sezione 12.1.4).

A tutti gli host viene impostato l'indirizzo IP virtuale come indirizzo di default gateway, cioè come indirizzo IP a cui gli host invieranno i pacchetti IP diretti al di fuori della LAN.

12.1.2 Instradamento asimmetrico del traffico

L'obiettivo dell'HSRP è "ingannare" l'host facendo credere ad esso di stare comunicando con l'esterno attraverso un singolo router caratterizzato da un indirizzo IP pari all'indirizzo di default gateway e da un indirizzo MAC pari all'indirizzo MAC ottenuto tramite il protocollo ARP, mentre in realtà l'HSRP in caso di guasto sposta il router active su un altro router senza che l'host se ne accorga:

1. ARP Request: quando un host si connette alla rete, invia una ARP Request all'indirizzo IP impostato come default gateway, che è l'indirizzo IP virtuale;[1]
2. ARP Reply: il router active invia in risposta una ARP Reply con il proprio indirizzo MAC virtuale;
3. traffico in uscita: l'host invia ogni pacchetto successivo all'indirizzo MAC virtuale, e solo il router active lo elabora, mentre i router stand-by e listen lo scartano.
   Poi, il router active inoltra il pacchetto secondo i protocolli di instradamento esterni (OSPF, BGP, ecc.) che sono indipendenti dall'HSRP ⇒ il pacchetto potrebbe anche attraversare i router stand-by e listen se i protocolli di instradamento ritengono che questo sia il percorso migliore;
4. traffico in entrata: ogni pacchetto proveniente dall'esterno e diretto all'host può entrare nella LAN da uno qualsiasi dei router ridondanti secondo i protocolli di instradamento esterni indipendenti dall'HSRP, e l'host lo riceverà con l'indirizzo MAC reale del router come indirizzo MAC sorgente.
   I protocolli di instradamento esterni sono anche in grado di rilevare i guasti dei router, compreso il default gateway, per il traffico in entrata ⇒ la protezione è ottenuta anche se la LAN è priva dell'HSRP.

[1] Si ricorda che la ARP Request è una trama di livello data-link con indirizzo MAC di destinazione broadcast e con l'indirizzo IP nel payload.

Server dual-homed

L'HSRP può essere utilizzato per ottenere la ridondanza di una macchina singola per migliorare la tolleranza ai guasti: un server può essere dotato di due interfacce di rete, una primaria e una secondaria, a cui l'HSRP assegna un indirizzo IP virtuale e un indirizzo MAC virtuale ⇒ il server continuerà a essere raggiungibile allo stesso indirizzo IP anche in caso di guasto del link che collega l'interfaccia primaria alla rete.

12.1.3 Pacchetti di Hello

I **pacchetti di Hello** sono dei messaggi generati dai router ridondanti per:

- eleggere il router active: nella fase di negoziazione, i router si scambiano dei pacchetti di Hello proponendosi come router active ⇒ il router active è quello con la priorità più alta (configurabile dall'amministratore di rete), o in caso di parità quello con l'indirizzo IP più alto;
- rilevare i guasti del router active: il router active invia periodicamente pacchetti di Hello come messaggi di "keep-alive" ⇒ in caso di guasto del router active, il router stand-by non riceve più il messaggio di "keep-alive" ed elegge se stesso come router active;
- aggiornare i filtering database: quando il router active cambia, il nuovo router active inizia a inviare dei messaggi di Hello segnalando ai bridge all'interno della LAN aziendale la nuova posizione dell'indirizzo MAC virtuale ⇒ tutti i bridge aggiorneranno i loro filtering database conformemente.
  Quando un router diventa active, invia anche una ARP Reply gratuita in broadcast (le ARP Reply normali sono unicast) con l'indirizzo MAC virtuale come indirizzo MAC sorgente.

Nel pacchetto di Hello l'intestazione HSRP è incapsulata nel formato seguente:

14 byte	20 byte	8 byte	20 byte
intestazione MAC	intestazione IP	intestazione UDP	
src: indirizzo MAC virtuale	src: indirizzo IP reale	src: porta 1985	intestazione HSRP
dst: 01:00:5E:00:00:02	dst: 224.0.0.2	dst: porta 1985	
	TTL: 1		

Tabella 12.2: Formato del pacchetto di Hello HSRP generato dal router active.

14 byte	20 byte	8 byte	20 byte
intestazione MAC	intestazione IP	intestazione UDP	
src: indirizzo MAC reale	src: indirizzo IP reale	src: porta 1985	intestazione HSRP
dst: 01:00:5E:00:00:02	dst: 224.0.0.2	dst: porta 1985	
	TTL: 1		

Tabella 12.3: Formato del pacchetto di Hello HSRP generato dal router stand-by.

Osservazioni

- indirizzo MAC sorgente: è l'indirizzo MAC virtuale per il router active, è l'indirizzo MAC reale per il router stand-by;
- indirizzo IP di destinazione: "224.0.0.2" è l'indirizzo IP del gruppo multicast "all routers"; è uno degli indirizzi multicast non filtrati da IGMP snooping e quindi mandati sempre in flooding dai bridge[2];

[2]Si veda il capitolo 10.

- indirizzo MAC di destinazione: “01:00:5E:00:00:02” è l’indirizzo MAC multicast derivato dall’indirizzo IP multicast;
- Time To Live (TTL): è pari a 1 in modo che i pacchetti vengano scartati subito dai router a cui giungono, perché essi possono essere propagati solo per la LAN;
- l’intestazione HSRP del pacchetto di Hello è incapsulata in UDP e non in TCP perché la perdita di un pacchetto di Hello non richiede la sua ritrasmissione;
- i router listen non generano pacchetti di Hello, a meno che non rilevano che il router stand-by è diventato router active e devono candidarsi affinché uno di essi diventi il nuovo router stand-by.

Formato dell’intestazione HSRP

L’intestazione HSRP ha il formato seguente:

8	16	24	32
Version	Op Code	State	Hello Time
Hold Time	Priority	Group	Reserved
Authentication			
Data			
Virtual IP Address			

Tabella 12.4: Formato dell’intestazione HSRP (20 byte).

dove i campi più significativi sono:

- campo Op Code (1 byte): descrive il tipo di messaggio contenuto nel pacchetto di Hello:
  - 0 = Hello: il router è in esecuzione ed è capace di diventare il router active o stand-by;
  - 1 = Coup: il router vuole diventare il router active;
  - 2 = Resign: il router non vuole più essere il router active;
- campo State (1 byte): descrive lo stato corrente del router mittente del messaggio:
  - 8 = Standby: il pacchetto HSRP è stato inviato dal router stand-by;
  - 16 = Active: il pacchetto HSRP è stato inviato dal router active;
- campo Hello Time (1 byte): è il tempo tra i messaggi di Hello inviati dai router (predefinito: 3 s);
- campo Hold Time (1 byte): è il tempo di validità del messaggio di Hello corrente, scaduto il quale il router stand-by si propone come router active (predefinito: 10 s);
- campo Priority (1 byte): è la priorità del router usata per il processo di elezione del router active/stand-by (predefinita: 100);
- campo Group (1 byte): identifica il gruppo a cui fa riferimento l’istanza di HSRP corrente (si rimanda alla sezione 12.1.4);
- campo Authentication Data (8 byte): contiene una password da 8 caratteri in chiaro (predefinita: “cisco”);
- campo Virtual IP Address (4 byte): è l’indirizzo IP virtuale utilizzato dal gruppo, cioè l’indirizzo IP utilizzato come indirizzo di default gateway dagli host della LAN aziendale.

Con i valori predefiniti per i parametri di Hello Time e di Hold Time, il tempo di convergenza è pari a circa 10 secondi.

12.1.4 Gruppi HSRP

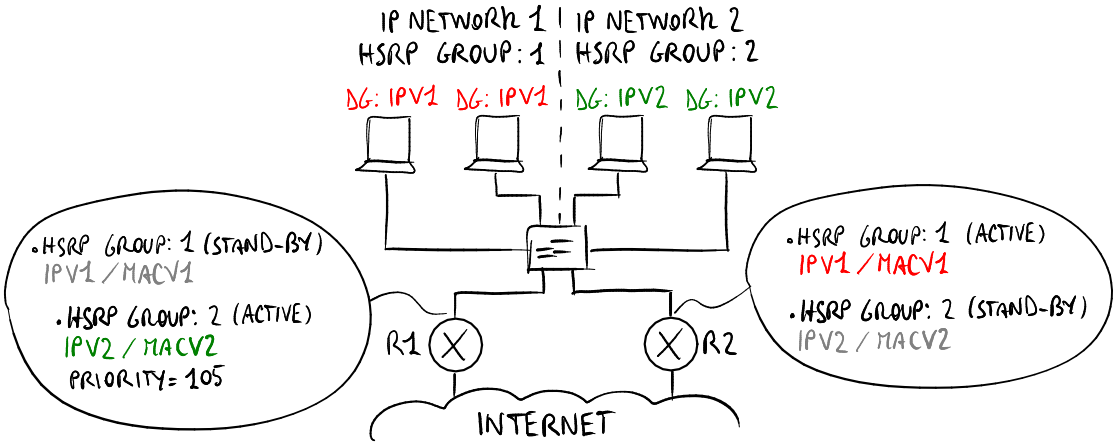


Figura 12.2: Esempio di rete con più gruppi HSRP.

I **gruppi HSRP** permettono di distinguere più reti IP logiche nella stessa LAN fisica: a ogni rete IP corrisponde un gruppo HSRP, con una coppia indirizzo MAC virtuale e indirizzo IP virtuale. Gli host di una rete IP hanno uno degli indirizzi IP virtuali impostato come indirizzo di default gateway, gli host di un'altra rete IP hanno un altro indirizzo IP virtuale impostato come indirizzo di default gateway, e così via.

Ogni router ridondante conosce più coppie indirizzo MAC virtuale e indirizzo IP virtuale, una per ogni gruppo ⇒ ogni router (tranne i router listen) genera un pacchetto di Hello per ogni gruppo, e risponde a uno degli indirizzi MAC virtuali alla ricezione di traffico da host di una rete IP, a un altro alla ricezione di traffico da host di un'altra rete IP, e così via.

Gli ultimi 8 bit dell'indirizzo MAC virtuale identificano il gruppo a cui l'indirizzo fa riferimento ⇒ l'HSRP è in grado di gestire fino a 256 gruppi diversi in una stessa LAN.

In presenza di VLAN

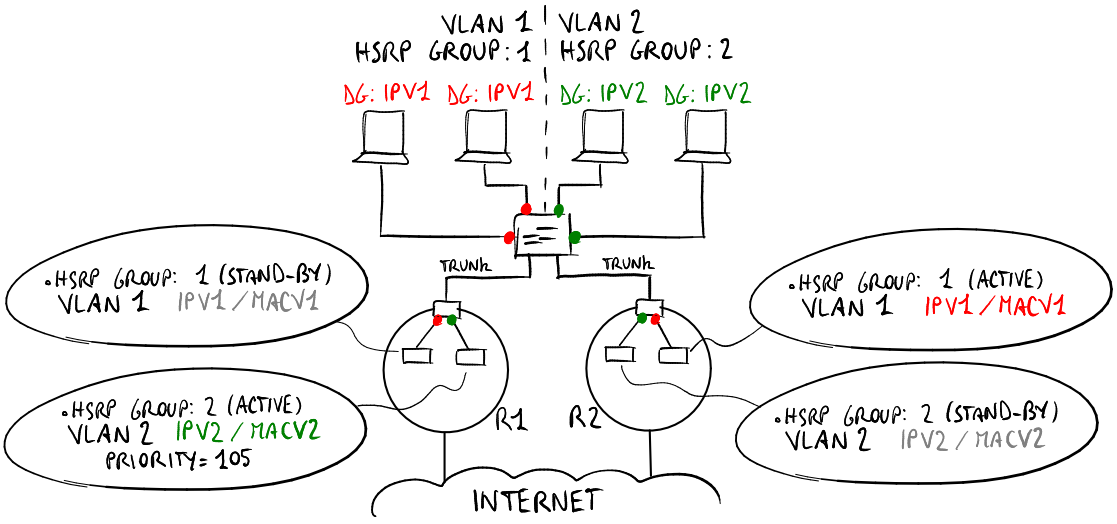


Figura 12.3: Esempio di rete con più gruppi HSRP in presenza di VLAN.

Definire più gruppi HSRP è obbligatorio in presenza di VLAN: ogni VLAN è infatti una LAN separata con il proprio default gateway ⇒ a ogni VLAN è assegnato un gruppo HSRP. Ogni

router a braccio singolo[3] ridondante ha più interfacce virtuali[4], una per ogni VLAN ⇒ i gruppi HSRP sono configurati sulla stessa interfaccia fisica ma ciascuno su interfacce logiche differenti.

Multi-group HSRP (mHSRP)

Tramite un'opportuna configurazione delle priorità, è possibile distribuire il traffico delle reti IP sui router ridondanti (**condivisione del carico**):

- figura 12.2: il traffico della rete IP 1 passa per il router R2, mentre il traffico della rete IP 2 passa per il router R1;
- figura 12.3: il traffico della VLAN 1 passa per il router R2, mentre il traffico della VLAN 2 passa per il router R1.

Vantaggi

- il mHSRP è molto conveniente quando il traffico in entrata nella LAN è simmetrico: un router a braccio singolo per l'interconnessione di VLAN può essere ridondato in modo che un router sostiene il traffico in entrata da una prima VLAN e in uscita in una seconda VLAN, mentre l'altro router sostiene il traffico in entrata dalla seconda VLAN e in uscita nella prima VLAN;
- migliore utilizzo delle risorse: in una rete con un singolo gruppo HSRP la larghezza di banda del router stand-by è del tutto inutilizzata ⇒ il mHSRP permette di utilizzare la larghezza di banda di entrambi i router.

Svantaggi

- il mHSRP non è così conveniente quando il traffico in entrata nella LAN è asimmetrico: la condivisione del carico infatti riguarda solo il traffico in uscita (il traffico in entrata è indipendente dall'HSRP), e il traffico in uscita (upload) generalmente è inferiore rispetto al traffico in entrata (download);
- la condivisione del carico non implica necessariamente il bilanciamento del traffico: il traffico proveniente da una LAN potrebbe essere molto maggiore del traffico proveniente da un'altra LAN;
- difficoltà di configurazione: gli host in ogni rete IP devono avere un indirizzo di default gateway diverso rispetto agli host delle altre reti IP, ma il server DHCP solitamente restituisce un singolo indirizzo di default gateway per tutti gli host.

12.1.5 Funzione di track

L'HSRP offre protezione dai guasti del link che collega il router default gateway alla LAN e dai guasti del router default gateway stesso, ma non dai guasti del link che collega il router default gateway a Internet: un guasto sul link WAN infatti obbliga i pacchetti a essere inviati al router active che a sua volta li invia tutti al router stand-by, invece di andare subito al router stand-by ⇒ ciò non comporta una reale perdita di connettività Internet, ma comporta un overhead aggiuntivo nel processo di inoltro dei pacchetti.

La **funzione di track** consente di rilevare i guasti sui link WAN e di scatenare il router stand-by a prendere il posto del router active tramite una diminuzione automatica della priorità del router active (predefinito: -10).

La funzione di track funziona solo se è attiva la **capacità di prelazione**: se la diminuzione della priorità del router active è tale da portarla al di sotto della priorità del router stand-by,

[3] Si veda la sezione 11.1.
[4] Si veda la sezione 11.3.2.

quest'ultimo può "sottrarre" lo stato di active al router active inviando un messaggio di Hello di tipo Coup.

Tuttavia, il rilevamento dei guasti avviene esclusivamente a livello fisico: la funzione di track non è in grado di rilevare un guasto avvenuto su un link più lontano al di là di un bridge.

12.1.6 Problemi

Resilienza del livello data-link

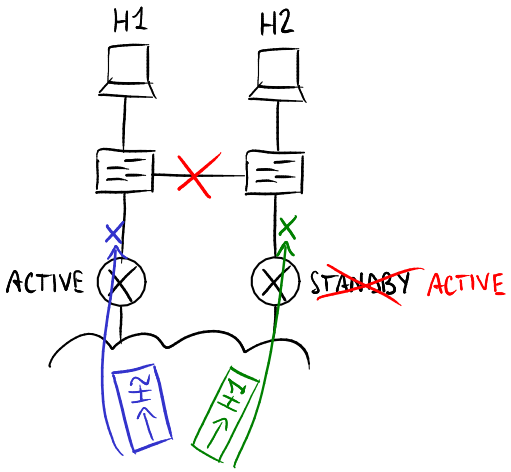


Figura 12.4: Esempio di guasto interno alla rete di livello data-link.

L'HSRP non protegge da tutti i guasti della rete di livello data-link. Per esempio, il guasto in figura 12.4 partiziona la rete aziendale in due parti, e siccome avviene tra due bridge non può essere rilevato dai router a livello fisico. Il router stand-by non riceve più i messaggi di Hello dal router active e si auto-promuove active ⇒ il traffico in uscita non viene influenzato per nulla dal verificarsi del guasto: ognuno dei due router serve il traffico in uscita di una delle due porzioni di rete.

Il guasto ha invece impatto sul traffico in entrata, in quanto alcune trame non possono raggiungere gli host di destinazione. I protocolli di instradamento di livello rete infatti operano esclusivamente tra router e router: si limitano a rilevare che il percorso tra i due router è stato spezzato da qualche parte, ma non sono in grado di rilevare le interruzioni di percorso tra un router e un host, perché il loro compito è inoltrare il pacchetto in modo che raggiunga uno qualsiasi dei router di frontiera, a cui poi spetta la consegna diretta della trama alla destinazione finale. Visti dall'esterno, entrambi i router appaiono avere la connettività alla stessa rete IP, perciò i protocolli di instradamento di livello rete assumeranno che tutti gli host appartenenti a quella rete IP siano raggiungibili da entrambe le interfacce e ne sceglieranno una in base al criterio del percorso più breve:

- se viene scelto il router che serve la porzione di rete a cui appartiene la destinazione, la trama è consegnata a destinazione senza problemi;
- se viene scelto il router che serve l'altra porzione di rete, il router effettua una ARP Request a cui nessun host risponderà e perciò la destinazione apparirà inesistente sulla rete.

È quindi importante ridondare tutti i link interni alla rete di livello data-link, mettendo più link in parallelo gestiti dal protocollo di spanning tree o configurati in link aggregation.

Flooding

In alcune topologie di rete, l'instradamento asimmetrico del traffico può far sì che in alcuni periodi di tempo aumenti in maniera considerevole il traffico in entrata dall'esterno mandato in flooding,

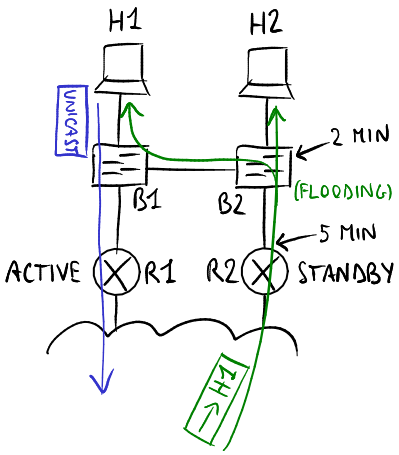


Figura 12.5: Esempio di topologia di rete affetta da flooding periodico.

mentre in altri il traffico in entrata venga inoltrato correttamente dai bridge. Ciò è dovuto al fatto che le ARP cache dei router generalmente durano più a lungo dei filtering database dei bridge.

Per esempio, in figura 12.5 i mapping nella ARP cache sul router ingress R2 scadono dopo 5 minuti, mentre le entry nel filtering database del bridge B2 scadono dopo soli 2 minuti:

1. il pacchetto unicast in uscita aggiorna solo il filtering database del bridge B1, perché non passa attraverso il bridge B2;
2. il pacchetto in entrata scatena l'invio da parte del router R2 di una ARP Request all'host H1 (in broadcast);
3. la ARP Reply che l'host H1 invia in risposta aggiorna sia la ARP cache del router R2 sia il filtering database del bridge B2;
4. per i primi 2 minuti, i pacchetti in entrata diretti all'host H1 vengono inoltrati senza problemi;
5. dopo 2 minuti dalla ARP Reply, scade la entry relativa all'host H1 nel filtering database del bridge B2;
6. per i successivi 3 minuti, il router R2, che ha ancora un mapping valido per l'host H1 nella sua ARP cache, invia i pacchetti in entrata verso il bridge B2, il quale li manda tutti in flooding perché non riceve alcuna trama avente l'host H1 come sorgente;
7. dopo 5 minuti dalla ARP Reply, scade anche il mapping nella ARP cache del router R2;
8. la successiva ARP Reply sollecitata dal router R2 finalmente aggiorna anche il filtering database del bridge R2.

Soluzioni possibili Non è facile identificare questo problema nella rete perché si manifesta in maniera intermittente; una volta identificato il problema, è possibile:

- forzare le stazioni a inviare trame in broadcast gratuite più spesso, con una frequenza minore dell'ageing time delle entry nei filtering database dei bridge;
- aumentare il valore di ageing time sui bridge lungo il percorso ingress ad almeno il tempo di durata delle ARP cache dei router.

Link unidirezionali

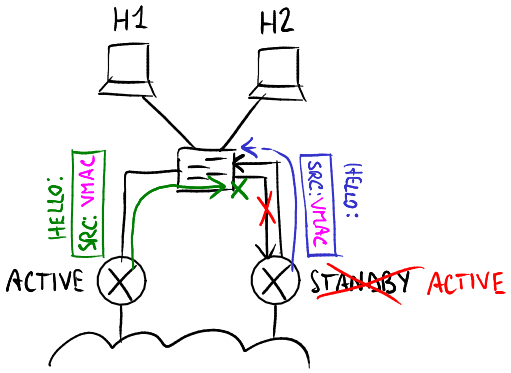


Figura 12.6: Esempio di guasto su un link unidirezionale.

L'HSRP non prevede la gestione di guasti su link unidirezionali: per esempio, in figura 12.6 avviene un guasto verso il router stand-by, il quale non riceve più i messaggi di Hello dal router active e si auto-elegge active, iniziando a inviare pacchetti di Hello con l'indirizzo MAC virtuale come indirizzo sorgente ⇒ il bridge riceve alternativamente pacchetti di Hello da entrambi i router active aventi lo stesso indirizzo MAC come indirizzo sorgente, e la entry relativa a quell'indirizzo MAC continuerà a oscillare periodicamente ⇒ se un host invia una trama al default gateway mentre la entry nel filtering database del bridge è associata all'ex-router stand-by, la trama, non potendo passare per il link unidirezionale guasto, andrà persa.

12.2 GLBP[5]

Il **Gateway Load Balancing Protocol** (GLBP) aggiunge alla ridondanza del default gateway la possibilità di distribuire automaticamente il traffico in uscita su tutti i router ridondanti.

Il GLBP elegge un Active Virtual Gateway (AVG) per ogni gruppo; gli altri membri del gruppo, chiamati Actual Virtual Forwarder (AVF), fungono da backup in caso di guasto dell'AVG. L'AVG eletto assegna quindi un indirizzo MAC virtuale a ogni membro del gruppo GLBP, compreso se stesso; ogni AVF si assume la responsabilità di inoltrare i pacchetti inviati al suo indirizzo MAC virtuale.

In caso di guasto di un AVF, l'AVG notifica uno degli AVF rimasti attivi affidando ad esso il compito di rispondere anche al traffico diretto verso l'indirizzo MAC virtuale dell'AVF guasto.

L'AVG risponde alle ARP Request inviate dagli host con indirizzi MAC che puntano a router diversi, in base a uno dei seguenti algoritmi di bilanciamento del carico:

- nessuno: l'AVG è l'unico forwarder (come nell'HSRP);
- weighted: a ogni router è assegnato un peso, che determina la percentuale di ARP Request risposte con l'indirizzo MAC virtuale di quel router, e quindi la percentuale di host che useranno quel router come forwarder ⇒ utile quando i link di uscita hanno differenti capacità;
- round robin: gli indirizzi MAC virtuali sono selezionati sequenzialmente in una coda circolare;
- host dependent: garantisce che un host rimanga associato sempre allo stesso forwarder, cioè se l'host effettua due ARP Request riceverà due ARP Reply con lo stesso indirizzo MAC virtuale ⇒ ciò evita problemi con i meccanismi di traduzione degli indirizzi dei NAT.

[5]Questa sezione contiene contenuti CC BY-SA dalla voce Gateway Load Balancing Protocol su Wikipedia in inglese.

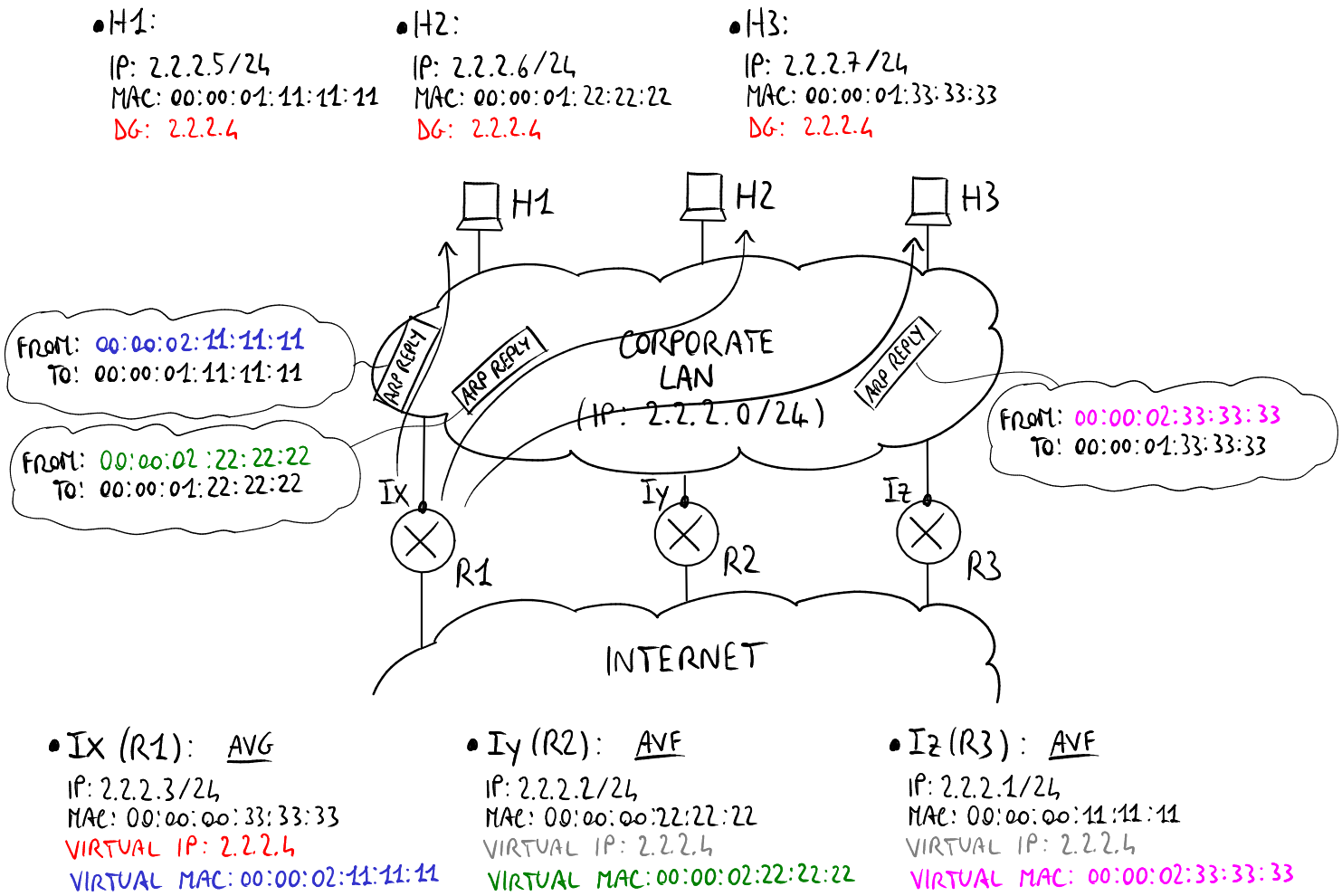


Figura 12.7: Esempio di rete aziendale con tre router ridondanti grazie al GLBP.

Capitolo 13

Il livello rete nelle LAN

I router sono una parte fondamentale di una LAN perché forniscono l'accesso a Internet e l'interconnessione di VLAN.

Vantaggi del livello data-link

- mobilità (sezione 3.2.2)
- trasparenza (sezione 3.2.2)
- algoritmi di inoltro semplici e veloci

Vantaggi del livello rete

- scalabilità (sezioni 3.2.4, 2.3.4)
- sicurezza e isolamento della rete (sezione 3.2.4)
- algoritmi di inoltro efficienti: indirizzamento gerarchico, alberi di inoltro multipli
- no recupero dai guasti lento a causa dello STP (sezione 6.6.1)

13.1 Evoluzioni degli apparati di interconnessione

13.1.1 Layer 3 switch

In una rete aziendale il router costituisce un collo di bottiglia per l'accesso a Internet e per l'interconnessione delle VLAN, perché implementa degli algoritmi complessi che girano su una CPU.

Il **layer 3 switch** è un router realizzato puramente in hardware per migliorare le prestazioni. La sua fabbricazione è meno costosa rispetto ai router tradizionali, ma manca di alcune funzionalità avanzate:

- nessun protocollo di instradamento sofisticato (ad es. BGP);
- insieme limitato di interfacce di rete;
- nessuna possibilità di applicare patch e aggiornamenti (ad es. supporto a IPv6, correzioni di bug, ecc.);
- nessuna funzionalità di protezione (ad es. firewall).

13.1.2 Multilayer switch

Il **multilayer switch** è un apparato che integra le capacità sia L2 sia L3 sulla stessa scheda hardware: il cliente può comprare un multilayer switch e poi configurare ogni interfaccia in

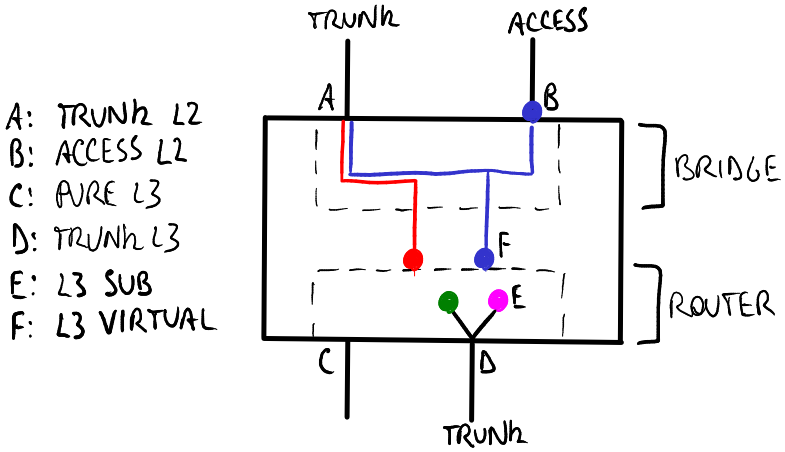


Figura 13.1: Esempio di multilayer switch.

modalità L2 o L3 a seconda delle sue necessità, per una maggiore flessibilità nella distribuzione della rete.

Su un multilayer switch possono essere configurate quattro tipi di interfacce:

- interfacce fisiche L2: in modalità trunk (A) o access (B);
- interfacce fisiche L3: possono terminare link L3 puro (C) o in modalità trunk (D);
- interfacce logiche per l'interconnessione delle VLAN:
  - sotto-interfacce L3 (E): un'interfaccia fisica L3 può suddividersi in più sotto-interfacce L3, una per ogni VLAN;
  - interfacce virtuali L3 (F): connettono il router interno con il bridge interno, una per ogni VLAN.

L'interconnessione di due VLAN tramite un router a braccio singolo richiede al traffico di attraversare due volte il link trunk verso il router ⇒ il multilayer switch, grazie all'integrazione delle funzionalità di instradamento e di commutazione, virtualizza il braccio singolo in modo che il traffico entri con il tag di una VLAN ed esca direttamente con il tag di un'altra VLAN (anche dalla stessa porta in cui era entrato), senza che il carico su un link sia raddoppiato:

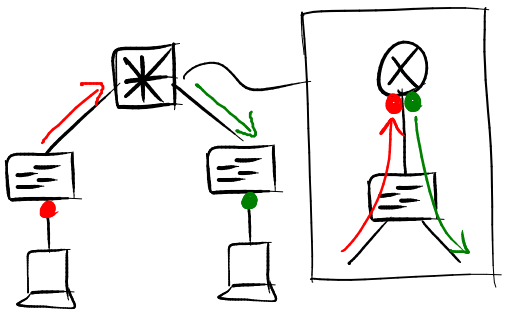
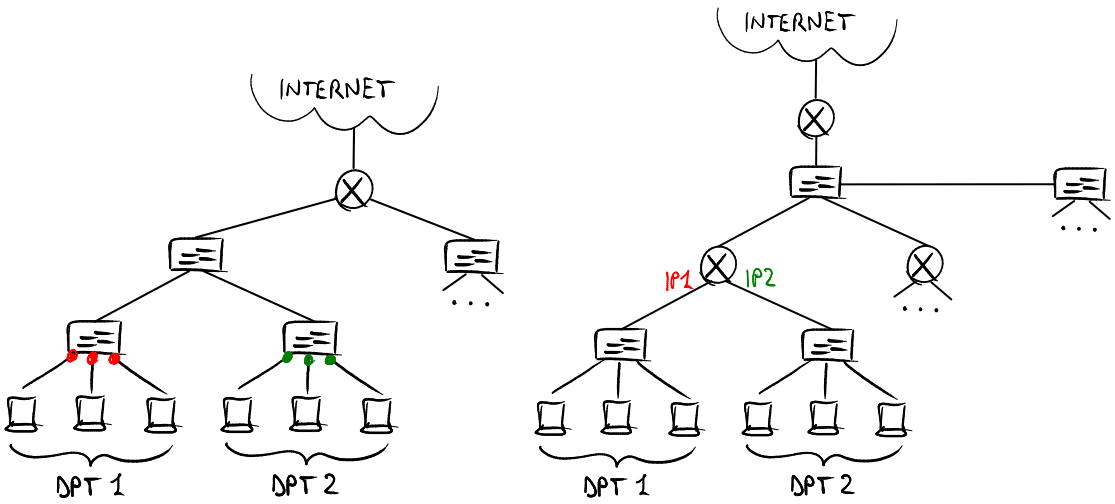


Figura 13.2: Il multilayer switch ottimizza il router a braccio singolo.

13.2 Posizionamento degli apparati di interconnessione

Dove è meglio posizionare i router in una rete aziendale?

- accesso: solo bridge (tipicamente multilayer switch) collegati direttamente agli host;



(a) Backbone con segmentazione VLAN. *(b) Backbone con segmentazione IP.*

Figura 13.3: Esempi di posizionamento dei router all'interno di una rete aziendale.

- backbone: esistono due soluzioni possibili:
  - segmentazione VLAN: l'intera rete aziendale è a livello data-link, e ad ogni zona (ad es. dipartimento dell'università) è assegnata una VLAN ⇒ la mobilità è estesa all'intera rete aziendale;
  - segmentazione IP: ogni bridge di accesso è collegato a un router (tipicamente layer 3 switch), e ad ogni zona è assegnata una rete IP ⇒ maggiore isolamento della rete e maggiore scalabilità.
    Spesso dei bridge interni connettono tutti i router di accesso tra loro e con il router gateway di uscita;
- edge: un router come gateway di uscita verso Internet, di solito un multilayer switch L4-7 dotato di funzioni a livello trasporto e superiori, come protezione (ad es. firewall), qualità del servizio, bilanciamento del carico, ecc.

13.3 Esempio di progettazione di LAN

- multilayer switch all'edge:
  - con dei semplici router ci sarebbe una rete IP diversa per ogni piano, a svantaggio della mobilità tra i piani;
  - sul router interno sono configurate tante interfacce virtuali quante sono le VLAN nell'edificio;
  - tutte le porte verso i bridge di piano sono configurate in modalità trunk, quindi ogni porta può accettare qualsiasi VLAN, a vantaggio della mobilità tra i piani;
  - è un multilayer switch L4-7 per le funzioni dei livelli superiori (in particolare le funzioni di sicurezza);
- traffico tra i router edge: una VLAN aggiuntiva è appositamente dedicata al traffico L3 che i router si scambiano tra loro (ad es. messaggi OSPF, messaggi HSRP), per separarlo dal traffico normale della LAN (altrimenti un host potrebbe fingersi router ed intercettare il traffico tra i router);

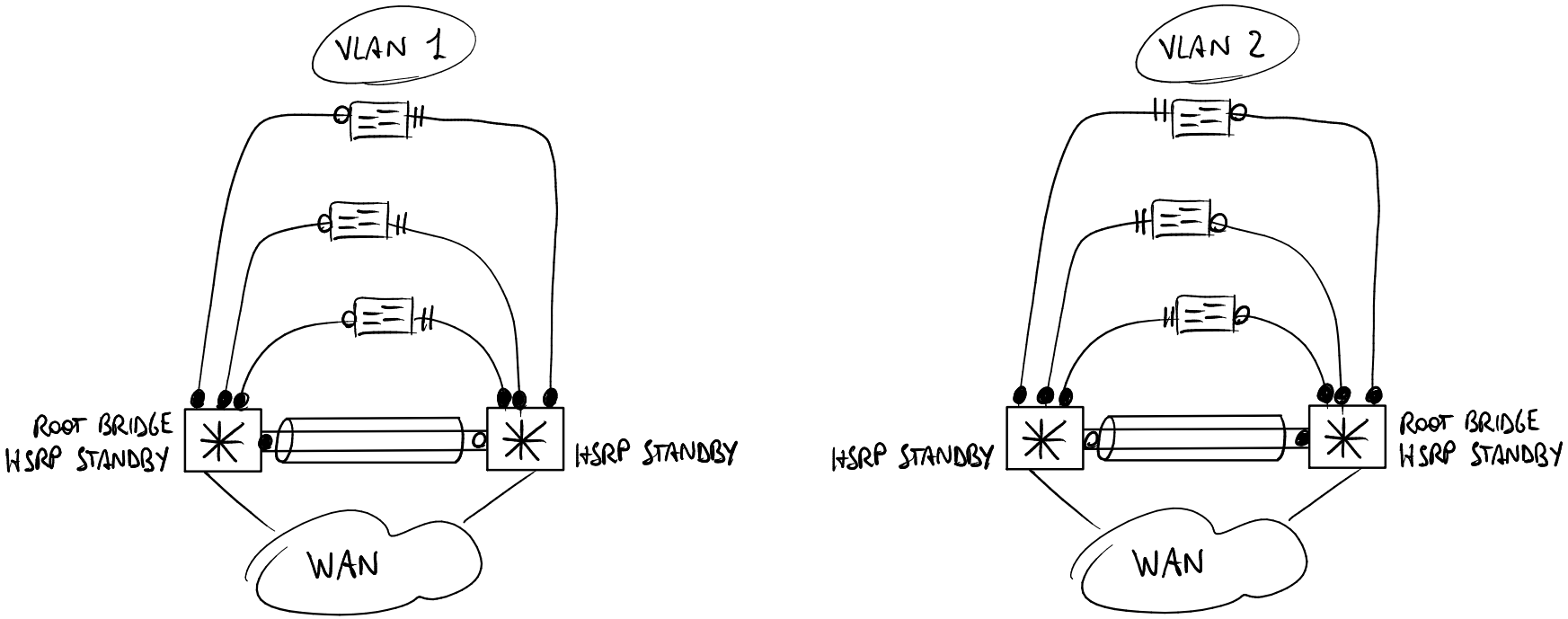


Figura 13.4: Esempio di progettazione di LAN.

- Multi-group HSRP (mHSRP): un multilayer switch può essere active per alcune VLAN e stand-by per altre;
- Per-VLAN Spanning Tree (PVST): un'istanza di protocollo di spanning tree è attiva per ogni VLAN, per ottimizzare i percorsi in base alla VLAN.
  Il root bridge deve essere sempre il router HSRP active, altrimenti alcuni percorsi non sono ottimizzati;
- link diretto tra i multilayer switch:
  - fornisce un percorso diretto per il traffico addizionale L3 tra i router;
  - alleggerisce il carico di traffico sui bridge di piano, che tipicamente sono dimensionati per supportare poco traffico;
  - le porte alle estremità del link sono configurate come porte L2, per dare la possibilità anche al traffico normale di attraversare questo link in caso di guasto di uno dei link a un bridge di piano;
  - è raddoppiato in link aggregation per una maggiore tolleranza ai guasti e per sfruttare la banda a disposizione su entrambi i link (evitando che lo STP disattivi uno dei due link).

Parte V

Argomenti aggiuntivi

Capitolo 14

Introduzione alle Storage Area Network

14.1 Architetture di archiviazione

Una società ha tipicamente bisogno di archiviare molti dati:

- **mainframe** (storico): l'accesso ai dati è centralizzato nella stessa macchina in cui sono fisicamente memorizzati;
- **modello client-server**: i vari client chiedono a una macchina server di recuperare dei dati memorizzati su dischi fissi;
- **modello peer-to-peer**: i dati sono distribuiti tra tutte le macchine collegate tra loro, e ogni macchina può chiedere a ogni altra macchina di avere dei dati.

Confronto

- costi: ogni macchina nella rete peer-to-peer non richiede un'elevata potenza di calcolo e un'elevata capacità di archiviazione, a differenza di un server che deve gestire richieste da più client allo stesso tempo ⇒ i server sono molto costosi;
- scalabilità: nel modello peer-to-peer i dati possono essere distribuiti su un numero illimitato di macchine, mentre la potenza di calcolo e la capacità di archiviazione di un server sono limitate;
- robustezza: un server è caratterizzato da un'elevata affidabilità, ma un guasto è più critico da risolvere; le macchine nella rete peer-to-peer invece sono molto più soggette a guasti perché sono macchine low-end e meno affidabili, ma il software che gestisce la rete peer-to-peer, consapevole di questa debolezza, è progettato per mantenere l'integrità dei dati, effettuando ad esempio dei backup automatici.

Datacenter Un **datacenter** è un luogo centralizzato in cui sono concentrati tutti i server, e permette di evitare di avere troppi server sparsi in giro per la società sotto il controllo di così tante organizzazioni diverse:

- accesso ai dati: i dati potrebbero essere disponibili, ma le persone che ne hanno bisogno potrebbero appartenere a un'altra organizzazione o potrebbero non disporre dei permessi necessari;
- integrità: è difficile effettuare i backup di tutti i server se questi sono sparsi in giro per la società;
- sicurezza: è facile rubare un disco fisso da un server non protetto.

14.2 DAS

In un sistema **Direct-Attached Storage** (DAS), ogni server ha accesso esclusivo al proprio insieme di dischi fissi:

- dischi interni: è una soluzione non adatta per i server in quanto, in caso di guasto, è necessario estrarre fisicamente i dischi fissi dall'interno della macchina;
- dischi esterni: i dischi sono collegati al server tramite SCSI; più insiemi di dischi possono essere collegati in cascata come un'architettura a bus.
  I dischi possono essere messi in un armadietto apposito chiamato **Just a Bunch Of Disks** (JBOD): il controller SCSI è in grado di esportare una struttura di unità virtuale che è differente da quella dei dischi fisici, aggregando o suddividendo le capacità dei dischi e fornendo servizi avanzati (ad es. RAID).

Lo standard **Small Computer System Interface** (SCSI) definisce una pila protocollare completa:

- interfacce fisiche (ad es. cavi e connettori): permettono di collegare fisicamente i dischi fissi ai server;
- protocolli: permettono di effettuare transazioni di lettura e scrittura indirizzando direttamente i blocchi su disco secondo lo schema Logical Block Addressing (LBA);
- comandi esportati alle applicazioni: permettono di effettuare operazioni di lettura e scrittura impartendo dei comandi del tipo READ, WRITE, FORMAT, ecc.

Vantaggi

- bassa latenza: è dell'ordine dei millisecondi attraverso un disco e dei microsecondi attraverso una cache;
- alta affidabilità: la probabilità di errore è molto bassa, e l'integrità dei dati è sempre garantita;
- ampia compatibilità: è ampiamente supportato dai sistemi operativi ed è utilizzato da molti dispositivi esterni oltre ai dischi.

Svantaggi

- recupero dagli errori lento: siccome gli errori capitano raramente, i meccanismi di recupero dagli errori non sono particolarmente efficienti dal punto di vista prestazionale;
- accesso ai dischi centralizzato: solo il server può accedere ai dischi $\Rightarrow$ in caso di guasto del server, i dischi non sono più accessibili;
- limiti di scalabilità: possono essere collegati in cascata al massimo 16 dispositivi per una lunghezza massima di 25 metri.

I NAS (sezione 14.3) e le SAN (sezione 14.4) permettono di disaccoppiare i dischi dai server connettendo tali entità tramite una rete $\Rightarrow$ a un disco possono accedere più server.

14.3 NAS

Un **Network-Attached Storage** (NAS) esporta file system, servendo file logici, anziché blocchi su disco, sulla rete (di solito LAN).

I file system sono condivisi con dei client di rete: sia i server sia i client collegati alla rete possono accedere ai file.

I protocolli tipici utilizzati per esportare i file system sono:

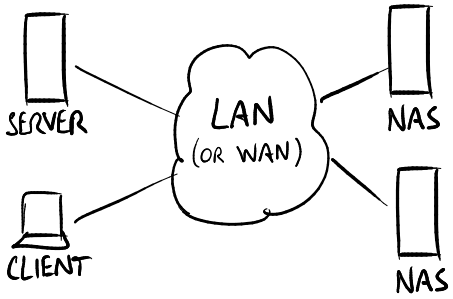


Figura 14.1: Esempio di NAS.

- **Network File System** (NFS): popolare sui sistemi UNIX;
- **Common Internet File System** (CIFS): usato dai sistemi Windows;

che operano su una rete TCP/IP:

NFS/CIFS
TCP
IP
Ethernet

Tabella 14.1: Pila protocollare NAS.

Vantaggi

- insieme al file system, possono essere esportati i permessi di utente e le protezioni di accesso (ad es. nome utente e password);
- compatibilità con i client di rete: un sistema NAS ha un impatto minimale sull'infrastruttura esistente: tutti i sistemi operativi sono in grado di montare un disco condiviso senza driver aggiuntivi.

Svantaggi

- compatibilità con le applicazioni: il disco grezzo è invisibile al client: i dischi non possono essere formattati né gestiti a livello di blocchi ⇒ alcune applicazioni che hanno bisogno di accedere direttamente ai blocchi su disco non possono funzionare sui dischi remoti: sistemi operativi, sistemi per la gestione delle basi di dati, file/partizioni di swap;
- l'appliance NAS richiede una potenza di calcolo sufficiente per la gestione dei permessi di utente e per la rimappatura delle richieste relative a file in richieste relative a blocchi;
- la pila protocollare non è progettata per i NAS: i meccanismi di recupero dagli errori del TCP possono introdurre un overhead prestazionale non trascurabile.

14.4 SAN

Una **Storage Area Network** (SAN) esporta dischi fisici, anziché volumi logici, e permette di indirizzare i blocchi su disco secondo lo schema LBA, proprio come se il disco fosse collegato direttamente al server tramite SCSI (sistema DAS).

I client possono accedere ai dati solo attraverso i server, a cui sono collegati tramite una rete locale o geografica. Tipicamente un datacenter segue un **modello a tre livelli**:

1. server Web: è il front-end esposto ai client;

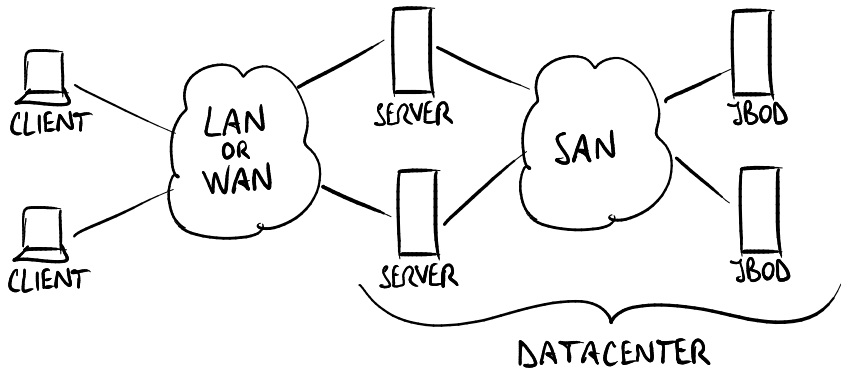


Figura 14.2: Esempio di SAN a due livelli.

2. application/database server: può montare un file system a disco condiviso che converte le richieste relative a file da parte dei client in richieste relative a blocchi da inviare ai dischi remoti tramite la SAN;

3. dischi fissi: sono spesso tenuti nei JBOD.

Le SAN non possono basarsi esclusivamente sul classico TCP/IP, poiché i meccanismi di recupero dagli errori del TCP possono introdurre un overhead prestazionale non trascurabile ⇒ sono stati progettati dei protocolli per le SAN volti a mantenere il più possibile l'alta velocità, la bassa latenza e l'elevata affidabilità tipiche di SCSI:

SCSI
Fibre Channel

(a) Fibre Channel (14.4.1)

SCSI
Fibre Channel
FCoE
10 Gigabit Ethernet

(b) FCoE (14.4.2)

SCSI
iSCSI
TCP
IP
Ethernet

(c) iSCSI (14.4.3)

SCSI
Fibre Channel
FCIP
TCP
IP
Ethernet

(d) FCIP (14.4.4)

Tabella 14.2: Pile protocollari SAN.

Tutti i protocolli per le SAN adottano SCSI come il livello superiore nella pila protocollare e operano al di sotto di esso ⇒ ciò garantisce la compatibilità con tutte le applicazioni basate su SCSI esistenti, con un impatto minimo per la migrazione dai DAS alle SAN.

14.4.1 Fibre Channel

Lo standard **Fibre Channel** è nato dalla necessità di avere un supporto affidabile per i collegamenti in fibra ottica tra i server e i dischi di archiviazione, ed è pensato per rimpiazzare il livello fisico di SCSI. Fibre Channel supporta elevate velocità di trasferimento: 1 Gbps, 2 Gbps, 4 Gbps, 8 Gbps, 16 Gbps.

Topologie

Lo standard prevede tre possibili topologie per le SAN:

- punto punto: connessione diretta tra un server e un JBOD, come in SCSI;
- arbitrated loop: topologia ad anello a scopo di affidabilità;

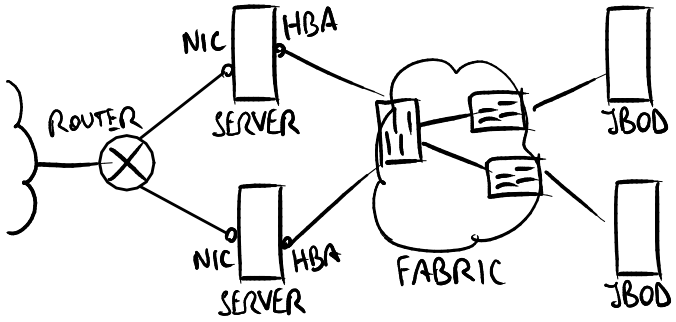


Figura 14.3: Esempio di SAN basata su Fibre Channel con topologia switched fabric.

- switched fabric: più server sono collegati a più JBOD attraverso una **fabbrica**, cioè una rete magliata di bridge.
  La topologia switched fabric è nuova nel mondo dell'archiviazione: SCSI permetteva solo un collegamento a cascata come un'architettura a bus.

Instradamento

L'instradamento è svolto dal protocollo **Fabric Shortest Path First** (FSPF), molto simile al protocollo OSPF delle reti IP. Non è previsto alcun protocollo di spanning tree per gli anelli in topologia.

A ogni porta di un nodo Fibre Channel (server o JBOD) è assegnato dinamicamente un indirizzo da 24 bit:

8	16	24
Domain ID	Area ID	Port ID

dove i campi sono:

- campo Domain ID (8 bit): identifica il bridge a cui è collegato il nodo;
- campo Area ID (8 bit): identifica il gruppo di porte a cui appartiene la porta del bridge a cui è collegato il nodo;
- campo Port ID (8 bit): identifica la porta del nodo.
  Ogni server è collegato alla fabbrica tramite un'interfaccia chiamata **Host Bus Adapter** (HBA).

Controllo di flusso

Fibre Channel migliora i meccanismi di recupero dagli errori di SCSI introducendo un controllo di flusso hop-by-hop basato su un **meccanismo a crediti**: ogni porta ha un numero di crediti, che viene decrementato ogni volta che viene inoltrato un pacchetto e viene incrementato ogni volta che viene ricevuto un acknowledge ⇒ se il numero di crediti a disposizione scende a 0, la porta non può inviare altri pacchetti e deve aspettare che l'hop successivo comunichi tramite un acknowledge che è pronto a ricevere altri dati nel suo buffer ⇒ questo meccanismo evita la congestione dei buffer dei nodi e quindi la perdita di pacchetti.

Il meccanismo a crediti inoltre permette la prenotazione delle risorse e garantisce la consegna in-order delle trame: il nodo di destinazione non ha bisogno di implementare un meccanismo per il riordino dei pacchetti (come nel TCP).

Problemi

- il traffico di un link può essere bloccato per un certo tempo per la mancanza di crediti ⇒ occorre impostare in modo appropriato il massimo numero di crediti di una porta in base alla capacità del buffer della porta che sta all'altra estremità del link;
- si possono verificare deadlock in una rete magliata con dipendenze cicliche.

Funzionalità avanzate

- Virtual SAN (VSAN): l'equivalente delle VLAN per le SAN;
- link aggregation;
- bilanciamento del carico;
- virtualizzazione: le funzionalità di virtualizzazione del controller SCSI possono essere spostate direttamente sul bridge a cui è collegato il JBOD.

14.4.2 FCoE[1]

La tecnologia **Fibre Channel over Ethernet** (FCoE) permette di incapsulare le trame Fibre Channel in trame Ethernet tramite lo strato di adattamento FCoE, che rimpiazza il livello fisico di Fibre Channel ⇒ ciò consente di utilizzare le reti 10 Gigabit Ethernet (o velocità più elevate) preservando il protocollo Fibre Channel.

Prima di FCoE, i datacenter utilizzavano Ethernet per le reti TCP/IP e Fibre Channel per le SAN. Con FCoE, Fibre Channel diventa un altro protocollo di rete operante su Ethernet, insieme al tradizionale traffico IP: FCoE opera direttamente al di sopra di Ethernet nella pila protocollare di rete, a differenza di iSCSI che opera in cima a TCP e IP:

- vantaggio: il server non deve più avere un'interfaccia HBA specifica per Fibre Channel, ma una singola interfaccia NIC può fornire la connettività sia alla SAN sia a Internet ⇒ minor numero di cavi e di bridge e minor consumo di energia;
- svantaggio: FCoE non è instradabile a livello IP, cioè non può andare sulla rete Internet al di fuori della SAN.

Siccome, a differenza di Fibre Channel, l'Ethernet classico non include alcun meccanismo di controllo di flusso, FCoE ha richiesto dei miglioramenti allo standard Ethernet per supportare un meccanismo per il controllo di flusso basato sulle priorità, per ridurre le perdite di trame da congestione.

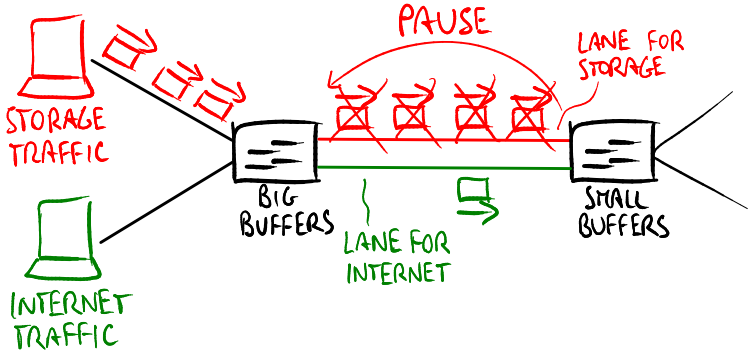


Figura 14.4: Controllo di flusso basato sulle priorità in FCoE.

L'idea di base è adottare i pacchetti PAUSE dallo standard 802.3x per il controllo di flusso su Ethernet[2], ma il canale Ethernet tra i due bridge è partizionato logicamente in **lane** (per

[1] Questa sezione contiene contenuti CC BY-SA dalla voce Fibre Channel over Ethernet su Wikipedia in inglese.
[2] Si veda la sezione 8.2.

esempio, una dedicata al traffico di archiviazione e un'altra dedicata al normale traffico Internet) ⇒ il pacchetto PAUSE, anziché bloccare tutto il traffico sul link interessato, blocca solo il traffico di una certa lane senza influenzare il traffico delle altre lane.

Tipicamente per i server con tecnologia FCoE si prediligono gli switch top of the rack (TOR) agli switch end of the row (EOR) usati con Fibre Channel, perché gli switch con tecnologia FCoE sono meno costosi rispetto agli switch con tecnologia Fibre Channel:

- switch end of the row: c'è un singolo switch principale e ogni server è collegato ad esso con un proprio cavo ⇒ cavi più lunghi;
- switch top of the rack: in cima a ogni armadio c'è uno switch, e ogni server è collegato allo switch dell'armadio, poi tutti gli switch degli armadi sono collegati allo switch principale ⇒ switch più numerosi, ma cavi più corti.

14.4.3 iSCSI

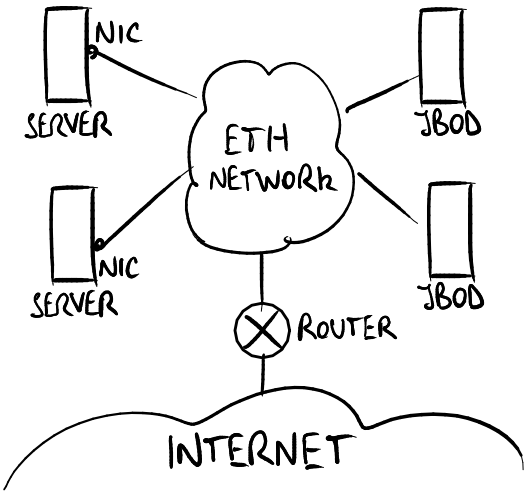


Figura 14.5: Esempio di SAN basata su iSCSI.

Il protocollo **Internet Small Computer System Interface** (iSCSI), proposto da Cisco per contrastare l'egemonia di Fibre Channel, permette di realizzare una SAN sfruttando la tecnologia di rete più diffusa, ovvero il TCP/IP: i comandi SCSI vengono incapsulati in pacchetti TCP tramite lo strato di adattamento iSCSI e attraversano la SAN su una rete Ethernet.

Vantaggi

- il server non deve più avere un'interfaccia HBA specifica per Fibre Channel, ma una singola interfaccia NIC può fornire la connettività sia alla SAN sia a Internet ⇒ minor numero di cavi e di bridge, e minor consumo di energia;
- i dischi sono raggiungibili anche dai client tramite Internet;
- non è necessario posare delle fibre ottiche appositamente adibite al collegamento della SAN.

Svantaggi

- è necessario dimensionare i buffer dei bridge nella SAN in modo da minimizzare le perdite di pacchetti per overflow dei buffer e quindi l'overhead prestazionale dovuto ai meccanismi per il recupero dagli errori del TCP;

- la tecnologia Ethernet non è molto conosciuta nel mondo dell'archiviazione, dove si è abituati a usare gli strumenti Fibre Channel ⇒ il protocollo iSCSI non ha avuto molto successo.

14.4.4 FCIP

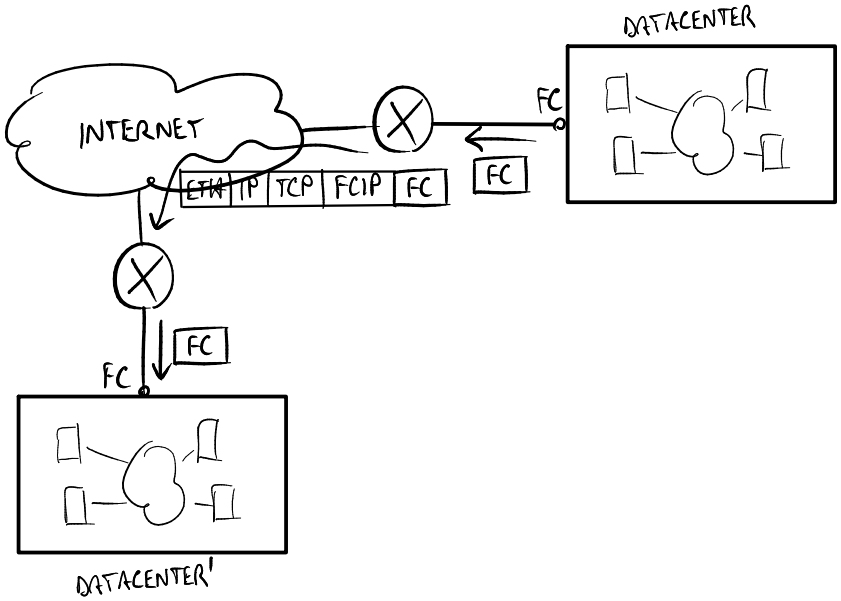


Figura 14.6: Esempio di SAN basata su FCIP.

Un datacenter è soggetto al rischio di perdita dei dati a causa di calamità naturali (quali terremoti, tsunami, ecc.) ⇒ al fine di migliorare la resilienza (continuità operativa), il datacenter può essere interamente replicato in un altro luogo, generalmente a qualche centinaio di chilometri di distanza. Il datacenter principale e il datacenter di backup potrebbero comunicare tra loro direttamente utilizzando Fibre Channel, ma sarebbe troppo costoso collegarli con una semplice fibra ottica a causa della lunga distanza.

La tecnologia **Fibre Channel over IP** (FCIP) consente a SAN geograficamente distribuite di essere interconnesse utilizzando l'infrastruttura TCP/IP esistente, ovvero Internet, senza che gli apparati interni ai datacenter siano consapevoli della presenza della rete IP:

1. il datacenter principale invia una trama Fibre Channel;
2. il router di frontiera incapsula la trama Fibre Channel in un pacchetto TCP, tramite lo strato di adattamento FCIP che rimpiazza il livello fisico di Fibre Channel, poi inoltra il pacchetto TCP sulla rete Internet, in una sorta di tunnel, fino all'altro router di frontiera;
3. l'altro router di frontiera estrae la trama Fibre Channel e la invia al datacenter di backup;
4. il datacenter di backup riceve la trama Fibre Channel.

La dimensione minima delle trame Fibre Channel, tuttavia, supera il limite di dimensione del payload Ethernet, e l'overhead per la frammentazione sarebbe eccessivo ⇒ le trame Ethernet devono essere estese a circa 2,2 KB affinché possano essere incapsulate trame Fibre Channel di dimensione minima.