Università degli Studi di Roma "La Sapienza"
Dipartimento INFOCOM

Aldo Roveri

RETEMATICA

volume primo:

funzioni e servizi di rete

Versione intermedia
ottobre 2001

I QUADRO DI RIFERIMENTO

Una comunicazione a distanza avviene solitamente sulla base di un rapporto di domanda e di offerta. Oggetto del rapporto è un *servizio di telecomunicazione*. Soggetti del rapporto sono il *cliente* (service customer), il *fornitore* (service provider) e il *gestore di rete* (network operator).

Il cliente del servizio ha l'esigenza di comunicare a distanza con altri clienti, secondo *modalità operative* e *prestazioni di qualità* definiti all'interno di un opportuno *quadro contrattuale*. Per fruire di un servizio deve interagire con il fornitore. Delega alla fruizione uno o più *utenti* (operatori umani o macchine di elaborazione). Gli utenti agiscono come *sorgenti* e/o *collettori* di informazione.

La fruizione di un servizio comporta il richiamo e l'esecuzione, in un ordine prestabilito, di *componenti funzionali,* che qualificano il servizio nei suoi vari aspetti e che debbono svolgersi rispettando opportune regole. Queste costituiscono la *logica del servizio.* Tra le componenti funzionali è sempre incluso un *trasferimento* e può esservi una *utilizzazione* dell'informazione. Il trasferimento deve avvenire da una sorgente ad almeno un collettore tra loro a distanza, ma senza interazione diretta tra queste parti. L'utilizzazione deve:

- rispondere alle *esigenze applicative* degli utenti posti in comunicazione;
- includere anche quanto è necessario per completare l'operazione di trasferimento, e cioè una *interazione diretta tra l'origine e la destinazione* della comunicazione per rendere possibile una fruizione dell'informazione scambiata.

Il fornitore del servizio ha il compito di *confezionare* la logica del servizio e di *rendere questa attivabile* secondo forme e modalità (ivi compresi gli aspetti di *qualità* e di *costo*) definite nei suoi impegni contrattuali con il cliente. Per rendere possibile il trasferimento dell'informazione tra l'origine e la destinazione della comunicazione, deve poter utilizzare le *risorse infrastrutturali* rese disponibili dal *gestore di rete*.

Una *rete di telecomunicazione* è la *piattaforma* su cui è possibile eseguire il programma contenente la logica di ogni servizio supportato. Consente quindi di:

- *trasferire* l'informazione a distanza secondo quanto richiesto nell'espletamento di ogni servizio;
- *controllare e gestire le sue parti componenti,* in modo che il trasferimento avvenga entro prefissati obiettivi di qualità e di costo;
- assicurare al cliente/utente e al fornitore *il controllo e la gestione dei servizi supportati.*

Il gestore di rete ha il compito di attivare e mantenere operativi i *mezzi tecnici e organizzativi* che sono atti ad assicurare il *supporto infrastrutturale* di servizi di telecomunicazione per una popolazione di utenti. I *vincoli da rispettare* sono il conseguimento di una accettabile *qualità* per ognuno dei servizi supportati e di un *costo di fornitura* commisurato al beneficio ottenibile.

In base alle potenzialità dell'ambiente di comunicazione che sono coinvolte nella loro fornitura, i servizi di telecomunicazione si distinguono in *servizi di rete* e in *servizi applicativi*. Le potenzialità che consentono di comunicare sono a loro volta localizzabili in una *rete* come piattaforma di fornitura e in *apparecchi terminali* (Terminal Equipment), questi ultimi costituenti il mezzo attraverso cui un utente usufruisce di uno o più servizi di telecomunicazione.

Ciascuna potenzialità è definita come un insieme di funzioni che possono essere suddivise in due livelli per sottolineare la dipendenza gerarchica di un livello dall'altro: cioè le *funzioni di basso livello* costituiscono la base per l'esecuzione delle *funzioni di alto livello* e queste ultime presuppongono lo svolgimento preventivo delle prime.

Le funzionalità di basso livello sono preposte al trasferimento dell'informazione attraverso la rete. Le funzionalità di alto livello riguardano invece gli aspetti connessi all'utilizzazione dell'informazione e possono anche includere funzionalità di controllo e di gestione. Nella trattazione seguente le prime

saranno chiamate *funzioni orientate al trasferimento*, mentre alle seconde sarà attribuito il termine di *funzioni orientate all'utilizzazione*.

Con queste precisazioni si può affermare che:

- i *servizi di rete* forniscono, fondamentalmente, la possibilità di trasferire informazioni tra due punti di accesso alla rete; coinvolgono quindi solo potenzialità di rete e cioè insiemi funzionali di basso livello;
- i *servizi applicativi*, così denominati in quanto rispondenti alle esigenze applicative degli utenti, forniscono una possibilità di comunicare in senso lato, e cioè comprendente oltre agli aspetti di puro trasferimento dell'informazione, anche quelli legati alla relativa utilizzazione; sono allora coinvolte, in aggiunta a potenzialità di rete, anche potenzialità di apparecchio terminale che sono caratterizzate da insiemi di funzioni di livello sia basso che alto.

La finalità di questo capitolo è fornire un riferimento per gli sviluppi della trattazione successiva. Conseguentemente verranno affrontati i seguenti argomenti:

- ♦ qual è la struttura generale di una rete per la fornitura di servizi di telecomunicazione (par. I.1);
- ♦ come è possibile caratterizzare le sorgenti di informazione che sono origine di una comunicazione (par. I.2);
- ♦ quali siano le possibilità di attuazione di una comunicazione e di caratterizzazione dei servizi (par.I.3);
- ♦ quale siano le attuali esigenze dell'utenza e le corrispondenti risposte in termini di evoluzione delle infrastrutture (par. I.4);
- ♦ quale sia la organizzazione in ambito mondiale ed europeo preposta a definire e ad aggiornare la normativa nell'ambito delle telecomunicazioni (par. I.5);

I.1 La struttura di una rete

La trattazione di questo paragrafo è dedicata alla struttura della piattaforma su cui si appoggia la fornitura dei servizi di rete e applicativi. Sono considerate:

- la distinzione tra funzioni aventi nature fisica e logica (§ I.1.1);
- le due sezioni della rete logica preposte all'accesso e al trasporto dell'informazione (§ I.1.2).

I.1.1 Reti logica e fisica

Il modello più semplice e intuitivo di una rete di telecomunicazione descrive la relativa *configurazione geometrica* (topologia). Elementi componenti di questa sono i *rami* e i *nodi*. Un *ramo*, rappresentato graficamente da un segmento di retta o di curva, costituisce elemento di connessione di due nodi. Un *nodo* è l'estremità comune di due o più rami convergenti nello stesso punto. Il significato di queste entità geometriche è diverso a seconda del tipo di operatività che si considera. Nell'operatività di una rete di telecomunicazione occorre infatti distinguere le funzioni di *natura logica* da quelle di *natura fisica*. Entrambe concorrono al trasferimento dell'informazione tra sorgente e collettore, ma con finalità ben distinte.

Nel caso delle *funzioni di natura logica*, l'attenzione è rivolta all'informazione come *insieme di stati logici* aventi significatività in un processo di comunicazione e quindi come *bene immateriale* per il quale è richiesto il trasferimento da una o più sorgenti a uno o più collettori tutti dislocati in posizioni tra loro remote.

Il trasferimento dell'informazione a distanza (o anche solo la sua memorizzazione in sede locale) richiede anche lo svolgimento di *funzioni di natura fisica*. Queste consentono di utilizzare i mezzi elettromagnetici disponibili (a propagazione libera o guidata) provvedendo al trasferimento dei *segnali* che supportano l'informazione. Si tratta quindi delle funzioni di tipo *trasmissivo.*

Il sottoinsieme di risorse funzionali (tra quelle presenti nella rete) che svolgono compiti di natura logica è chiamato *rete logica;* il sottoinsieme preposto a compiti di natura fisica è chiamato *rete fisica.*

La rete logica è quindi un'infrastruttura che consente il trasferimento di informazione da una o più sorgenti a uno o più collettori, tutti dislocati in posizioni tra loro remote. È la sede di funzioni di natura logica, aventi come fine comune la fornitura di servizi di rete.

Le due reti fisica e logica sono in stretta relazione gerarchica, dato che le funzioni di natura logica debbono utilizzare quelle di natura fisica e queste ultime sono al servizio delle prime. Entrambe le funzioni sono preposte al trasferimento e quindi al servizio di quelle di utilizzazione. Il rapporto tra rete fisica e rete logica segue il modello di interazione "*client-server*", in cui la rete logica agisce come "client" e quella fisica come "server".

Circa la *topologia della rete logica*, un ramo rappresenta il *percorso diretto* che l'informazione segue per essere trasferita da un'estremità all'altra. Un nodo descrive il *mezzo di scambio* tra due o più rami che ad esso fanno capo. Rami e nodi sono coinvolti nella formazione di *percorsi di rete.* Un ramo è in corrispondenza con gli apparati di rete che svolgono la funzione di *multiplazione* (cfr. par. IV.2). Un nodo è in corrispondenza con gli apparati di rete (commutatori) che svolgono la funzione di *commutazione* (cfr. par. IV.3).

La rete fisica è l'infrastruttura preposta al trasferimento dei segnali che supportano l'informazione. È la sede di funzioni di natura fisica, quali sono quelle di tipo trasmissivo. È l'infrastruttura di base su cui si definisce la rete logica.

Circa la *topologia della rete fisica,* i rami rappresentano le vie per il trasferimento dei segnali e sono in corrispondenza modellistica con i *sistemi trasmissivi di linea*. I nodi rappresentano i punti di trasmissione e/o ricezione dei segnali e sono in corrispondenza modellistica con gli *apparati terminali di rice-trasmissione.*

Le topologie della rete logica e della rete fisica in generale non coincidono.

I.1.2 Sezioni della rete logica

In una rete logica si distinguono una sezione *dorsale* o *interna*, chiamata *rete di trasporto*, e una sezione di ingresso/uscita, denominata *rete di accesso*.

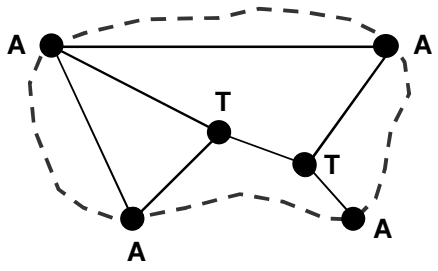


Figura I.1.1 Topologia della rete di trasporto con i relativi nodi

Tra i nodi della rete di trasporto si distinguono nodi *di accesso* e nodi *di transito* (Fig. I.1.1). La rete di trasporto ha il ruolo di trasferire l'informazione tra nodi di accesso, utilizzando, se necessario,

anche nodi di transito. È la sede di *risorse condivise* (di trasferimento e di elaborazione). È supportata da una rete fisica oggi orientata ad un uso, sempre più esteso, delle fibre ottiche.

La rete di accesso ha il ruolo di consentire l'accesso alla rete da parte dei suoi utenti. Il suo supporto fisico (mezzo trasmissivo) può essere *a filo* (wired) in rame/fibra o *senza filo* (wireless) su portante radio. È la sede di risorse che in alcuni casi sono indivise e che in altri casi sono condivise. Comprende l'interfaccia *utente-rete,* e cioè il *punto di accesso* alla rete (Fig. I.1.2).

I.2 Le sorgenti di informazione

Le sorgenti emettono informazione con il supporto di entità fisiche che sono correntemente denominate *segnali*. La caratterizzazione delle sorgenti è effettuabile con riferimento ai segnali emessi (§ I.2.1) e alle relative modalità di emissione (§ I.2.2).

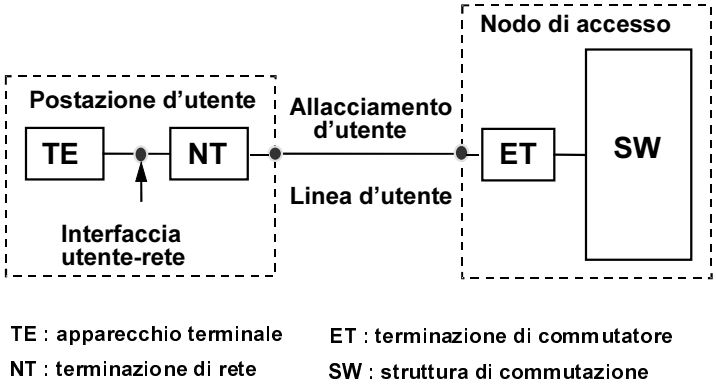


Figura I.1.1 Elemento di rete di accesso con linea di utente individuale

I.2.1 I segnali emessi

I segnali che supportano l'informazione in applicazioni telematiche sono normalmente *in forma numerica* in quanto così vengono emessi dalle loro sorgenti o così vengono ottenuti a seguito di una conversione analogico-numerica. Si distinguono *segnali numerici isocroni* e *anisocroni*.

Un *segnale numerico* è costituito da una successione di segnali elementari (*elementi di segnale*), che si susseguono nel tempo e a ciascuno dei quali è associata una porzione dell'informazione emessa dalla sorgente (ad esempio, una cifra binaria). Questa associazione è attuata facendo assumere a una *grandezza caratteristica* di ogni elemento di segnale (ad esempio, alla sua ampiezza, durata, ecc.) valori discreti che sono in corrispondenza con l'informazione da rappresentare e che sono individuati sull'asse dei tempi da un *istante significativo*.

In un segnale numerico isocrono gli intervalli di tempo tra istanti significativi consecutivi hanno, almeno in media, le stesse durate ovvero durate che sono multipli interi della durata più breve; gli scostamenti dalla media debbono essere di valore massimo contenuto entro limiti specificati. In un segnale numerico anisocrono non sono verificate queste condizioni.

Al fine di rappresentare fisicamente la successione degli istanti significativi di un segnale numerico isocrono, a questo deve essere associato, in emissione e in ricezione, un *crono-segnale*, e cioè un segnale periodico, in cui gli istanti significativi sono individuati da particolari valori dell'ascissa

temporale. Ad esempio, se il crono-segnale è sinusoidale, gli istanti significativi sono rappresentati dagli attraversamenti dello zero con pendenza positiva (o negativa), mentre, se il crono-segnale è ad onda quadra periodica, gli istanti significativi sono individuati dalla posizione temporale del fronte di salita (o di discesa).

Una sorgente numerica (ovvero il suo equivalente a valle di una conversione analogico-numerica) ha un comportamento che è descrivibile mediante l'andamento nel tempo del *ritmo binario* (bit rate) *di emissione*, e cioè del numero di cifre binarie che sono emesse nell'unità di tempo durante intervalli in cui questo numero rimanga invariato.

Quando l'emissione è sostenuta da un segnale numerico isocrono (Fig. I.2.1), in cui ogni elemento di segnale è portatore di una singola cifra binaria e il cronosegnale associato ha periodo T_s , il ritmo binario di emissione è costante e uguale a $1/T_s$.

Se invece l'emissione è con segnale numerico anisocrono, in cui ogni elemento di segnale sia ancora portatore di una singola cifra binaria (Fig. I.2.2), si possono normalmente individuare tratti di segnale in cui l'intervallo tra due istanti significativi successivi è costante e che si susseguono con altri tratti in cui detto intervallo assume un valore diverso. Si può allora definire un ritmo binario di emissione che varia da tratto a tratto e, tra i valori assunti da questi ritmi, si può individuare un valore massimo, che è chiamato *ritmo binario di picco*. Questo è quindi il numero massimo di cifre binarie che la sorgente emette nell'unità di tempo.

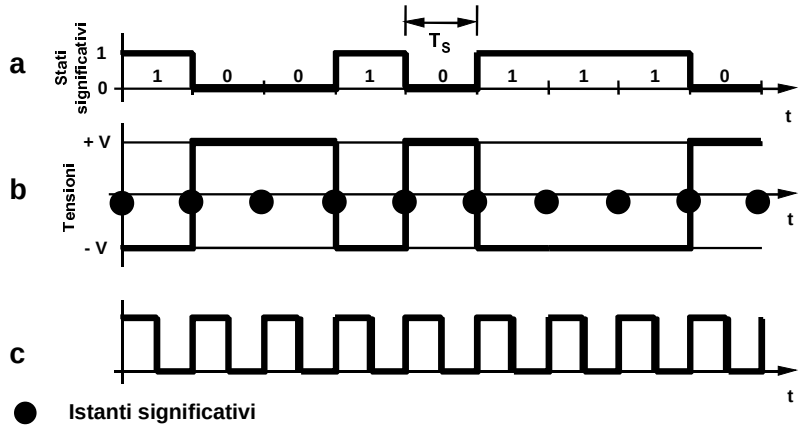


Figura I.2.1 - Esempio di segnale numerico isocrono (b) con l'informazione da rappresentare (a) e con l'associato crono-segnale (c).

I.2.2 Caratteristiche di emissione

Dal punto di vista delle loro caratteristiche di emissione, le sorgenti di informazione possono essere a *ritmo binario costante* (CBR - Constant Bit Rate) o a *ritmo binario variabile* (VBR - Variable Bit Rate). In questo secondo caso, il ritmo binario emesso varia nel tempo tra un valore massimo (il *ritmo binario di picco*) e un valore minimo, che può essere anche nullo. Tale variabilità è legata a:

- la possibile variazione nel tempo del contenuto informativo all'ingresso del codificatore;
- la convenienza di realizzare una codifica, che tenga conto di questa variazione con l'obiettivo di utilizzare in modo più efficiente la capacità di canale richiesta per trasferire a distanza l'informazione emessa.

Il caso delle sorgenti CBR può invece essere caratterizzato dal solo ritmo binario di picco.

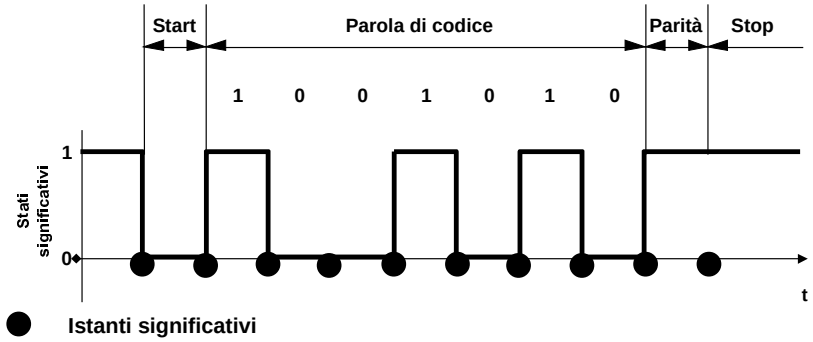


Figura I.2.2 - Esempio di segnale numerico anisocrono

In base al loro ritmo binario di picco R (Tab. I.2.1) si distinguono sorgenti *a bassa velocità* ($R \leq$ 100 kbit/s); *a media velocità* (100 kbit/s < $R \leq$ 10 Mbit/s) e ad *alta velocità* ($R >$ 10 Mbit/s).

CLASSE	SERVIZI	RITMO BINARIO DI PICCO (Mbit/s)
Bassa	Telemetria Dati, Testi Voce, Dati Immagini fisse Videotelefono Videoconferenza	0,0001 0,01 0,1 0,1 0,1 0,1
Media	 TV convenzionale Dati, Immagini fisse	1 da 1 a 10 da 1 a 10
Alta	TV ad alta definizione Dati, Immagini fisse	da 10 a 100 da 10 a 1000

Tabella I.2.1 - Classi di velocità

Un esempio di ritmo binario variabile è fornito da una sorgente "*tutto o niente*" (on-off), nella quale la sequenza dei dati emessi è strutturabile in intervalli di *attività* e di *silenzio*. Gli intervalli di attività sono chiamati *tratti informativi,* mentre quelli di silenzio sono detti *pause* (Fig. I.2.3). Nelle sorgenti "tutto o niente" le durate dei tratti informativi e delle pause sono, in generale, quantità variabili in modo aleatorio e quindi descrivibili solo in modo probabilistico. Le distribuzioni di queste quantità sono strettamente legate al tipo di sorgente considerata e quindi al servizio nell'ambito del quale la sorgente opera.

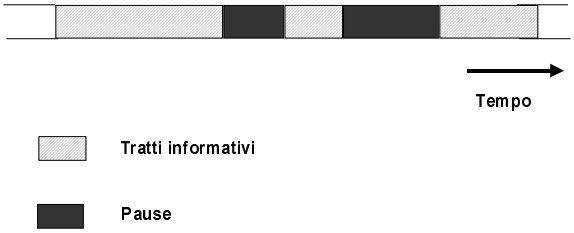


Figura I.2.3 - Emissione di una sorgente VBR "tutto o niente"

Nel caso di sorgenti VBR più generali di quelle "tutto o niente", l'emissione si presenta come una successione di tratti informativi in ognuno dei quali il ritmo binario emesso è costante e si modifica di

valore passando da un tratto ad un altro. Circa le durate di questi tratti con i relativi ritmi binari emessi, valgono considerazioni analoghe a quelle fatte con riferimento alle sorgenti "tutto o niente": si è in presenza cioè di quantità variabili in modo aleatorio.

Conseguentemente, una descrizione completa di una sorgente VBR richiederebbe modelli probabilistici che possono essere di identificazione e di utilizzazione anche molto complesse. Se pero' la descrizione può essere ristretta al solo *comportamento in media*, si può fare riferimento a un *ritmo binario medio*, ottenuto come media dei valori dei ritmi binari che la sorgente emette nel tempo. Inoltre per ogni sorgente la variabilità del ritmo binario può essere sommariamente descritta dal *grado di intermittenza* (burstiness), e cioè dal *rapporto tra il ritmo binario di picco e quello medio*.

I.3 Servizi e infrastrutture

Per fornire un idea più precisa delle svariate possibilità di attuazione di una comunicazione,

- sono definiti i tipi di flussi di informazione che sono scambiati a sostegno dell'erogazione di un servizio (§ I.3.1);
- è considerata la varietà di infrastrutture che hanno trovato attuazione per soddisfare specifiche esigenze operative o per assecondare le opportunità offerte da evoluzioni tecnologiche (§ I.3.2);
- è fornita una caratterizzazione dei servizi di rete e applicativi (§ I.3.3);
- è presentata una classificazione dei servizi basata sul grado di cooperazione che si stabilisce tra le parti in comunicazione (§ I.3.4).

I.3.1 Le informazioni scambiate

Nell'erogazione di un servizio vengono in generale trasferiti tre tipi di informazione: quelle di *utente*, di *controllo* e di *gestione*.

L'*informazione di utente* include quanto viene emesso da una sorgente ed è destinato a uno o più collettori di informazione per le finalità di una particolare applicazione, ma comprende anche l*'extra-informazione* (overhead) che viene in generale aggiunta al flusso informativo di sorgente per scopi di procedura o di protezione (*informazione di coordinamento*). Lo scambio dell'informazione di utente è quindi l'obiettivo primario di un servizio di telecomunicazione.

Costituiscono l'informazione di utente, in alternativa o in unione parziale o totale, le forme codificate di *voce*, di *suoni musicali*, di *dati*, di *testi*, di *immagini fisse* o *in movimento*. A parità di natura dell'informazione scambiata, è la sua *forma codificata* che pone requisiti al relativo trattamento infrastrutturale. Al riguardo è oggi d'uso parlare di *mezzo di rappresentazione* con riferimento a uno specifico tipo di informazione, come descritto dalla sua forma a valle di un operazione di codifica con o senza riduzione di ridondanza. In relazione poi alla capacita' di gestire un solo mezzo di rappresentazione o una pluralità di questi, una comunicazione si dice *monomediale* nel primo caso o *multimediale* nel secondo; gli stessi attributi sono utilizzati con riferimento a un servizio o a una applicazione.

L'informazione di utente può essere scambiata tra due o più utenti o tra utenti e centri di servizio e, nello scambio, può essere trattata dall'infrastruttura in *modo trasparente* oppure *può essere elaborata*, come accade nei casi di una archiviazione, di una conversione di mezzo di rappresentazione (ad esempio, da testo a voce sintetizzata), di una codifica crittografica svolta all'interno della rete.

L'*informazione di controllo* (o di segnalazione) è di supporto affinché possa avvenire lo scambio dell'informazione di utente. Essa ha lo scopo di consentire le interazioni tra cliente/utente e fornitore di servizi nell'ambito di quanto previsto: (i) per *inizializzare* la comunicazione, per *negoziarne le caratteristiche* qualitative e quantitative iniziali e per *modificare* tali caratteristiche nel corso della

comunicazione; (ii) per *ottenere un arricchimento* dei servizi di base con il coinvolgimento di risorse di elaborazione accessibili nell'ambiente di comunicazione.

L'*informazione di gestione* ha lo scopo di consentire il complesso di operazioni necessarie per gestire la fornitura dei servizi e i mezzi necessari allo scopo; in questi ultimi sono ovviamente incluse le risorse preposte al trasferimento delle informazioni di utente e di controllo. Tra le operazioni di gestione vanno menzionate quelle connesse all'erogazione del servizio (*esercizio*), al suo mantenimento (*manutenzione*) e al suo addebito (*amministrazione*). Per lo svolgimento di tali operazioni deve essere previsto uno scambio di informazioni tra le *apparecchiature di rete* e quelle *terminali*. Oggetto dello scambio è l'informazione di gestione, che è quindi anch'essa di supporto al trattamento infrastrutturale dell'informazione di utente.

Il trasferimento delle informazioni di utente, di controllo e di gestione può essere attuato nell'ambito di un'*unica infrastruttura*. Questa è la soluzione adottata in passato per le reti dedicate a un servizio. Gli attuali orientamenti sono a favore dell'impiego di infrastrutture separate. Per l'informazione di utente si attua allora una infrastruttura, separata da quella di segnalazione (*rete di segnalazione a canale comune*) e da quella di gestione (*rete di gestione delle telecomunicazioni*).

I.3.2 Tipi di infrastrutture

Esistono varie modalità per classificare i numerosi tipi di infrastrutture a supporto della fornitura di servizi. In base al *tipo di utenza* si distinguono: *reti pubbliche* e *reti private*. Una rete è *pubblica* se l'accesso è consentito a chiunque provveda a stabilire un accordo contrattuale con il fornitore di servizi. È invece *privata* quando gli utenti abilitati all'accesso costituiscono un insieme chiuso con specifiche esigenze di comunicazione, che richiedono accordi tra cliente e fornitore non assimilabili a quelli in ambito pubblico.

In base alla *mobilità del terminale*, si distinguono *reti fisse* e *mobili*. Una rete è *fissa* se i servizi supportati dalla rete sono accessibili solo da parte di utenti che, ogniqualvolta desiderino comunicare, siano in posizione statica o che, pur in movimento, rimangano in un intorno relativamente ristretto di un sito di riferimento (abitazione, ambiente di lavoro, ecc.). È invece *mobile* se l'accesso è consentito ad utenti che sono in movimento senza limitazioni alle loro possibilità di deambulazione (a piedi o su veicoli).

In base alla loro *estensione* si distinguono *reti in area geografica* e *in area locale*. Una rete si dice *in area geografica* (Wide Area Network, WAN) quando gli utenti sono distribuiti su un'area molto estesa (una nazione, un continente, l'intero globo terrestre). È detta invece in *area locale* (Local Area Network, LAN) quando l'area interessata è ristretta ad un singolo edificio o a un complesso di insediamenti entro il raggio di qualche chilometro.

Infine, in base alla *gamma dei servizi supportati*, si distinguono *reti dedicate e reti integrate.* Le *reti dedicate a un servizio* sono state concepite e realizzate in passato per la fornitura di un singolo servizio e possono oggi essere utilizzate anche per un insieme ristretto di altri servizi, seppure con limitazioni severe per ciò che concerne la qualità conseguibile; esempi significativi di reti di questo tipo sono la *rete telefonica* e le *reti per dati*. Le *reti integrate nei servizi* sono invece la realtà attuale; il loro obiettivo è rendere possibile la fornitura di una *vasta gamma di servizi* con prestazioni di qualità e di costo decisamente migliori rispetto a quelle ottenibili con le reti dedicate.

I.3.3 Caratterizzazione dei servizi

I servizi applicativi e di rete sono *di base* o *supplementari.* I servizi *di base* possono essere singolarmente offerti in modo autonomo da altri servizi dello stesso tipo. I servizi *supplementari*

modificano o complementano uno o più servizi di base; ne segue che un servizio supplementare non è offribile in modo autonomo da uno di base.

La *caratterizzazione* dei servizi di base è fondata su un elenco di *attributi*, che riguardano il *trasferimento dell'informazione di utente*, le *modalità di accesso* e altre proprietà generali. Tra i sette attributi relativi al trasferimento dell'informazione, se ne distinguono quattro, che sono detti *dominanti* e che sono utilizzati per identificare una particolare categoria di servizi. I rimanenti tre sono gli *attributi secondari* e sono impiegati per identificare un particolare servizio entro una categoria. Infine, per caratterizzare le *modalità di accesso* e altre proprietà generali, sono utilizzati gli *attributi qualificanti*, che meglio specificano un singolo servizio.

Ognuno degli attributi può assumere un valore scelto in una gamma di alternative possibili. A questo riguardo conviene premettere alcune definizioni riguardanti l'informazione d'utente, che viene scambiata durante la fruizione di un servizio, e la sua relazione con il *punto di accesso alla rete* (cfr. § I.1.2).

Nell'ambito di una comunicazione, la rete che ne consente lo svolgimento è attraversata in primo luogo da *flussi informativi di utente*, che interessano almeno due punti di accesso: l'uno adiacente alla sorgente e l'altro prossimo al collettore. Per ognuno di questi flussi, l'interessamento di un punto di accesso è sommariamente caratterizzabile con:

- il *verso di scorrimento*, che è quello dell'informazione trasportata dal flusso: tale verso può essere *uscente* o *entrante* a seconda che il trasferimento sia dall'apparecchio terminale al punto di accesso o viceversa;
- il *ritmo di flusso*, che ha significatività quando il flusso è composto da cifre binarie con cadenza regolare e che è misurato dal numero di cifre binarie componenti il flusso che attraversano il punto di accesso nell'unita' di tempo;
- la *portata media di flusso*, che è un parametro significativo ogniqualvolta il flusso considerato è composto da cifre binarie che non hanno cadenza regolare e che, in queste condizioni, misura (ad esempio in bit/s) la quantità di informazione che viene mediamente trasportata dal flusso nell'unita' di tempo attraverso il punto di accesso; come è evidente ritmo e portata media di flusso coincidono quando il flusso è composto da cifre binarie con cadenza regolare.

Tra gli attributi dominanti è significativo citare il *modo di trasferimento*: questo descrive le modalità operative che sono seguite nella fornitura di un servizio per trasferire l'informazione d'utente da un punto di accesso ad un altro; se ne parlerà in dettaglio nel Cap. IV°. Un altro attributo dominante è il *ritmo di trasferimento*; questo riguarda il flusso informativo che il servizio attiva attraverso il punto di accesso con verso uscente e caratterizzabile con un parametro dipendente dal tipo di modo di trasferimento utilizzato.

Gli attribuiti secondari sono la simmetria, la configurazione e l'inizializzazione.

La *simmetria di una comunicazione* riguarda la relazione tra verso di scorrimento e portata media del relativo flusso attivati dal servizio tra due o più punti di accesso. In particolare, la comunicazione è:

- *unidirezionale*, quando si ha scorrimento solo in un verso;
- *bidirezionale simmetrica*, quando sono attivati entrambi i versi di scorrimento e le portate medie dei relativi flussi hanno valori paragonabili;
- *bidirezionale asimmetrica*, quando, pur in presenza di entrambi i versi di scorrimento, la portata media di un flusso è di valore decisamente prevalente rispetto a quello della portata media dell'altro flusso.

La *configurazione di una comunicazione* fa riferimento alla dislocazione spaziale e al numero dei punti di accesso che sono coinvolti nella fornitura di un servizio. Si parla allora di comunicazione:

- *punto-punto*, quando vengono interessati solo due punti di accesso;

- *multipunto*, quando i punti di accesso coinvolti sono in numero maggiore di due;
- *diffusiva*, quando, come nel caso precedente, i punti di accesso sono in numero maggiore di due, ma con la differenza che, in questo caso, l'informazione fluisce da un unico punto verso gli altri in modo unidirezionale.

E' da aggiungere che il caso multipunto include le configurazioni *punto-multipunto* (una singola origine con una molteplicità di destinazioni), *multipunto-punto* (più origini con una singola destinazione) e *multipunto-multipunto* (più origini e più destinazioni).

Le modalità da seguire per dare inizio e conclusione a un trasferimento di informazione nell'ambito dell'utilizzazione di un particolare servizio di telecomunicazione (*inizializzazione di una comunicazione*) possono essere su basi *chiamata*, *prenotazione* e *permanente*.

In una *comunicazione su base chiamata* (nel seguito denominata, per brevità, *chiamata*) si distinguono tre fasi: una fase iniziale di *richiesta* del servizio, una fase intermedia di *utilizzazione* e una fase finale di *chiusura*. In una chiamata è sempre possibile individuare, come attori, almeno due utenti: da un lato *l'utente chiamante*, che presenta la richiesta di fornitura di un servizio e, dall'altro, *l'utente chiamato*, con cui il primo desidera stabilire uno scambio di informazione. Il termine "*chiamata*" viene normalmente utilizzato con riferimento a una interazione utente-rete in cui la rete risponde alla richiesta dell'utente *non appena possibile*, e cioè con un ritardo contenuto entro quanto consentito dalla tecnologia realizzativa e dalle condizioni di carico della rete. In queste condizioni la comunicazione può essere iniziata, non appena possibile, dopo la presentazione della richiesta da parte dell'utente e termina, non appena possibile, su richiesta di una delle parti in comunicazione.

Esiste però un'altra possibilità, che, seppure su una diversa scala temporale, ha una organizzazione in fasi del tipo di quella ora definita. È questo il caso di una *comunicazione su base prenotazione*, in cui la comunicazione può iniziare a un istante che è stato definito in precedenza al momento di una prenotazione dell'utente e termina dopo un tempo che è stato prefissato al momento della prenotazione ovvero richiesto durante la svolgimento della comunicazione.

Nei due casi di comunicazioni su basi chiamata e prenotazione, è anche necessario uno scambio di *informazione di segnalazione* (cfr. § I.3.1) tra le apparecchiature coinvolte nell'espletamento del servizio (apparecchi terminali, terminazioni di rete, apparati della rete di trasporto). In particolare, nel caso di comunicazioni su base chiamata, l'informazione di segnalazione è elemento essenziale per consentire agli apparati di rete di svolgere la funzione di *trattamento di chiamata*. Questa ha lo scopo di mettere a disposizione degli utenti, quando ne fanno richiesta all'inizio della chiamata, quanto loro occorre per comunicare con altri utenti; supervisionare lo svolgimento della comunicazione e prendere atto della conclusione della chiamata.

Infine, in una *comunicazione su base permanente* non esiste una organizzazione in fasi. Esiste invece un contratto tra cliente e fornitore per la erogazione di un servizio senza vincoli sulla sua durata di fruizione. Nell'ambito di tale rapporto la comunicazione può iniziare a un istante qualunque successivo alla stipula del contratto e può continuare, a discrezione dell'utente, fino al termine stabilito contrattualmente.

I.3.4 Cooperazione tra le parti

Dal punto di vista del *grado di cooperazione* che si stabilisce tra le parti in comunicazione, si distinguono servizi *interattivi* e servizi *distributivi*. I servizi interattivi a loro volta possono essere: *di conversazione*, *di messaggistica* e *di consultazione*. Tra i servizi distributivi si distinguono quelli *con controllo di presentazione* e *senza controllo di presentazione*.

Nei *servizi interattivi* la comunicazione coinvolge due o più utenti (operatori umani, macchine di elaborazione, banche di informazione), che interagiscono tra loro per il conseguimento di uno scopo definito, quale un dialogo in tempo reale, un trasferimento di messaggi in tempo differito o una

consultazione di informazione archiviata in appositi centri di servizio. Nei *servizi distributivi* esiste una sorgente centralizzata che distribuisce l'informazione a un gran numero di utenti senza richieste individuali.

I *servizi di conversazione* sono di tipo interattivo e forniscono il mezzo per un dialogo a distanza tra due o più utenti; richiedono, tra queste parti, un trasferimento di informazione *in tempo reale*, e cioè con un ritardo di transito che non pregiudichi la possibilità e l'efficacia del dialogo.

Anche i *servizi di messaggistica* sono di tipo interattivo e offrono una comunicazione da utente a utente per mezzo di uno scambio, in tempo differito, di messaggi aventi come contenuto, in alternativa o in unione, testi, voce o immagini. La comunicazione avviene per il tramite di dispositivi di memorizzazione, che possono svolgere funzioni di immagazzinamento e rilancio, di casella postale e di trattamento di messaggio.

Nei *servizi di consultazione*, sempre di tipo interattivo, l'utente ha la possibilità di reperire l'informazione memorizzata in appositi centri di servizio; l'informazione è inviata all'utente solo a sua domanda e può essere consultata su base individuale; l'istante in cui la sequenza delle informazioni richieste deve iniziare è sotto controllo dell'utente.

I *servizi con controllo di presentazione* sono servizi distributivi, in cui l'informazione, diffusa circolarmente, è strutturata in una sequenza di unità con ripetizione ciclica; l'utente ha la possibilità di accedere individualmente a tale informazione e può controllarne sia l'inizio che l'ordine di presentazione.

Infine i *servizi senza controllo di presentazione* sono servizi distributivi, in cui la sorgente centralizzata emette, senza soluzione di continuità, un flusso informativo che l'utente riceve senza però avere la possibilità di controllarlo; l'utente non può determinare l'inizio e l'ordine di presentazione dell'informazione diffusa.

I.4 Le esigenze di comunicazione

Da circa un quarto di secolo l'utenza dei servizi di telecomunicazione, con riferimento specifico a quella degli ambienti di lavoro (*utenza-affari*) e seppure in modo graduale, sta manifestando il bisogno di allargare le proprie possibilità di scambi informativi. Si è così passati da un quadro di esigenze che potevano essere soddisfatte dalla disponibilità di servizi singoli nell'ambito di *comunicazioni monomediali* su infrastrutture fisse ad un altro in cui la richiesta sta riguardando un'ampia gamma di *comunicazioni multimediali*, sotto forma di una pluralità di servizi da fruire in *modo integrato* e con possibilità di *accesso fisso* o *mobile* a seconda delle esigenze del momento.

In una prima fase l'esigenza di multimedialità ha riguardato servizi che, orientativamente, appartenevano alle classi a bassa e a media velocità, cioè con esigenze di ritmo binario di trasferimento non superiori a 2 Mbit/s. E' questo un ambiente di comunicazione "*a banda stretta*".

In una fase successiva, si sono manifestati ulteriori esigenze di multimedialità, che comprendono:

- comunicazioni su base chiamata con possibilita' di trasferire anche informazioni di dati e di immagine con movimento completo, nell'ambito di servizi ad alta velocita';
- comunicazioni senza connessione per trasferire dati tra reti in area locale e stazioni di lavoro ad alta velocità elaborativa;
- comunicazioni multimediali con trasferimento, in forma correlata, di flussi informativi composti di voce, di dati e di immagini fisse o in movimento;
- comunicazioni con configurazione multipunto per servizi sia interattivi che distributivi.

L'ambiente di comunicazione in cui possono essere soddisfatte queste modalità di scambio informativo è "*a banda larga*".

Al bisogno di multimedialità a banda stretta o larga, negli ultimi dieci anni si è aggiunta l'ulteriore esigenza della *ubiquità* e della *personalizzazione* della comunicazione. Agli utenti deve cioè essere consentito di comunicare

- in *qualunque luogo* di ambienti pubblici o privati e di un'area locale o geografica;
- in *qualunque momento* delle loro attività lavorative e/o ricreative e con le stesse caratteristiche di *reperibilità* e di *sicurezza* a loro offerte nei loro ambienti domestico o lavorativo;
- con *potenzialità di banda* tali, almeno in prospettiva, da non creare limitazione alla forma di comunicazione;
- secondo modalità che sono ritagliate sulle loro *esigenze specifiche* e che possono essere direttamente sotto il loro controllo.

Per completare il quadro delle attuali esigenze di comunicazione, occorre considerare anche gli obiettivi in relazione:

- alla *sicurezza* nello scambio e nella manipolazione dell'informazione;
- a una migliore qualità dei servizi e a un contenimento dei costi di fornitura, come è possibile conseguire con attività di *gestione della rete e dei servizi*.

Alla *multimedialità* (§ I.4.1), alla *ubiquità* e alla *personalizzazione dei servizi* (§ I.4.2), alla *sicurezza/protezione delle informazioni* (§ I.4.3) e alla *gestione della rete e dei servizi* (§ I.4.4) sono dedicati i contenuti del presente paragrafo, che si conclude con una sintetica presentazione dell'attuale stato di evoluzione delle infrastrutture (§ I.4.5) come risposta alle esigenze di comunicazioni multimediali, mobili e personali.

I.4.1 La multimedialità

Lo sviluppo delle telecomunicazioni è stato caratterizzato per oltre mezzo secolo dalla richiesta e dall'offerta di servizi *monomediali*. Ognuno di questi consente il trattamento di un singolo mezzo di rappresentazione, che lo caratterizza rispetto a altri servizi e che è ovviamente diversificabile passando da un servizio all'altro.

Oggi, in aggiunta a servizi monomediali, si manifesta l'esigenza di nuove forme di comunicazione, in cui siano trattati due o più mezzi di rappresentazione e in cui questi ultimi conservino, nell'operazione di trasporto, le caratteristiche di concatenazione fisica e logica presentate all'origine. Si parla di servizi *multimediali*.

I.4.2 L'ubiquità e la personalizzazione dei servizi

L'*ubiquità dei servizi* è la possibilità per l'utente di essere chiamante o chiamato in qualsiasi momento della giornata o in qualsiasi posto si trovi nell'ambito dei suoi spostamenti a piedi o su mezzi di trasporto e di fruire di prestazioni di qualità di servizio indipendenti da queste diverse condizioni di comunicazione. L'ubiquità del servizio comporta il soddisfacimento del requisito di *mobilità*, nel quale possono distinguersi due aspetti principali, connessi alla persona che comunica e all'apparecchio terminale che viene utilizzato.

La *mobilita' della persona* consente a un utente di usufruire di tutti i servizi che sono di suo interesse, indipendentemente dalla terminazione di rete a cui è fisicamente connesso. La *mobilita' del terminale* consente all'utente di comunicare in condizioni di effettivo movimento in un'area geografica che dovrebbe essere il più possibile vasta e, in particolare, in un'area più o meno estesa nell'intorno di singole terminazioni di rete.

La *personalizzazione dei servizi* è la possibilità per l'utente di poter comunicare secondo modalità che sono ritagliate sulle sue esigenze specifiche e che possono essere direttamente sotto il suo

controllo. A tale scopo è necessario soddisfare la *larga varietà di interessi di ogni persona*, che da un lato manifesta necessita' di razionalizzare le sue funzioni di lavoro e che dall'altro cerca di coltivare i suoi interessi culturali e ricreativi. Per attuare la personalizzazione di una comunicazione i tre attori principali che intervengono nel conseguimento di questo obiettivo, e cioè il *cliente*, il *gestore di rete* e il *fornitore dei servizi*, assumono ruoli specifici.

In particolare, al *cliente* deve essere assicurata una maggiore quota di controllo sulle modalità di creazione e di gestione dei servizi. Esempi al riguardo, e in ordine di crescente complessità di controllo, possono essere la possibilità per l'utente di:

- modificare i parametri che descrivono il suo profilo entro i limiti stabiliti dal contratto con il fornitore dei servizi;
- variare, in funzione delle sue esigenze di qualità di servizio nella fruizione di informazioni audio-visive, i parametri degli algoritmi di codifica;
- costruire uno specifico servizio mediante la composizione dinamica di "elementi" di servizio, assemblati anche in tempo reale durante la svolgimento di una comunicazione.

Circa poi il punto di vista del *gestore di rete*, un primo obiettivo da raggiungere è poter manipolare, in modo flessibile e sotto il suo controllo, le componenti funzionali presenti nell'infrastruttura di rete. Inoltre un altro obiettivo oggi giudicato di particolare interesse è riuscire a integrare i servizi all'utenza con le applicazioni gestionali.

Riguardo infine al *fornitore di servizi*, è da sottolineare la sua convenienza a operare in un contesto di *modularità* e di *inter-operabilità* degli elementi di servizio atti a comporre servizi personalizzati. Le applicazioni di interesse per l'utenza, anche in una prospettiva a lungo termine, potranno diventare entità da comporre e da integrare tra loro per dare origine ad altri servizi più complessi, a loro volta ulteriormente componibili e integrabili con altri.

I.4.3 La sicurezza e la protezione delle informazioni

Le informazioni scambiate o manipolate in una comunicazione debbono essere adeguatamente protette in modo da conseguire sicurezza relativamente alla *autenticazione*, alla *integrità*, alla *confidenzialita'* e alla *imputabilità*.

L'*autenticazione* dei dati o della loro origine è necessaria per garantire che le parti impegnate in una comunicazione siano effettivamente quelle che dichiarano di essere.

L'integrità dei dati rappresenta, invece, la garanzia contro accidentali o indebite modifiche del flusso informativo di utente durante il suo trasferimento.

La *confidenzialità* è poi la garanzia che il contenuto informativo di una comunicazione o la conoscenza delle parti in questa coinvolte non siano divulgate a terze parti, se non debitamente autorizzate. Questa garanzia rappresenta quindi il fine tradizionale delle tecniche di *cifratura dell'informazione* quali sono trattate dalla *crittografia*.

La *imputabilità* rappresenta infine la garanzia, per l'utente, di vedersi attribuito un addebito effettivamente commisurato ai servizi fruiti e, per il fornitore, di poter effettuare una tariffazione che sia incontestabile.

I.4.4 La gestione della rete e dei servizi

La *gestione* di un ambiente di comunicazione comprende l'insieme delle attività volte al perseguimento della migliore qualità dei servizi offerti e al contenimento dei costi di fornitura. Questo secondo obiettivo può essere conseguito con una più efficiente utilizzazione delle risorse impiantistiche e operative. Le principali *funzioni di gestione* riguardano:

- l'esercizio e la manutenzione degli elementi di rete;

- la gestione dei servizi, compreso il controllo, ove richiesto, da parte dell'utente; questi agirà sui parametri di gestione dei servizi a lui forniti;
- la supervisione, la misura e la gestione del traffico di telecomunicazione;
- la gestione della tassazione, e cioè degli addebiti al cliente per i servizi forniti;
- la pianificazione e la progettazione di reti e di servizi.

Per lo svolgimento di queste funzioni, la tendenza ormai consolidata è quella di sovrapporre alla infrastruttura gestita una *rete di gestione* (TMN-Telecommunication Management Network). Questa comprende *entità funzionali* situate negli elementi di rete e *sistemi di mediazione*, che cooperano mediante l'impiego di appropriate funzioni di trasporto dell'informazione di gestione. A questa rete sovrapposta si agganciano i *sistemi di gestione* (OS-Operation System), che hanno quindi tutte le informazioni provenienti dalla infrastruttura gestita o contenute nella rete di gestione: possono pertanto inviare opportuni comandi a tutti gli apparati che sono sotto le loro cure gestionali.

I.4.5 Evoluzione delle infrastrutture

La risposta al mutato quadro di esigenze connesse alla transizione dalla monomedialità alla multimedialità dei servizi si è manifestata in vari passi. In primo luogo si è attuata una *integrazione delle tecniche di rete*: cioè da mezzi di comunicazione realizzati con tecnologia completamente analogica si è passati, seppure gradualmente, ad un impiego generalizzato di tecnologie numeriche, basate su componentistica di tipo elettronico e, ormai, anche di tipo ottico. La convenienza di questa trasformazione di natura tecnica risiedeva (e risiede tuttora) in una migliore qualità dei servizi offerti e in ridotti costi di fornitura.

Parallelamente è iniziata una *integrazione dei servizi*, e cioè una ulteriore trasformazione finalizzata a rendere fruibile un accesso comune in forma numerica (*integrazione degli accessi*) per tutti i servizi di base relativi allo scambio di voce, di dati e di immagini fisse o in movimento: l'obiettivo era lo svolgimento di comunicazioni multimediali.

La realizzazione di ambienti di comunicazione multimediale "a banda stretta" è iniziata assumendo come punto di partenza la rete telefonica integrata nelle tecniche e ha condotto alla cosiddetta *N-ISDN* (Narrowband Integrated Services Digital Network). Relativamente alle integrazione degli accessi, i risultati oggi sono nelle possibilità di scelta delle utenze domestica e affari.

La ulteriore evoluzione verso il soddisfacimento di esigenze di comunicazione a banda larga richiede interventi su entrambe le sezioni di rete (accesso e trasporto) ed è tuttora in atto. Le linee-guida dello sviluppo hanno ipotizzato in una prima fase (prima metà degli anni '90) la realizzazione di una *B-ISDN* (Broadband Integrated Services Digital Network) in qualche modo legata allo sviluppo del paradigma telefonico. Solo successivamente (seconda metà degli anni '90) dette linee-guida si sono decisamente orientate sullo sviluppo del paradigma Internet.

Infatti attualmente viene riconosciuto in modo unanime il ruolo centrale di Internet nell'attuale e futuro sviluppo delle infrastrutture per telecomunicazioni: è avvenuta cioè, da parte di Internet, la sostituzione dei tradizionali paradigmi (in primo luogo di quello telefonico), che hanno finora guidato l'evoluzione delle comunicazioni.

Sulla base di queste premesse l'attenzione allo sviluppo delle infrastrutture nell'ultima decade è stata rivolta:

- in primo luogo a *Internet* per le ragioni sopra chiarite;
- all'integrazione degli accessi, con le potenzialità per il sostegno di comunicazioni voce/dati (servizi *Frame Relaying*) e con apertura anche alle comunicazioni multimediali a larga banda (*rete ATM*);
- alle *reti mobili,* come supporto all'erogazione di servizi personalizzati con mobilità del terminale;

- infine alle *reti intelligenti*, come risposta alle esigenze di personalizzazione e di mobilità della persona.

Internet, integrazione degli accessi, reti ATM e reti mobili saranno oggetto di trattazione specifica nel seguito dell'opera. In questa sezione ci limitiamo a fornire qualche breve nota sulle reti intelligenti, dato che queste non verranno ulteriormente considerate.

Alcuni servizi (come ad esempio quello supplementare di numerazione abbreviata) sono forniti impegnando le risorse del solo nodo che è di accesso per l'utente richiedente. Altri servizi (più evoluti dei precedenti e esemplificativi tramite il servizio supplementare di identificazione della linea chiamante/chiamata ovvero quello di gruppo chiuso di utenti) coinvolgono invece due o più nodi, ma sempre senza distinzione tra la logica del servizio e le modalità di fornitura.

Per lo sviluppo di servizi supplementari ancora più evoluti (come ad esempio quelli rispondenti alle esigenze di personalizzazione e di *mobilità* della persona) si è riconosciuta l'importanza di *separare la logica del servizio dalle modalità di fornitura.*

Su questo principio è basata una infrastruttura, chiamata *rete intelligente* (IN-Intelligent Network), che costituisce la piena valorizzazione delle funzionalità di controllo. I servizi supplementari fornibili tramite una IN offrono quanto è assicurato dalle *funzionalità di controllo* con il coinvolgimento di *risorse di elaborazione condivise*; sono chiamati *servizi di rete intelligente*.

La IN comprende due tipi di nodi e alcuni sistemi periferici specializzati. I nodi sono quelli *"intelligenti"* e quelli *di accesso*. Mentre i primi sono apparecchiature centralizzate e in poche unita' nell'ambito di una IN, i secondi sono disseminati nella rete in relazione alla consistenza numerica dell'utenza. I *nodi "intelligenti"* o SCP (Service Control Point) costituiscono la *parte controllante* della IN e, come tali:

- contengono la logica dei servizi offerti sotto forma di appositi programmi (SLP-Service Logic Program) e i dati relativi al profilo degli utenti che possono accedere alla IN;
- eseguono l'SLP rispondendo a una richiesta esplicita da parte dei nodi di accesso;
- istruiscono questi ultimi circa il trattamento di chiamata.

I *nodi di accesso* o SSP (Service Switching Point) sono invece la *parte controllata* della IN e costituiscono il tramite con gli utenti per la fornitura dei servizi; quindi:

- riconoscono la richiesta di un servizio di rete intelligente;
- interrogano conseguentemente i nodi "intelligenti";
- svolgono il trattamento di chiamata sulla base delle istruzioni ricevute da questi ultimi.

I.5 La normativa nelle telecomunicazioni

La normativa nelle telecomunicazioni è curata da organismi internazionali in ambito *mondiale* ed *europeo*. Tra gli organismi mondiali si citano:

- *ITU* (International Telecommunication Union);
- *ISO* (International Organization for Standardization);
- *IEC* (International Electrotechnical Commission).

L'*ITU* è un'agenzia specializzata delle Nazioni Unite, con sede in Ginevra e con il compito di armonizzare tutte le iniziative mondiali e regionali nel settore delle Telecomunicazioni. Tiene tre tipi di conferenze amministrative tra cui si menziona *WARC* (World Administrative Radio Conference); questa considera e approva tutti i cambiamenti alle regolamentazioni sulle radio comunicazioni con particolare riferimento all'uso dello spettro delle frequenze radio. Dagli inizi degli anni '90 l'ITU è organizzato in tre settori: *sviluppo, standardizzazione, radiocomunicazioni*.

L'*ITU-TS* è il settore dell'ITU preposto ad attività di standardizzazione. Include le precedenti attività svolte dal *CCITT* e parte dell'attività di normativa precedentemente svolta dal *CCIR*. L'ITU-TS è il principale organismo internazionale per la produzione di standard tecnici nel campo delle

Telecomunicazioni. La sua attività è organizzata in *Gruppi di Studio* (Study Group-SG), che sono costituiti per trattare le cosiddette "*Questioni*". Produce *Raccomandazioni*: queste hanno carattere volontario, ma costituiscono di fatto un linea-guida fondamentale per le attività dei diversi attori nel mondo delle Telecomunicazioni.

L'*ITU-RS* è il settore preposto ad attività sulle Radiocomunicazioni. Include il resto delle precedenti attività del CCIR e quella del IFBR (International Frequency Registration Board).

L'*ISO* è un Ente delle Nazioni Unite, creato con l'obiettivo di promuovere lo sviluppo della normativa internazionale per facilitare il commercio di beni e servizi nel mondo. Relativamente alle Telecomunicazioni e alle aree collegate, opera tramite un comitato tecnico congiunto con l'IEC: Joint Technical Committee on Information Technology (*JTC1).* L'ISO-IEC JTC1 è organizzato in *Comitati Tecnici* (TC) e in *Sotto-comitati* (SC).

Si fa anche riferimento allo *sviluppo di Internet*. Questo è sotto il controllo di quattro gruppi:

- *ISOC* (Internet Society) è una società professionale per facilitare, supportare e promuovere l'evoluzione e la crescita di Internet;
- *IAB* (Internet Architecture Board) è l'organismo di supervisione e di coordinamento tecnico; è composto di una quindicina di volontari internazionali e opera nell'ambito di ISOC;
- *IETF* (Internet Engineering Task Force) è il gruppo preposto alla definizione degli standard nel mondo Internet con obiettivi di breve termine; opera nell'ambito dell'IAB;
- *IRTF* (Internet Research Task Force) sviluppa progetti di ricerca di lungo termine; opera nell'ambito dell'IAB.

Tra gli organismi europei si citano:

- *ETSI* (European Telecommunication Standards Institute);
- *CEN* (European Committee for Standardization);
- *CENELEC* (European Committe for Electrotechnical Standardization).

In *ETSI* la preparazione degli standard è effettuata da *Comitati Tecnici* (TC), che trattano argomenti specifici e che riferiscono all'*Assemblea Tecnica* (TA). *CEN* è l'equivalente dell'ISO in ambito europeo. Infine *CENELEC* prepara standard elettrotecnici di interesse europeo.

In ambito mondiale esistono anche *gruppi di interesse*, tra i quali si menzionano: l'*ATM Forum*, il *Frame Relay Forum* e l'*NM* (Network Management) *Forum*. Si cita anche l'*ECMA* e cioè l'associazione tra enti industriali che sviluppano, producono e commercializzano componenti H&S e servizi nel campo IT&C.

II ARCHITETTURE DI COMUNICAZIONE

La comunicazione tra due o più parti richiede *cooperazione*, e cioè collaborazione per il conseguimento di uno scopo comune. In particolare:

- occorre assicurare il rispetto di opportune *regole procedurali* nel trasferimento dell'informazione e negli adempimenti richiesti per l'utilizzazione di questa;
- mettere in atto, quando è necessario, *provvedimenti protettivi* per fronteggiare eventi di natura aleatoria (ad es. disturbi trasmissivi, errori procedurali, guasti di apparecchiatura, etc.) che potrebbero compromettere lo scambio di informazione.

Occorre in definitiva consentire l'evoluzione di un *processo di comunicazione.* Questo consiste nello svolgere, in forma collaborativa tra le parti coinvolte, una sequenza di funzioni che rendano possibile a una parte, non solo di essere fisicamente connessa con un'altra parte, ma anche di comunicare con quest'ultima nonostante impedimenti di natura varia, quali errori di origine fisica o logica, diversità di linguaggi, etc.

Per realizzare e per gestire processi di comunicazione si è dimostrata indispensabile la disponibilità di una *descrizione astratta delle modalità di comunicazione* tra due o più parti in posizioni tra loro remote e attraverso una infrastruttura di rete. Questa descrizione è un *modello funzionale* che è di riferimento per una rappresentazione dell'*ambiente di comunicazione.* L'identificazione del modello si svolge in vari passi logici.

Il *primo passo*, che corrisponde al più elevato livello di astrazione, riguarda la definizione degli *oggetti* che sono utilizzati per descrivere il processo di comunicazione sotto esame, delle *relazioni generali* tra questi oggetti e dei *vincoli generali* tra questi tipi di oggetti e di relazioni. Come conclusione del primo passo debbono essere definite le funzioni da svolgere e le relative *modalità organizzative* per permetterne uno svolgimento coordinato. Il risultato è *l'architettura della comunicazione*. Elemento distintivo di questa è la presenza costante di *interazioni tra due o più parti.*

L'*ultimo passo,* che corrisponde al più basso livello di astrazione, riguarda la *descrizione dettagliata* delle modalità di esecuzione delle funzioni identificate nel primo passo e consente di specificare le *procedure operative* che debbono essere seguite per ognuna delle interazioni tra le parti in gioco nella architettura di comunicazione. Tali procedure sono i *protocolli di comunicazione*; elementi costituenti sono:

- la *semantica*, e cioè l'insieme dei comandi, delle azioni conseguenti e delle risposte attribuibili alle parti;
- la *sintassi*, ossia la struttura dei comandi e delle risposte;
- la *temporizzazione*, ovvero le sequenze temporali di emissione dei comandi e delle risposte.

Per l'evoluzione di un processo di comunicazione in un ambiente prevalentemente rivolto ad applicazioni telematiche (ma non in modo esclusivo) devono essere svolte svariate funzioni, che possono essere suddivise nei *due insiemi* già considerati all'inizio del precedente cap. I e comprendenti le funzioni *orientate al trasferimento* e quelle *orientate all'utilizzazione*, in cui i termini "trasferimento" e "utilizzazione" hanno come oggetto comune l'*informazione* da scambiare nel corso del processo. Tra le funzioni orientate al trasferimento si citano, a titolo di esempio, la connessione fisica, la trasmissione, il controllo di errore, la multiplazione, la commutazione, il controllo di flusso e il controllo della qualità di servizio. Alle funzioni orientate all'utilizzazione, sempre a titolo di esempio, appartengono la gestione del dialogo, l'adattamento sintattico e gli adempimenti semantici.

Nella trattazione che segue in questo capitolo si supporrà implicitamente che l'informazione da scambiare nel corso del processo di comunicazione non contempli una distinzione tra informazioni di utente, di controllo e di gestione (cfr. § I. 3.1). Si parlerà quindi di flussi informativi aventi un'origine

(sorgente) e una o più destinazioni (collettori), senza alcuna precisazione di quale sia la loro finalità nell'ambito del processo di comunicazione. Tuttavia la distinzione in parola che, come già detto (cfr. § I.3.1), guida gli attuali sviluppi delle infrastrutture per telecomunicazioni, verrà discussa nel par. II.6.

II.1 Architettura a strati

Le architetture di comunicazione attualmente utilizzate per descrivere un qualunque ambiente di comunicazione sono del tipo *a strati* (layered architecture). Ciò corrisponde a una specifica modalità nell'organizzare le funzioni svolte per l'evoluzione dei corrispondenti processi di comunicazione. Per chiarire questa modalità conviene fare riferimento ai criteri del *raggruppamento* e della *gerarchizzazione* di insiemi funzionali.

Sia $\Im$ l'insieme delle funzioni da svolgere per consentire l'evoluzione di un processo di comunicazione. Il punto-chiave consiste nel definire criteri per *partizionare* $\Im$ in sottoinsiemi funzionali e per *organizzare* le modalità di interazione tra questi sottoinsiemi.

Il criterio del *raggruppamento* consiste nel

- considerare appartenenti allo stesso sottoinsieme funzioni simili per logica e per tecnologia realizzativa;
- identificare i sottoinsiemi in modo da minimizzare la complessità e la numerosità delle interazioni tra funzioni appartenenti a sottoinsiemi diversi.

Circa il criterio della *gerarchizzazione*, tre sottoinsiemi *A*, *B* e *C* appartenenti a $\Im$ si dicono in *ordine gerarchico crescente* se:

- lo svolgimento di *B* presuppone la preventiva esecuzione di *A;*
- l'unione di *A* e di *B* costituisce il presupposto per l'esecuzione di *C*.

Se i sottoinsiemi *A*, *B* e *C* sono in ordine gerarchico crescente, il sottoinsieme *B* offre un "*servizio*" a *C* e, per questo scopo, opera in modo da *aggiungere valore* al "servizio" che gli è offerto da *A* (Fig. II.1.1.). Il risultato è che il "servizio" offerto a *C* si presenta con caratteristiche arricchite rispetto a quello offerto a *B (principio del valore aggiunto)*.

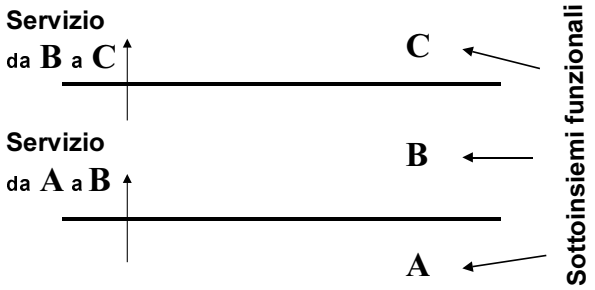


Figura II.1.1 - Il principio del valore aggiunto

Supponiamo allora che l'insieme $\Im$

- sia partizionato in sottoinsiemi funzionali, identificati applicando il criterio del raggruppamento;
- sia organizzato in modo che questi sottoinsiemi operino in un ordine gerarchico.

Coerentemente con queste ipotesi ogni sottoinsieme viene identificato da un *numero* crescente al crescere del livello gerarchico. Supponiamo ulteriormente che la partizione e l'organizzazione di $\Im$ siano effettuate in modo che ciascun sottoinsieme funzionale:

- interagisca solo con i sottoinsiemi che gli sono gerarchicamente "adiacenti";

- si comporti nei confronti di questi secondo il principio del “valore aggiunto”;
- svolga le sue funzioni con modalità indipendenti dagli analoghi svolgimenti negli altri sottoinsiemi.

Il soddisfacimento di queste ipotesi significa applicare il *principio della stratificazione* ad un ambiente di comunicazione. Secondo questo principio ogni sottoinsieme funzionale:

- *riceve “servizio”* dal sottoinsieme che gli è immediatamente inferiore nell’ordine gerarchico;
- *arricchisce questo “servizio”* con il valore derivante dallo svolgimento delle proprie funzioni;
- *offre il nuovo “servizio”* a valore aggiunto al sottoinsieme che gli è immediatamente superiore nell’ordine gerarchico.

L’applicazione di questo principio nella definizione di un’architettura di comunicazione consente di sezionare il complesso problema della organizzazione funzionale di un processo di comunicazione in un insieme di problemi più semplici, ognuno dei quali si riferisce ad un particolare sottoinsieme.

II.2 Elementi architetturali

Elementi fondamentali di un’architettura di comunicazione (Fig. II.2.1) sono

- i *sistemi*, capaci di effettuare trattamento e/o trasferimento di informazione in vista di specifiche applicazioni;
- i *processi applicativi*, limitatamente a quegli aspetti che sono coinvolti da esigenze di interazione con altri processi;
- i *mezzi trasmissivi*, rappresentanti la struttura fisica di interconnessione tra i sistemi.

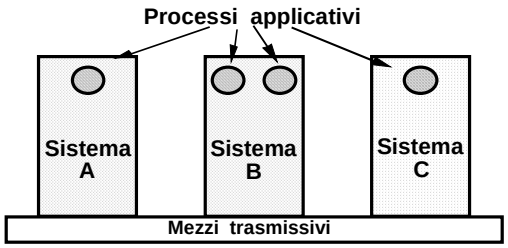


Figura II.2.1 - Elementi architetturali

Per ognuno dei sistemi interconnessi, l’architettura considera solo gli aspetti che riguardano il comportamento *verso l’esterno* e cioè quelli volti alla cooperazione con altri sistemi.

La trattazione che segue comprende le definizioni di altri elementi architetturali, e precisamente:

- degli *strati architetturali* (§ II.2.1);
- del *servizio di strato* (§ II.2.2);
- delle *primitive di servizio* e dei *protocolli di strato* (§ II.2.3);
- della *funzione di indirizzamento* (§ II.2.4);
- dei *sistemi terminali e intermedi* (§ II.2.5).

II.2.1 Gli strati architetturali

Si faccia ora riferimento ad un ambiente di comunicazione, al cui insieme di funzioni $\mathfrak{I}$ si applica il principio della stratificazione; i sistemi interconnessi che compongono l’ambiente vengano posti in corrispondenza con i sottoinsiemi funzionali in cui viene partizionato e organizzato l’insieme $\mathfrak{I}$. Ogni

sistema è allora visto come logicamente composto da una successione ordinata di *sottosistemi*, ad ognuno dei quali è associato un sottoinsieme funzionale. L'ordine definisce il *rango* di ogni sottosistema; il rango di un sottosistema è il numero dell'associato sottoinsieme funzionale nell'organizzazione gerarchica dell'insieme $\Im$. Un sottosistema è quindi quella parte dell'intero sistema che è preposta a svolgere un associato sottoinsieme funzionale tra quelli identificati applicando il principio della stratificazione e che interagisce solo con i sottosistemi di rango immediatamente superiore e immediatamente inferiore.

Tutti i sottosistemi che appartengono a qualunque sistema tra quelli interconnessi e che sono caratterizzati da uguale rango (*sottosistemi omologhi*) formano uno *strato*. Uno strato è quindi l'unione di tutti i sottosistemi omologhi appartenenti a sistemi interconnessi (Fig. II.2.2.). Il *rango di uno strato* è quello dei sottosistemi componenti. Uno strato, nel suo ruolo di sede ove si svolgono funzioni identificate applicando all'insieme $\Im$ il principio della stratificazione, presenta un'operatività che è indipendente da quella degli altri strati dell'architettura.

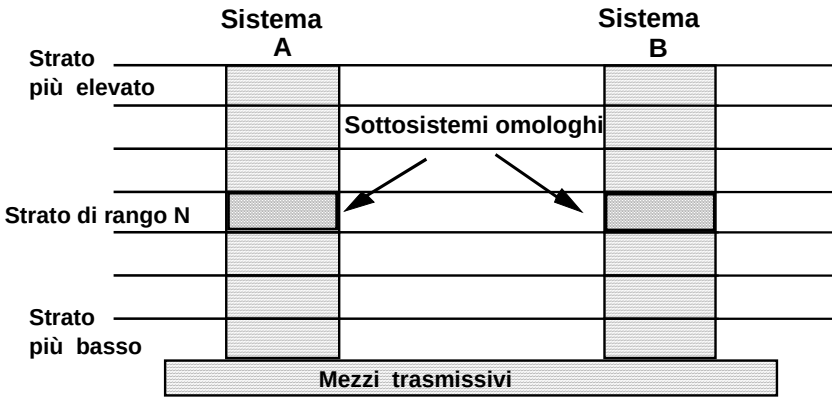


Figura II.2.2 - Definizione di strato architetturale

All'interno di un sottosistema possono identificarsi una o più *entità* che rappresentano *gli elementi attivi* di ogni singolo strato. Un'entità corrisponde a quella parte di un sottosistema che provvede all'esecuzione di una o più tra le funzioni dello strato. Nello svolgimento del processo di comunicazione, entità appartenenti allo stesso strato e residenti nei sistemi interconnessi (*entità alla pari*) interagiscono tra loro per il corretto espletamento delle funzioni a cui sono preposte.

Nel seguito, per indicare qualsiasi elemento della architettura, sarà usata una particolare notazione in cui il nome dell'elemento è preceduto dal numero di rango dello strato a cui si riferisce. Ad esempio, il termine *(N)-entità* identificherà una entità appartenente allo strato di rango *N*. In Fig. II.2.3 sono forniti esempi di questa notazione architetturale.

In taluni casi è necessario estendere la stratificazione anche all'interno di un singolo strato. In tal modo si possono identificare *sottostrati*, corrispondenti a particolari raggruppamenti di funzioni che possono eventualmente non essere eseguite nello svolgimento del processo di comunicazione.

II.2.2 Servizio di strato

Ogni (N)-strato fornisce un *(N)-servizio* all'(N+1)-strato. Per questo scopo utilizza l'(N-1)-servizio e lo arricchisce con lo svolgimento di un particolare sottoinsieme delle (N)-funzioni.

Una *(N)-funzione* è parte delle attività di una (N)-entità ed è svolta sempre mediante la cooperazione di due o più (N)-entità alla pari. Esistono due tipi di (N)-funzioni:

- quelle rientranti tra le funzionalità che caratterizzano l'(N)-servizio e che quindi sono *visibili* all'*(N)-interfaccia* tra l'(N+1)-strato e l'(N)-strato; la visibilità si manifesta con una esplicita notifica dell'esecuzione di queste funzioni;
- quelle finalizzate a conseguire gli obiettivi dell'(N)-strato al di fuori della fornitura dell'(N)-servizio e svolte all'*interno dell'(N)-strato* senza richieste specifiche da parte dell'(N+1)-strato.

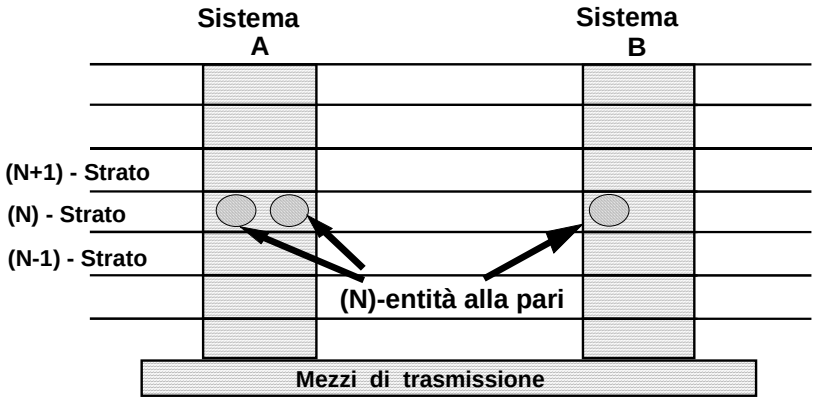


Figura II.2.3 – Esempi di notazioni architetturali

Riguardo alla definizione di (N)-servizio, va sottolineato che:

- lo strato funzionale con rango più elevato non offre servizio ad alcuno strato (dato che ad esso non corrisponde uno strato superiore), ma al contempo riceve la somma dei servizi di tutti gli altri strati;
- lo strato funzionale di rango più basso non riceve servizio da alcuno strato (dato che esso si interfaccia direttamente con i mezzi trasmissivi).

Gli utenti di un (N)-servizio, e cioè gli *(N)-utenti*, sono le (N+1)-entità che ne usufruiscono per i loro scopi di cooperazione. Fornitore di un *(N)*-servizio, e cioè l'*(N)-fornitore*, è invece costituito dall'insieme delle (N)-entità che sono in corrispondenza con le (N+1)-entità agenti come utenti e che cooperano tra loro per svolgere le funzioni caratterizzanti l'(N)-servizio. È da sottolineare come i termini "utente" e "fornitore" hanno qui un significato che è limitato al rapporto fra due strati adiacenti e che ben si differenzia da quello introdotto nel precedente capitolo.

Si osserva poi che i ruoli di utente e di fornitore nei confronti di un servizio di strato possono invertirsi quando si passa da uno strato a quello immediatamente inferiore ovvero a quello immediatamente superiore. Infatti le (N)-entità che cooperano nella fornitura dell'(N)-servizio possono diventare utenti dell'(N-1)-servizio, mentre le (N+1)-entità che agiscono come utenti dell'(N)-servizio possono cooperare nella fornitura dell'(N+1)-servizio.

La corrispondenza tra gli (N)-utenti e le (N)-entità che costituiscono l'(N)-fornitore (Fig. II.2.4) è di natura logica ed è attuata attraverso specifici punti dell'(N)-interfaccia, chiamati *(N)-punti d'accesso al servizio* [*(N)-SAP*, *Service Access Point*]. Un (N)-SAP è quindi il varco attraverso cui avviene la comunicazione tra una (N+1)-entità nel suo ruolo di utente di un (N)-servizio e una (N)-entità che risiede nello stesso sistema e che coopera per la fornitura di quel servizio.

Per ragioni legate alla univocità dell'indirizzamento (cfr. § II.2.4), un (N)-SAP può essere servito da una sola (N)-entità e essere utilizzato da una sola (N+1)-entità. Tuttavia una (N)-entità può servire vari (N)-SAP e una (N+1)-entità può utilizzare vari (N)-SAP (Fig. II.2.5).

II.2.3 Primitive di servizio e protocolli di strato

Un (N)-servizio è in generale logicamente composto da un insieme di *elementi di servizio*, ognuno dei quali corrisponde ad un particolare sottoinsieme delle (N)-funzioni. Le interazioni elementari tra utente e fornitore del servizio attraverso un (N)-SAP, necessarie per attuare un elemento di servizio, sono comunemente chiamate *primitive di servizio*. Queste possono essere di quattro tipi:

- *primitiva di richiesta*, che è emessa da un (N)-utente per richiedere l'attivazione di un particolare elemento di servizio;
- *primitiva di indicazione,* che è emessa dall'(N)-fornitore per indicare che l'(N)-utente remoto ha presentato una richiesta di attivazione di un elemento di servizio; lo stesso tipo di primitiva può essere utilizzata dall'(N)-fornitore per indicare all'(N)-utente l'attivazione di una particolare procedura interna al servizio stesso (ad esempio nel caso di notifica d'errore);
- *primitiva di risposta,* che è emessa dall'(N)-utente per completare la procedura relativa ad un elemento di servizio precedentemente attivato da una primitiva di indicazione;
- *primitiva di conferma,* che è emessa dall'(N)-fornitore per completare la procedura relativa ad un elemento di servizio precedentemente attivato da una primitiva di richiesta.

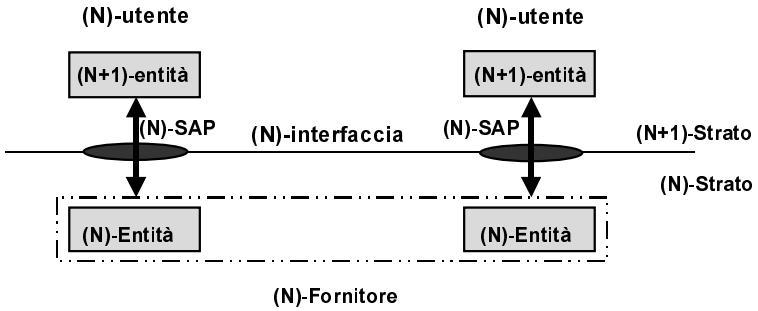


Figura II.2.4 - Definizione di servizio di strato

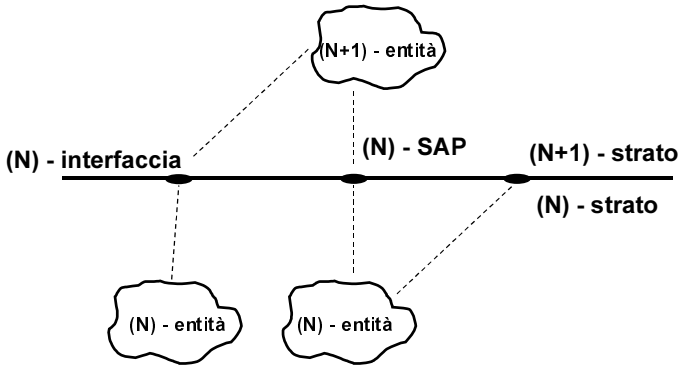


Figura II.2.5 - Definizione di punto di accesso al servizio

È necessario osservare che la definizione delle primitive di servizio non si riferisce necessariamente ad un tipo particolare di realizzazione, ma rappresenta esclusivamente un modello astratto secondo il quale l'utente e il fornitore del servizio interagiscono tra loro.

Un elemento di servizio, in relazione alle primitive necessarie alla sua attuazione, può essere classificato nel modo seguente:

- *servizio confermato*, se è richiesto dall'(N)-utente e se necessita di un esplicita conferma da parte dell'(N)-fornitore; la sua attuazione può richiedere sia l'impiego di tutti i tipi di primitive, sia essere limitata all'impiego delle primitive di richiesta e di conferma;
- *servizio non-confermato*, se è richiesto dall'(N)-utente, ma non necessita di conferma da parte dell'(N)-fornitore; richiede l'utilizzazione delle sole primitive di richiesta e di indicazione;
- *servizio iniziato dal fornitore*, se è attivato autonomamente dall'(N)-fornitore; richiede l'utilizzazione delle sole primitive di indicazione.

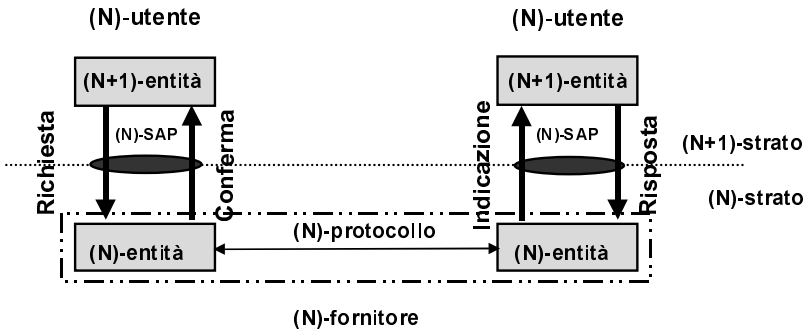


Figura II.2.6 - Modello e primitive dell'(N)-servizio

La Fig. II.2.6 illustra il modello astratto delle interazioni tra utente e fornitore di un generico (N)-servizio, mentre la Fig. II.2.7 illustra gli esempi di servizio confermato, non confermato e iniziato dal fornitore.

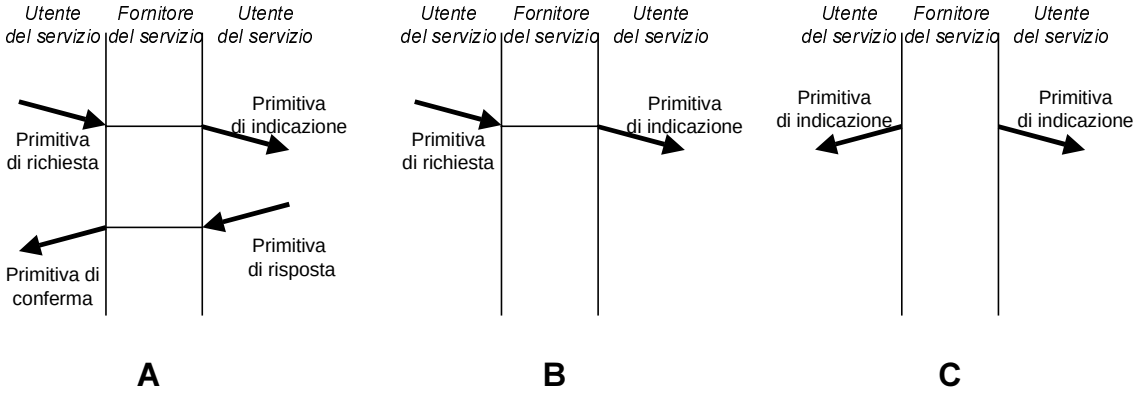


Figura II.2.7 - Classificazione dei servizi di strato:(A) servizio confermato; (B) servizio non confermato; (C) servizio iniziato dal fornitore

La cooperazione tra (N)-entità residenti in sistemi diversi (Fig. II.2.8) è governata da un insieme di regole che prende il nome di *(N)-protocollo.* Questo è un insieme di regole necessarie perché l'(N)-servizio sia realizzato. Le regole definiscono in particolare i meccanismi che consentono di trasferire l'informazione nell'ambito dell'(N)-strato operante come fornitore del'(N)-servizio.

Conseguentemente una architettura di comunicazione è caratterizzata da uno o più protocolli per ognuno degli strati che compongono l'architettura. Vale la pena aggiungere che:

- ogni protocollo di strato, come insieme di comandi e di risposte che sono scambiate tra le parti interagenti, include una *semantica*, una *sintassi* e una *temporizzazione*: il significato di questi termini è stato chiarito all'inizio di questo capitolo;

- la comunicazione diretta tra (N)-entità residenti nello stesso sistema non è visibile all'esterno di questo e non è quindi oggetto di interesse nella definizione dell'architettura;
- può non esistere una corrispondenza tra primitive di servizio e procedure dei protocolli di strato.

In Fig. II.2.9 sono rappresentati gli elementi architetturali che sono stati definiti in questa sezione e in quella precedente.

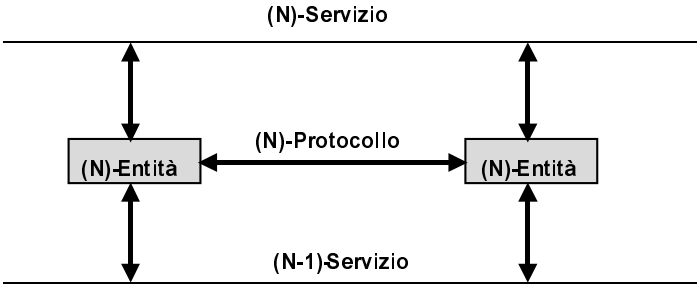


Figura II.2.8 - Interazioni tra entità alla pari

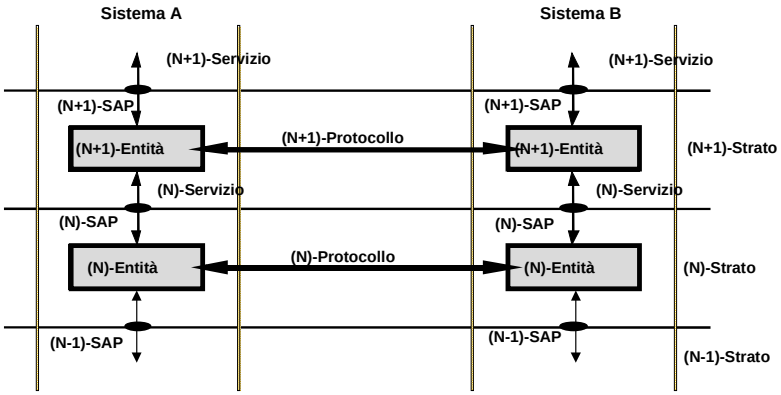


Figura II.2.9 - Elementi di strato

II.2.4 Funzione di indirizzamento

Gli oggetti entro uno strato o alla frontiera tra due strati adiacenti debbono essere identificati in modo univoco. A tale scopo vengono definiti *identificatori* per le entità e per i SAP.

Una (N)-entità è identificata con una *denominazione globale*, che la individua in modo univoco in tutto l'insieme dei sistemi interconnessi. Entro un dominio più limitato, una (N)-entità può essere identificata con una *denominazione locale*, che la individua in modo univoco entro quel dominio. Per esempio, entro il dominio corrispondente all'(N)-strato, le (N)-entità sono identificate con (N)-denominazioni locali, che sono univoche entro quello strato.

Ogni (N)-SAP è identificato da un *(N)-indirizzo*, che localizza in modo univoco l'(N)-SAP a cui è allacciata una specifica (N+1)-entità. Data la corrispondenza uno ad uno tra un (N)-SAP e una (N+1)-entità, l'uso dell'indirizzo di questo (N)-SAP per identificare questa (N+1)-entità è il meccanismo di indirizzamento più efficiente, se può essere assicurato un allacciamento permanente tra questa (N+1)-entità e questo (N)-SAP.

Gli allacciamenti tra (N)-entità e gli (N-1)-SAP che esse utilizzano per comunicare tra loro sono definiti in una particolare (N)-funzione, detta di *(N)-repertorio* (directory). Questa traduce le denominazioni globali delle (N)-entità negli (N-1)-indirizzi, attraverso i quali dette entità possono essere raggiunte e quindi cooperare tra loro.

Infine l'interpretazione della corrispondenza tra gli (N)-indirizzi serviti da una (N)-entità e gli (N-1)-indirizzi utilizzati per accedere agli (N-1)-servizi che l'(N)-entità utilizza (e quindi per identificare quest'ultima) è effettuata da un'altra particolare (N)-funzione, che è detta di *corrispondenza di indirizzo* (address mapping).

II.2.5 Sistemi terminali e intermedi

Un sistema impegnato in un processo di comunicazione (Fig. II.2.10) può essere *terminale* o *intermedio* (o di rilegamento). Un *sistema terminale* (End System) è origine o destinazione finale delle informazioni. Un *sistema intermedio* (Relay System) provvede ad assicurare il rilegamento fisico o logico tra due o più sistemi terminali.

Ogni sistema terminale o intermedio comprende una successione ordinata di sottosistemi, con rango che va dal *valore più basso* della gerarchia ad un *valore massimo*, che non è necessariamente quello più elevato dell'architettura. Un sistema terminale comprende di norma *tutti i ranghi* dell'architettura; un sistema intermedio ne comprende di norma solo un *sottoinsieme*.

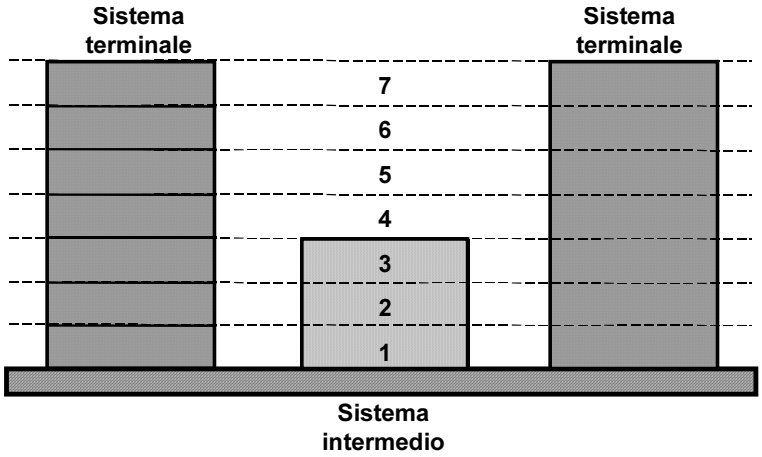


Figura II.2.10 - Sistemi interconnessi

Il numero di sottosistemi in un sistema intermedio è quindi solitamente inferiore a quello di un sistema terminale. Partendo dal sottosistema di rango più basso e procedendo verso sottosistemi di rango via via crescente, si perviene al sottosistema che è in grado di svolgere una *funzione di rilegamento*. Questa, se svolta nell'(N)-strato, consente ad una (N)-entità di porre in corrispondenza una (N)-entità a monte con una (N)-entità a valle. Il numero di sottosistemi inclusi in un sistema intermedio dipende quindi dalla scelta architetturale di quale sia lo strato in cui svolgere la funzione di rilegamento.

II.3 Trattamenti dell'informazione

Si analizzano i trattamenti dell'informazione in una architettura di comunicazione relativa un insieme di *sistemi interconnessi*.

Ogni sistema terminale svolge il ruolo di *emittente* quando è *origine* di informazioni e di *ricevente* quando ne è *destinazione*. Un sistema intermedio svolge, sempre e contemporaneamente, i ruoli di emittente e di ricevente in relazione al verso del flusso informativo che lo interessa: quando il verso del flusso è *entrante* nel sistema, questo svolge il ruolo di *ricevente*; quando il verso è invece *uscente*, il ruolo svolto è di *emittente*.

Indipendentemente dal loro ruolo, i sistemi (terminali o intermedi) hanno il compito di *trattare l'informazione* loro consegnata; il trattamento nei sistemi terminali ha in generale lo scopo di trasferire e utilizzare l'informazione, mentre quello nei sistemi intermedi può limitarsi a soli aspetti di trasferimento. Il trattamento in un sistema consiste nello svolgimento delle funzioni attribuite ai sottosistemi componenti.

Nel seguito del paragrafo vengono utilizzati i concetti architetturali messi a fuoco in precedenza per chiarire:

- quale sia l'*ordine logico* di svolgimento delle funzioni che concorrono all'evoluzione di un processo di comunicazione (§ II.3.1);
- come questa evoluzione possa essere analizzata facendo riferimento ad un insieme di comunicazioni di strato aventi natura virtuale (*comunicazioni intra-strato*) (§ II.3.2);
- come nell'operatività del modello siano identificabili due tipi di *flussi informativi*, che sono il veicolo di supporto delle *informazioni di utenza* e *di coordinamento* (§ II.3.3);
- quale sia l'evoluzione dei protocolli di strato in relazione a differenti *strategie di gestione* (§ II.3.4);
- quali modi alternativi di servizio di strato siano offerti per la realizzazione di una comunicazione intra-strato(§ II.3.5).

II.3.1 Ordine logico di trattamento

Lo svolgimento di una (N)-funzione (in quanto è sempre risultato della cooperazione di due o più parti) richiede la *compartecipazione* di due o più (N)-sottosistemi omologhi. La compartecipazione si manifesta in

- uno *svolgimento in apertura*, quando l'entità di un sottosistema prende l'iniziativa e attiva la funzione che le è pertinente;
- uno *svolgimento in chiusura*, quando viene completata l'attivazione della funzione da parte della (o delle) entità alla pari che è (o sono) in corrispondenza con l'entità iniziatrice.

Uno svolgimento in apertura viene effettuato da un sottosistema appartenente ad un sistema con il *ruolo di emittente*; uno svolgimento in chiusura viene invece effettuato da un sottosistema appartenente ad un sistema con il *ruolo di ricevente*.

Quando un sistema (terminale o intermedio) ricopre il ruolo di emittente, il suo trattamento dell'informazione (consistente nello svolgere in apertura le funzioni attribuite ai suoi sottosistemi componenti) segue uno *specifico ordine logico*, che è quello dei *ranghi decrescenti*. L'ordine di trattamento è quindi il seguente (Fig. II.3.1):

- *si inizia* dal sottosistema di *rango più elevato* che è presente nel sistema;
- *si scende* via via attraverso i sottosistemi di rango intermedio fino al sottosistema di *rango più basso* nell'architettura;
- *si chiude consegnando* l'informazione al *mezzo trasmissivo.*

Quando un sistema (terminale o di rilegamento) svolge invece il ruolo di ricevente, il suo trattamento dell'informazione (consistente nello svolgere *in chiusura* le funzioni attribuite ai suoi sottosistemi componenti) segue uno *specifico ordine logico* che è quello dei *ranghi crescenti.* L'ordine di trattamento è quindi il seguente (Fig. II.3.2):

- *si parte* estraendo l'informazione dal mezzo trasmissivo;
- *si inizia* il trattamento dal sottosistema di rango più basso nell'architettura;
- *si sale* via via attraverso i sottosistemi di rango intermedio fino al sottosistema di *rango più elevato* che è presente nel sistema.

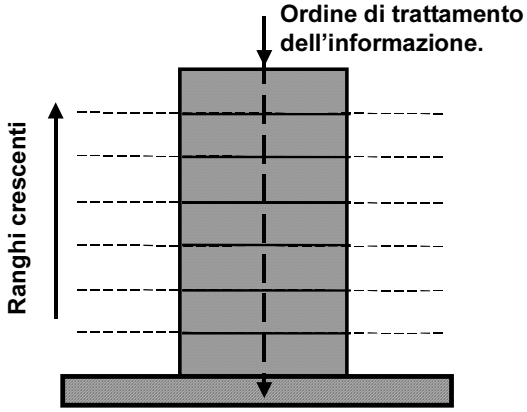


Figura II.3.1 - Sistema emittente

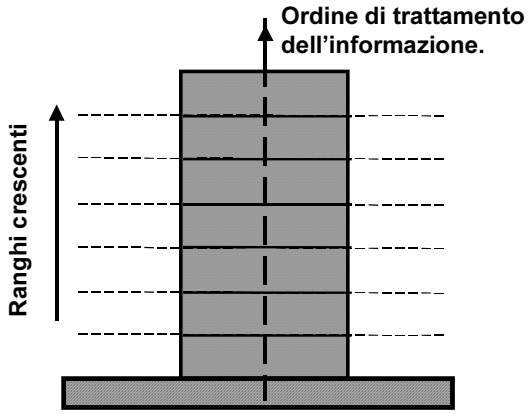


Figura II.3.2 - Sistema ricevente

Con queste premesse, definiamo l'ordine logico di trattamento in *due o più sistemi interconnessi*. A questo scopo individuiamo per ogni coppia di sistemi (tra quelli interconnessi)

- qual'è l'emettitore e qual'è il ricevitore;
- quali sono i sottosistemi omologhi in essi inclusi che *interagiscono direttamente* nello svolgimento delle funzioni loro pertinenti.

Allora, detto *A* un sistema terminale emittente e *B* quello terminale ricevente, ad uno svolgimento in apertura da parte di un sottosistema incluso in *A* corrisponde uno svolgimento in chiusura nel sottosistema omologo al precedente e incluso in *B* (Fig. II.3.3). La stessa conclusione si applica nel caso di sistemi intermedi, ove si hanno sempre interfacce con ruolo ricevente e altre con ruolo emittente (Fig. II.3.4).

II.3.2 Comunicazioni di strato

Nel trattamento dell'informazione che si svolge nei sistemi interconnessi (terminali o intermedi), due o più (N+1)-entità alla pari, se coinvolte nello svolgimento cooperativo di una *(N+1)-funzione*, hanno necessità di comunicare tra loro per il conseguimento delle loro finalità (*comunicazione intra-strato).* L'informazione scambiata in questa comunicazione è quella necessaria a consentire cooperazione nello svolgimento della (N+1)-funzione. La comunicazione avviene tra le (N+1)-entità per il tramite dell'(N)-strato e rientra tra le opportunità offerte dall'(N)-servizio. In questa loro comunicazione le (N+1)-entità svolgono il ruolo di (N)-utenti, mentre l'(N)-fornitore è rappresentato dalle (N)-entità che sono in corrispondenza con le (N+1)-entità comunicanti tramite i pertinenti (N)-SAP.

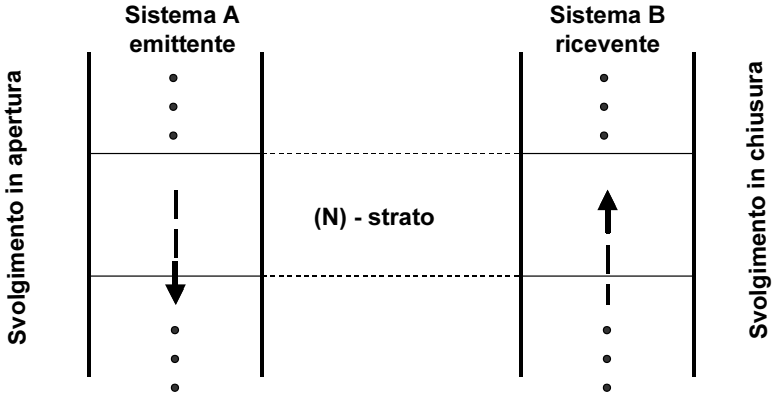


Figura II.3.3 - Ordine di trattamento in sistemi emittenti e riceventi

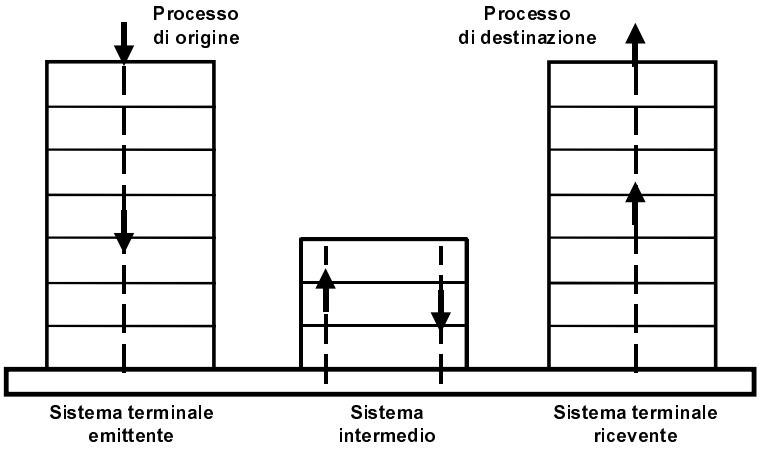


Figura II.3.4 - Ordine di trattamento in sistemi terminali e intermedi

La (*N+1)-entità emittente* consegna l'informazione all'(N)-fornitore attraverso il proprio (N)-SAP; l'informazione perviene così all'*(N)-entità di origine* che è in corrispondenza con la (N+1)-entità emittente attraverso l'(N)-SAP in comune. L'(N)-fornitore consegna detta informazione all'*(N+1)-entità ricevente* attraverso il suo (N)-SAP; l'informazione viene consegnata per il tramite della *(N)-*

entità di destinazione che è in corrispondenza con la (N+1)-entità ricevente attraverso l'(N)-SAP in comune (Fig. II.3.5).

Per effetto dell'(N)-servizio i due (N)-utenti, che percepiscono solo la presenza dei loro (N)-SAP, vedono un'*immagine* della comunicazione intra-strato che li pone in *corrispondenza diretta* nell'ambito dell'(N+1)-strato; questa corrispondenza è però solo *virtuale*, dato che la comunicazione intra-strato avviene in *modo indiretto* per l'opera intermediaria dell'(N)-fornitore. L' (N)-servizio consente quindi di stabilire tra gli (N)-utenti una *comunicazione virtuale* che è immagine (per questi utenti) della *comunicazione effettiva* assicurata dall'(N)-fornitore.

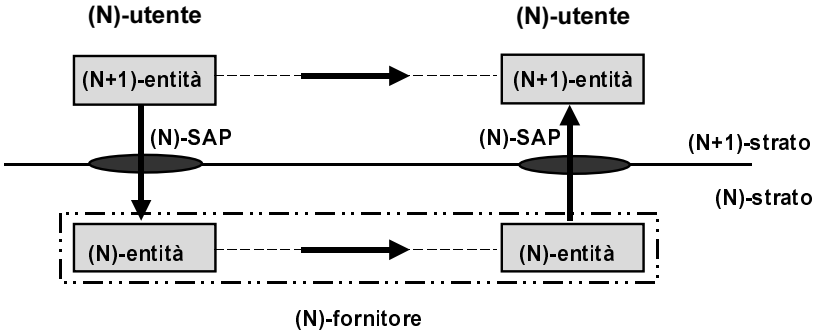


Figura II.3.5 - Svolgimento della comunicazione

L'(N)-fornitore è in grado di svolgere questo compito in quanto si avvale dell'*(N-1)-servizio;* più in particolare le (N)-entità, che partecipano alla comunicazione effettiva, sono poste in comunicazione virtuale del tutto analoga a quella riguardante le (N+1)-entità; questa mappatura

- *da comunicazione effettiva* tra entità alla pari basata sulle funzionalità degli strati inferiori
- *a comunicazione virtuale* che si svolge direttamente nell'ambito dello strato a cui appartengono le entità comunicanti,

è di *natura iterativa*, interessando via via gli strati di rango decrescente fino a raggiungere lo strato più basso dell'architettura. Le entità dello strato più basso, essendo direttamente a contatto con i mezzi trasmissivi, sono le uniche dell'architettura che possono svolgere effettivamente una comunicazione diretta.

Ricaviamo ora la relazione tra la comunicazione intra-strato e quella più generale tra processi applicativi residenti in sistemi diversi. Quest'ultima per facilità di riferimento, verrà indicata come *comunicazione tra processi*.

Sulla base delle considerazioni precedenti, la comunicazione intra-strato costituisce al tempo stesso

- il rapporto che si stabilisce in ogni coppia di strati contigui tra utenti del servizio di strato e il relativo fornitore;
- la riproduzione "in piccolo" della *comunicazione tra processi*.

La comunicazione tra processi può allora essere vista come partizionata in un insieme di comunicazioni intra-strato in corrispondenza con gli strati dell' architettura dal rango più elevato a quello più basso. Le componenti di questo insieme hanno le seguenti proprietà:

- si presentano con caratteristiche generali *omogenee,* pur con diversità che non ne compromettono l'inter-lavoro;
- *cooperano* secondo i meccanismi dei servizi di strato, per conseguire l'obiettivo della comunicazione tra processi;

- *evolvono in modo indipendente* l'una dall'altra, come è assicurato dalla proprietà di indipendenza degli strati architetturali.

II.3.3 Flussi informativi

Una comunicazione tra processi coinvolge le potenzialità di trattamento dell'informazione che sono rese disponibili dai sistemi interconnessi (terminali o intermedi). In questo coinvolgimento si dicono *adiacenti* i sistemi che interagiscono *direttamente* nello svolgimento delle funzioni loro pertinenti.

Per ogni insieme di sistemi adiacenti, il coinvolgimento si manifesta

- in *termini virtuali*, tramite tanti *scambi informativi intra-strato* quanti sono gli strati comuni presenti in detti sistemi e con trasferimento tra *sottosistemi omologhi*;
- in *termini effettivi*, tramite *scambi informativi inter-strato* con attraversamento, all'*interno di ogni sistema,* degli strati comuni con gli altri sistemi adiacenti.

Tali scambi informativi determinano due tipi di *flussi* (Fig. II.3.6):

- uno tra entità alla pari (residenti quindi in sistemi diversi); è il *flusso intra-strato*;
- l'altro tra entità appartenenti a strati contigui nell'ambito dello stesso sistema; è il *flusso inter-strato*.

Il flusso intra-strato viene trasferito in *modo indiretto* usando il servizio offerto dallo strato inferiore; ha il verso che va dal sistema con il ruolo di emittente a quello con il ruolo di ricevente. Il flusso inter-strato viene trasferito in *modo diretto*; ha il verso dell'ordine logico di trattamento dell'informazione nell'ambito del sistema considerato.

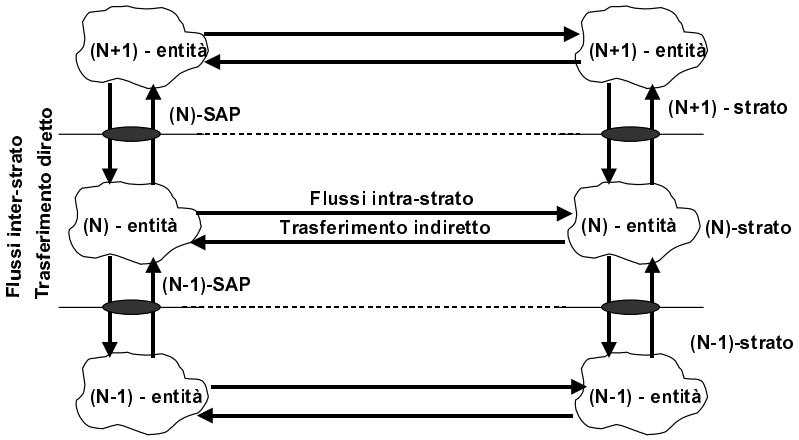


Figura II.3.6 - Trasferimenti di flusso

Dato che ad ogni sistema con il ruolo di emittente si accompagna almeno un sistema adiacente con il ruolo di ricevente, lo scambio informativo tra questi due sistemi comporta, congiuntamente, due tipi di *flusso inter-strato* (Fig. II.3.7): uno, riguardante il sistema emittente, *discende* i ranghi degli strati dell'architettura; l'altro, riguardante il sistema ricevente, *risale* invece questi ranghi.

Circa la *natura* di questi flussi in relazione all'evoluzione di un processo di comunicazione, conviene ancora una volta fare riferimento alla relazione (N)-utente/(N)-fornitore nell'ambito dello svolgimento di una comunicazione intra-strato tra (N)-utenti. Si distinguono allora due tipi di informazione:

- quella di *utenza*, che è l'oggetto dello scambio in quella comunicazione intra-strato e che è quindi nell'interesse degli utenti dell'(N)-servizio nel loro ruolo di compartecipazione alla comunicazione tra processi;
- quella *di coordinamento*, che ha lo scopo di gestire le azioni da svolgere in modo cooperativo nell'(N)-strato secondo gli obiettivi e le esigenze prestazionali di quel processo di comunicazione.

Le informazioni di utenza si distinguono in:

- *dati intra-strato*, che sono trasferiti nell'ambito di un flusso intra-strato;
- *dati inter-strato*, che sono trasferiti nell'ambito di un flusso inter-strato.

Le informazioni di coordinamento si distinguono in:

- informazioni *di protocollo (PCI, Protocol Control Information)* che sono scambiate *tra entità alla pari* e corrispondono alle regole di interazione previste nel pertinente protocollo di strato; compongono anch'esse i flussi intra-strato;
- le informazioni di *interfaccia (ICI, Interface Control Information),* che sono scambiate tra entità residenti nello stesso sistema e appartenenti a *strati contigui*; fanno parte dei flussi inter-strato.

E' da sottolineare che le ICI sono scambiate per coordinare lo svolgimento di funzioni del servizio di strato; un ruolo di questo tipo è ricoperto dalla primitive di servizio.

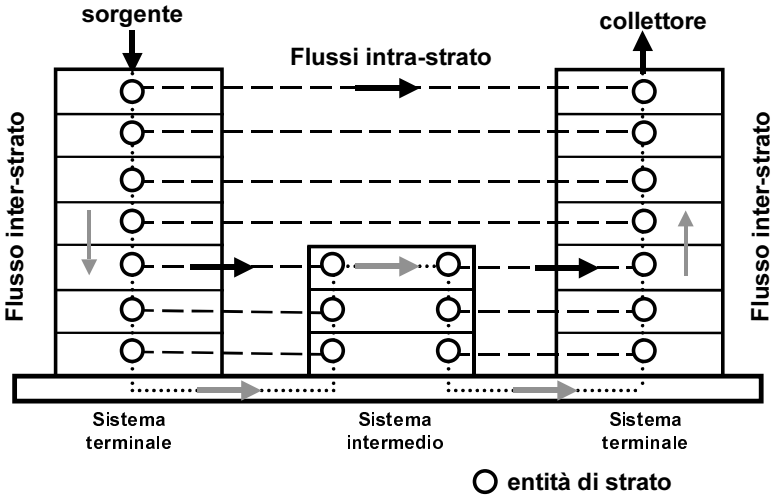


Figura II.3.7 - Flussi informativi inter-strato e intra-strato

II.3.4 Gestione dei protocolli

L'interazione tra due o più entità alla pari (effettuata per coordinare la loro azione nello svolgimento delle funzioni loro pertinenti) deve rispettare le regole procedurali che sono definite in un *protocollo di strato*; in particolare, se si fa riferimento alla interazione tra due (N)-entità, deve verificarsi tra queste lo scambio della pertinente informazione di un (N)-protocollo o, più in breve, di (N)-PCI. Nel caso in cui le due entità siano incluse in sistemi adiacenti *A* e *B* (Fig. II.3.8), l'interazione si manifesta con una *apertura di istanza* e con una corrispondente *chiusura*:

- l'*apertura di istanza* è effettuata dall'*entità "iniziatrice"* e consiste nell'emissione di un *comando* rivolto all'altra entità, che per comodità indichiamo come "*remota*";
- la *chiusura di istanza* è effettuata dall'entità remota che, ricevuto il comando, provvede ad attuare quanto ivi richiesto e, eventualmente, ad inviare una *risposta* all'entità iniziatrice.

L'evoluzione di un (N)-protocollo avviene allora attraverso *aperture* e successive *chiusure di istanza*; tali aperture e chiusure, rispettando la sintassi e la temporizzazione del protocollo, si ripresentano nel tempo fino al conseguimento dello scopo desiderato. L'attuazione di questa evoluzione verrà qui chiamata "*gestione del protocollo*".

Un protocollo di strato può essere gestito in vari modi in relazione alla collocazione reciproca delle entità interagenti nell'ambito dei sistemi interconnessi. Si possono avere una *gestione da estremo a estremo* ovvero una *gestione di sezione*.

Si ha una *gestione da estremo a estremo* quando le regole protocollari riguardano l'interazione diretta tra entità alla pari residenti in sistemi terminali che sono origine e destinazione dell'informazione, e cioè quando si prescinde dalla presenza di eventuali sistemi intermedi. Una gestione siffatta riguarda solitamente protocolli di strato alto nell'architettura (Fig. II.3.9).

Si ha invece una *gestione di sezione* quando, essendo presenti uno o più sistemi intermedi sul percorso dall'origine alla destinazione dell'informazione, le regole protocollari riguardano l'interazione diretta tra entità alla pari residenti in sistemi (terminali e/o intermedi) che sono adiacenti su detto percorso. Una gestione siffatta riguarda solitamente protocolli di strato basso nell'architettura. (Fig. II.3.9).

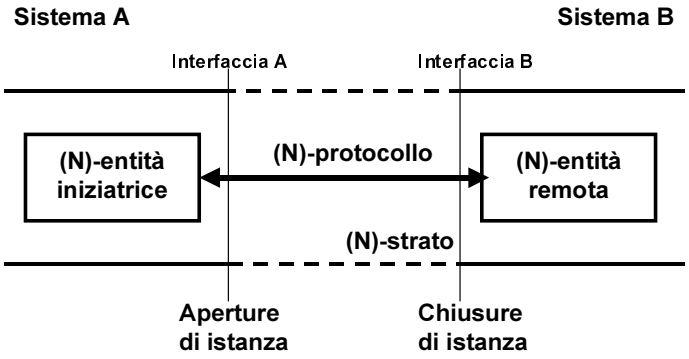


Figura II.3.8 – Interazione protocollare tra entità alla pari

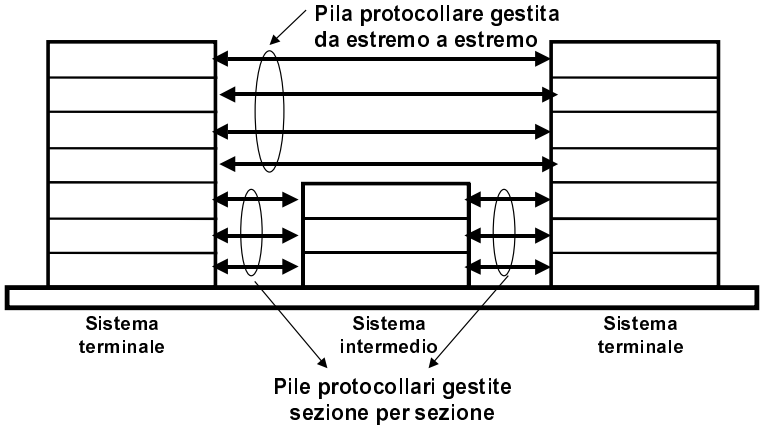


Figura II.3.9 – Pile protocollari gestite da estremo a estremo e sezione per sezione

Si ha poi una *gestione sezione per sezione* nel caso di una *sequenza* di gestioni protocollari di sezione, lungo il percorso dall'origine alla destinazione. In questo caso, si consideri il passaggio da una sezione ad un'altra. Occorre allora trattare le due sezioni in modo separato, facendo riferimento alle

due interfacce coinvolte, quella *a monte* e quella *a valle*, e ai due corrispondenti protocolli, che chiamiamo α e β. Ad una apertura/chiusura di istanza per il protocollo α relativo all'*interfaccia a monte* deve corrispondere una apertura/chiusura di istanza per quello β relativo all'*interfaccia a valle.* Se i protocolli α e β sono *omogenei*, il passaggio da una sezione all'altra avviene per semplice riproposizione dello stesso protocollo (*rigenerazione*). Se invece i protocolli α e β sono *diversi*, il passaggio in questione richiede una *conversione* tra il protocollo α e quello β.

Compito di un sistema intermedio (Fig.II.3.10) che interconnette ambienti di comunicazione caratterizzati da pile protocollari tra loro omogenee è quindi la *rigenerazione* dei protocolli. Occorre invece una *conversione* dei protocolli quando il passaggio è tra ambienti di comunicazione tra loro diversi.

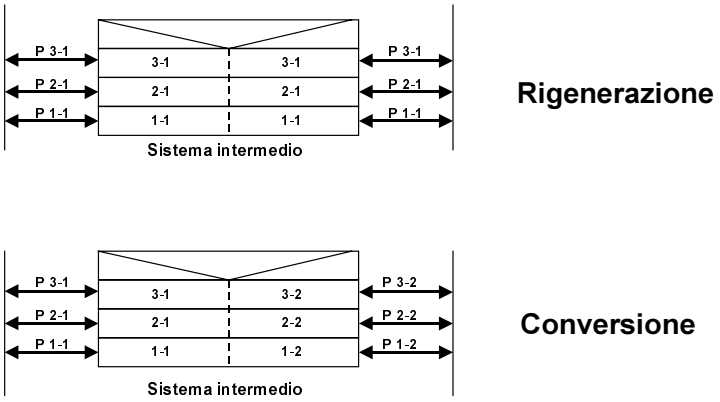


Figura II.3.10 – Interconnessione con rigenerazione o con conversione:
[Pi-j: protocollo di strato i appartenente alla pila j]

II.3.5 Modi di servizio

Un servizio di strato può essere fruito dalle parti interessate con o senza una loro intesa preliminare. Nel caso in cui l'intesa sussista si parla di *servizio con connessione* (connection oriented), facendo riferimento con questo termine a un legame, almeno *logico* e in alcuni casi anche *fisico*, che viene stabilito tra le parti in comunicazione. Nel caso contrario si tratta di un *servizio senza connessione* (connectionless). La scelta di una di queste possibilità nel caso di uno specifico strato *non condiziona* in alcun modo la scelta relativa agli strati contigui.

Un *servizio di strato con connessione* ha alcune caratteristiche distintive:

- una strutturazione in *tre fasi temporali*;
- un accordo tra *tre parti*;
- una *negoziazione* dei parametri di trasferimento;
- un uso di indirizzamento con *identificatori di connessione;*
- un *legame logico* tra i segmenti informativi scambiati.

Circa l'*organizzazione temporale* del servizio, questo si espleta attraverso tre distinte fasi in sequenza: l'*instaurazione della connessione*, il *trasferimento dell'informazione* e l'*abbattimento della connessione*. La vita di una connessione può essere di lunga durata e manifestarsi in molti scambi separati tra le parti connesse; ovvero può essere compressa in una interazione di durata molto breve, in cui l'informazione necessaria per instaurare la connessione, quella da trasferire e quella necessaria per abbattere la connessione sono trasportate in un piccolo numero di scambi.

L’instaurazione di una (N)-connessione richiede il preventivo *accordo tra almeno tre parti*: i due o più (N)-utenti che desiderano comunicare e l’(N)-fornitore che mette a disposizione i mezzi per soddisfare questa esigenza. Queste parti debbono preliminarmente manifestare la loro volontà di partecipare alla comunicazione. Perciò, per tutta la durata della connessione, tali parti sono vincolate all'accordo iniziale.

Nell'ambito di questo accordo, debbono essere negoziati i parametri e le opzioni che governeranno il trasferimento di informazione tra gli (N)-utenti. Perciò una richiesta di instaurazione può essere rifiutata da una parte se i parametri e le opzioni scelti dalla controparte sono per lei inaccettabili. La negoziazione può consentire alle parti di scegliere specifiche procedure, quali quelle riguardanti la qualità del servizio (QoS, Quality of Service), la sicurezza e la protezione dell'informazione. Inoltre l'accordo che risulta dalla negoziazione può, in alcuni casi, essere modificato (e cioè rinegoziato) dopo che la connessione sia stata instaurata e sia iniziata la fase di trasferimento dell'informazione.

Una volta che una connessione sia stata instaurata, essa può essere utilizzata per trasferire segmenti di una sequenza informativa fintantoché la connessione non venga rilasciata da una delle parti in comunicazione. Questi segmenti sono mutuamente legati in virtù del fatto che sono trasferiti su una particolare connessione. Più specificamente, per effetto del trasferimento ordinato sulla stessa connessione

- possono essere facilmente rivelate e recuperate condizioni di *fuori-sequenza*, di *perdita* e di *duplicazione* riguardanti differenti segmenti della sequenza informativa;
- possono essere impiegate tecniche di *controllo di flusso* per assicurare che il ritmo di trasferimento tra le parti in comunicazione non superi quello che queste parti sono in grado di trattare.

Passando ora ai servizi di strato *senza connessione* e alle loro caratteristiche distintive, queste sono:

- *una sola fase temporale*;
- un accordo tra *solo due parti*;
- un’*assenza di negoziazione*;
- un uso di *indirizzi* per l’*origine* e la *destinazione*;
- una *indipendenza* e *autoconsistenza* dei segmenti informativi scambiati.

In questo caso esiste infatti la sola fase di trasferimento dell’informazione e deve sussistere tra gli (N)-utenti del servizio solo una conoscenza mutua a priori. Esistono invece accordi individuali tra ogni utente e il fornitore del servizio, che deve essere sempre disponibile al trasferimento richiesto salvo esplicito avviso contrario. Tra le parti non viene scambiata, preliminarmente, alcuna informazione che riguardi la loro volontà di impegnarsi in una comunicazione. Inoltre tali parti possono non essere contemporaneamente attive. È sufficiente infatti che l’(N)-utente di origine sia attivo solo per il tempo necessario all'emissione delle informazioni, mentre l’(N)-utente di destinazione lo deve essere solo al momento della ricezione. Non c’è quindi possibilità di negoziare opzioni o prestazioni desiderate dagli (N)-utenti ivi compresi gli aspetti di QoS.

Gli indirizzi devono identificare in modo completo l’origine e la destinazione dell’informazione, con conseguente esigenza di un maggior volume di extra-informazione rispetto a quella utile. Infine in un servizio senza connessione debbono essere sottolineate la *indipendenza* e l’*autoconsistenza* dei segmenti informativi trasferiti. Circa l'indipendenza, questa implica che una sequenza di segmenti informativi può essere recapitata a destinazione in un ordine diverso da quello di consegna all'origine. Per ciò che riguarda poi l'autoconsistenza, ogni segmento informativo deve contenere tutta l'informazione necessaria per essere consegnata a destinazione. Questa caratteristica, se da un lato migliora la robustezza del servizio, dall'altro comporta una minore efficienza (a causa del maggior peso della extra-informazione) e una ridotta possibilità di controllo della QoS.

II.4 Unità informative

Indipendentemente dalla loro natura, le informazioni scambiate nell'ambito di una architettura di comunicazione sono strutturate in *unità informative,* e cioè in segmenti di informazione per i quali vengono precisati il formato e la finalità.

Occorre distinguere tra le informazioni (di utenza e/o di coordinamento) che compongono il flusso intra-strato e quello inter-strato (cfr. § II.3.3). Con riferimento allora all'(N)-protocollo e all'(N)-interfaccia come in tabella II.4.1, si possono avere

- l'*unità di dati dell'(N)-protocollo* [*(N)-PDU, Protocol Data Unit*], che è elemento componente del flusso intra-strato nell'ambito dell'(N)-strato;
- l'*unità di dati dell'(N)-interfaccia* [*(N)-IDU, Interface Data Unit*], che è elemento componente del flusso inter-strato fra l'(N)-strato e l'(N+1)-strato.

Entrando in maggiori dettagli, una (N)-PDU è una unità informativa che è specificata in un (N)-protocollo. Essa contiene in generale *informazioni di coordinamento* e *dati di utenza*; le prime sono a loro volta strutturate in (N)-PCI.

	INFORMAZIONI DI COORDINAMENTO	**INFORMAZIONI DI UTENZA**	**UNITA' INFORMATIVE**
(N)-(N) Entita' alla pari	**Informazione di protocollo (N)-PCI**	**Dati intra-strato**	**Unita' di dati dell'(N)-protocollo (N)-PDU**
(N+1)-(N) Entita' di strati adiacenti	**Informazione di interfaccia (N)-ICI**	**Dati inter-strato**	**Unita' di dati dell'(N)-interfaccia (N)-IDU**

Tabella II.4.1 - Relazioni tra tipi di informazione e unità informative

Una (N)-IDU è un unità informativa che è scambiata attraverso un (N)-SAP in una singola interazione tra una (N+1)-entità e una (N)-entità. E' costituita in generale dalla combinazione di informazioni di coordinamento e di dati di utenza; le prime sono a loro volta strutturate in (N)-ICI.

In entrambe le (N)-PDU e le (N)-IDU, i dati di utenza sono strutturati in *unita' di dati dell'(N)-servizio* [*(N)-SDU, Service Data Unit*]. Queste includono combinazioni di dati di utenza e di informazioni di coordinamento, che hanno origine nell'(N+1)-strato e sono trattate come dati di utenza dalle entità di questo. Un (N+1)-utente consegna queste unità all'(N)-fornitore perché siano trasferite all'(N+1)-utente remoto nell'ambito di quanto consentito dall'(N)-servizio.

Per chiarire il ruolo delle unità informative precedentemente definite analizziamo la composizione e il trattamento dei flussi inter-strato e intra-strato. Facciamo riferimento ad un insieme di sistemi interconnessi e indichiamo con *A* e *B* due sistemi adiacenti. Con riguardo ad esempio, al sistema *A* consideriamo dapprima un *flusso inter-strato* con *verso discendente* (cfr. Fig. II.3.5) e fissiamo l'attenzione sull'attraversamento dell'(N)-strato (Fig. II.4.1).

Quando una (N)-IDU viene consegnata all'(N)-strato si effettuano su di essa due operazioni:

- viene separata la (N)-ICI;
- viene enucleata la (N)-SDU.

La (N)-ICI svolge il suo ruolo di coordinamento dell'(N)-interfaccia e viene poi scartata. Alla (N)-SDU, nell'ipotesi che sia in corrispondenza uno a uno con una (N)-PDU, viene aggiunta una (N)-PCI, con la conseguente *apertura* di una istanza protocollare. Il risultato dell'aggiunta, che costituisce un *incapsulamento*, è una (N)-PDU. Quest'ultima, combinata con una (N-1)-ICI, dà luogo ad una (N-1)-

IDU. A questo punto si entra nel dominio dell'(N-1)-strato e si può ripetere quanto già detto per l'(N)-strato.

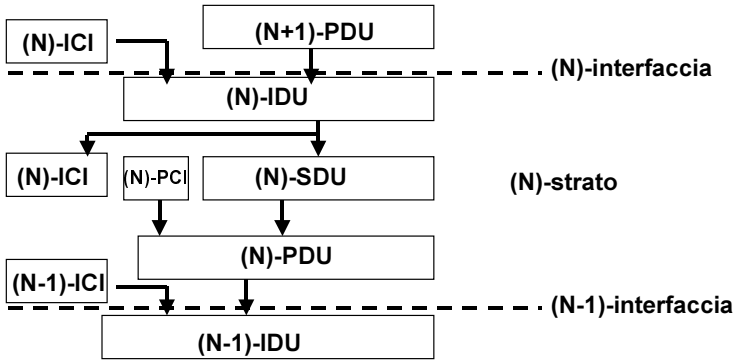


Figura II.4.1 – Unità informative nell'attraversamento discendente dell'(N)-strato

Se ora si considera, nel sistema *B*, un *flusso inter-strato* con *verso ascendente* e si ripete l'esame dell'attraversamento dell'(N)-strato, il punto di partenza del trattamento è rappresentato dalla consegna di una (N-1)-IDU all'(N)-strato (Fig. II.4.2). Su questa unità informativa si effettuano due operazioni:

- viene separata la (N-1)-ICI;
- viene enucleata la (N)-PDU.

La (N-1)-ICI coordina le azioni dell'(N-1)-interfaccia e viene poi scartata. Dalla (N)-PDU viene quindi separata la (N)-PCI. Il risultato della separazione, che costituisce un *decapsulamento*, è una (N)-SDU.

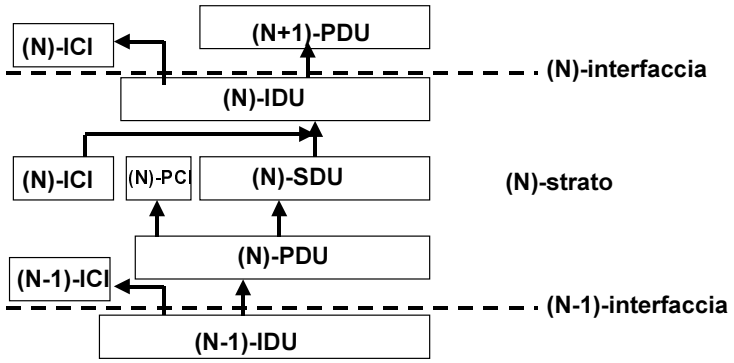


Figura II.4.2 – Unità informative nell'attraversamento ascendente dell'(N)-strato

La (N)-PCI *chiude* l'istanza protocollare che è stata aperta con la sua aggiunta nel sistema *A.* La (N)-SDU viene infine combinata con una (N)-ICI per dare luogo ad una (N)-IDU. Si perviene così all'(N)-interfaccia e quindi all'(N+1)-strato Per il trattamento in questo vale quanto già detto per l'(N)-strato.

Consideriamo ora il *flusso intra-strato*. Questo è costituito da una molteplicità di (N)-PDU, ognuna delle quali è ottenuta nel sistema emittente tramite aggiunta di una (N)-PCI ad una (N)-SDU e da ognuna delle quali nel sistema ricevente è estratta una (N)-SDU tramite sottrazione di una (N)-PCI. L'(N)-servizio trasferisce quindi le (N)-SDU come dati dell'(N)-utente nelle (N)-PDU. La lunghezza delle (N)-SDU non è però vincolata da quella delle (N)-PDU. A questo riguardo si distinguono tre casi:

- le (N)-SDU sono in *corrispondenza uno a uno* con le (N)-PDU, come si è supposto nelle Figg. II.4.1 e II.4.2;
- una singola (N)-SDU può essere *suddivisa* in segmenti più corti e trasferita in una molteplicità di (N)-PDU;
- più (N)-SDU possono essere *raggruppate* in segmenti più lunghi che sono trasferiti in una singola (N)-PDU.

Il caso di corrispondenza uno a uno è considerato in Fig. II.4.3, ove viene mostrata l'aggiunta di una (N)-PCI ad una (N)-SDU per formare, tramite incapsulamento, una (N)-PDU. Questa operazione è ulteriormente oggetto della Fig. II.4.4, ove si mostra l'incapsulamento in un sistema emittente e il decapsulamento in quello ricevente: per chiarezza grafica sono state omesse le unità informative di interfaccia.

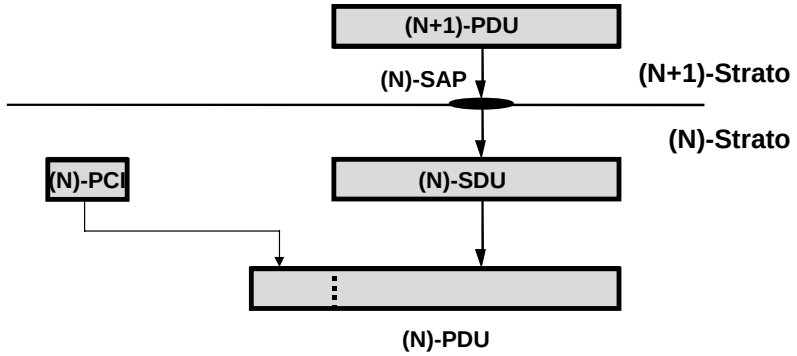


Figura II.4.3 – L'operazione di incapsulamento

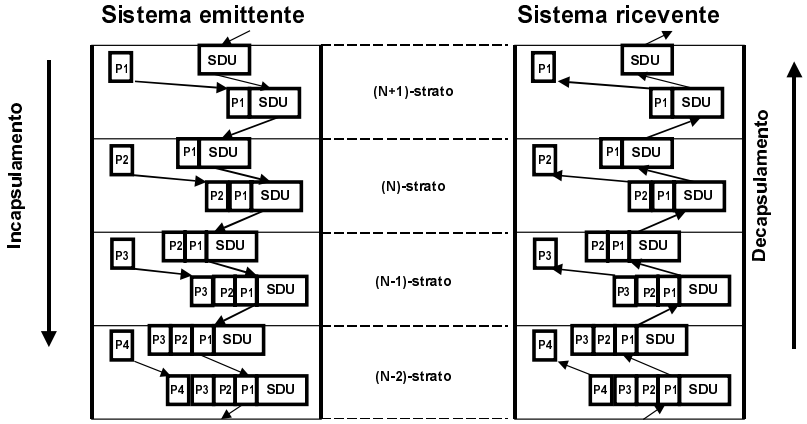


Figura II.4.4 – Incapsulamento in un sistema emittente e decapsulamento in un sistema ricevente

Gli altri due casi, che corrispondono a situazioni di incompatibilità tra la lunghezza delle (N)-SDU e quella delle (N)-PDU, sono trattati dallo strato pertinente con l'esecuzione di due funzioni.

- la *segmentazione*, se, in emissione, un'unica (N)-SDU deve essere suddivisa in diverse (N)-PDU; in ricezione, deve allora essere eseguita la funzione complementare, e cioè *l'assemblaggio* (Fig. II.4.5);
- la *aggregazione*, se il formato delle (N)-PDU e' maggiore di quello delle (N)-SDU; in questo caso e' necessario riunire varie (N)-SDU in un'unica (N)-PDU; la funzione complementare, svolta all'estremità ricevente e' quella di *disaggregazione* (Fig. II.4.6).

Nel caso di segmentazione, ogni (N)-PDU, in cui e' suddivisa una (N)-SDU, contiene la stessa (N)-PCI, mentre, nel caso di aggregazione, ogni (N)-PDU e' formata da due o più (N)-SDU, ognuna delle quali mantiene la propria (N)-PCI. Questi accorgimenti assicurano il mantenimento dell'identità di ogni (N)-SDU segmentata o aggregata.

Un'ulteriore funzione, detta di *concatenazione*, permette di unire in un'unica (N-1)-SDU due o più (N)-PDU; in questo caso la funzione complementare da svolgere in ricezione è quella di *separazione* (Fig. II.4.7). La funzione di concatenazione è simile a quella di aggregazione, con la differenza che, nella concatenazione, PDU più corte sono collocate in una SDU più lunga, mentre nell'aggregazione SDU più corte sono raggruppate in PDU più lunghe. In altre parole la concatenazione avviene tra due strati, mentre l'aggregazione è operazione svolta all'interno di uno strato.

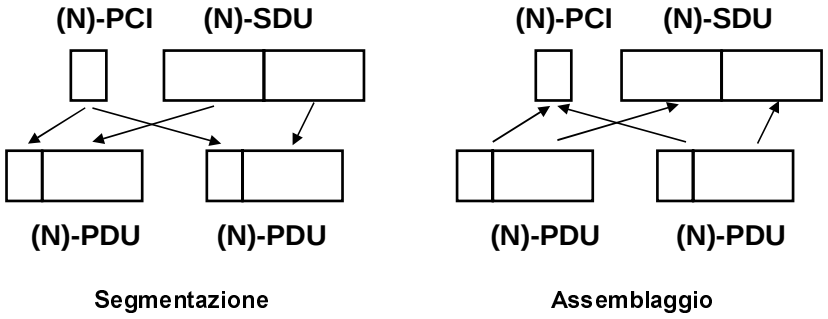


Figura II.4.5 - Segmentazione e assemblaggio

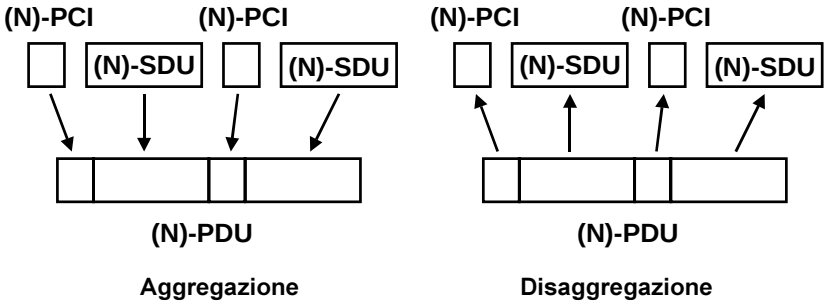


Figura II.4.6 – Aggregazione e disaggregazione

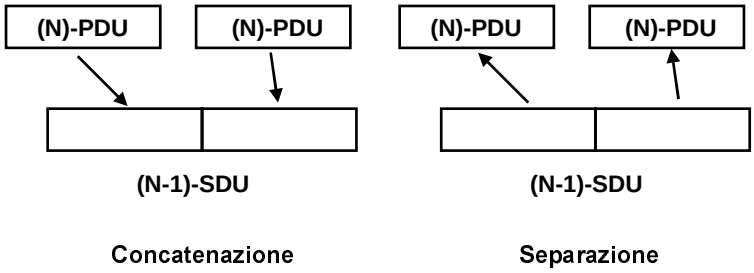


Figura II.4.7 – Concatenazione e separazione

II.5 Il servizio con connessione

Su un servizio nel modo *con connessione* si possono fornire alcuni ulteriori dettagli, riguardanti:

- la definizione e la identificazione delle connessioni di strato (§ II.5.1);
- le modalità di *gestione delle connessioni*, con riferimento alle funzioni di *instaurazione/abbattimento* e a quelle di *multiplazione* o di *suddivisione* (§ II.5.2);
- le funzioni di *trasferimento delle informazioni*, con riguardo ai dati *normali* o *veloci* e con ulteriore riferimento al *controllo di flusso* e alla *sequenzializzazione* (§ II.5.3);
- le modalità per fronteggiare condizioni di errore, che possono manifestarsi nel trasferimento delle informazioni su una connessione di strato (§ II.5.4).

II.5.1 Le connessioni di strato

Nel caso di (N)-servizio con connessione, tra le (N+1)-entità interessate a comunicare viene preventivamente stabilita una *associazione logica*. Questa è realizzata per mezzo dei servizi offerti dall'(N)-strato ed è detta *(N)-connessione*. Una (N)-connessione è quindi l'associazione dinamica stabilita tra due o più (N+1)-entità per controllare il trasferimento dell'informazione tra queste. Più in particolare, una (N)-connessione congiunge i due o più (N)-SAP a cui sono allacciate le (N+1)-entità. La Fig. II.5.1 mette in evidenza le relazioni tra entità e relative connessioni.

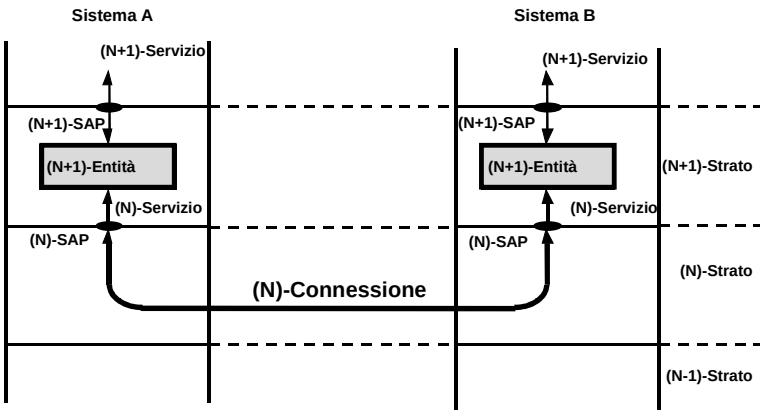


Figura II.5.1 - Connessioni di strato

In un servizio con connessione, durante la fase di instaurazione, ogni parte memorizza informazioni riguardanti le altri parti, come ad esempio gli indirizzi e le caratteristiche di QoS. Per effetto di questa memorizzazione e una volta che sia iniziata la fase di trasferimento, le informazioni di utenza possono essere accompagnate da informazioni di protocollo in misura più contenuta rispetto a quanto richiederebbe un indirizzamento completo delle parti.

Al riguardo occorre tenere conto che, all'atto in cui una (N)-connessione deve essere instaurata, ogni (N+1)-entità è identificata, per ciò che riguarda l'(N)-servizio, dall'indirizzo dell'(N)-SAP attraverso il quale l'(N+1)-entità interagisce con l'(N)-servizio. Quest'ultimo utilizza quindi gli (N)-indirizzi per instaurare la connessione richiesta. Invece, nella fase di trasferimento dell'informazione e in quella di abbattimento della connessione, non vengono utilizzati gli indirizzi degli (N)-SAP connessi. Si impiega, in loro luogo, un *identificatore di connessione*, che individua in modo univoco una connessione durante tutta la sua durata. Ciò consente di ridurre la quota di extra-informazione che sarebbe necessaria per trasferire la totalità degli (N)-indirizzi (quelli dell'origine e della destinazione).

Sulla modalità per identificare una connessione si aggiunge che:

- una (N)-connessione può essere instaurata tra due o più (N)-SAP; se sono interessati solo due (N)-SAP, la connessione è detta *punto-punto*; nel caso di un maggior numero di (N)-SAP interessati, è detta *punto-multipunto*;
- una (N+1)-entità può gestire simultaneamente una molteplicità di connessioni con altre (N+1)-entità.

In generale quindi su uno stesso (N)-SAP si attestano più (N)-connessioni. L'attestazione per ciascuna di queste è detta *punto terminale di una (N)-connessione* [*(N)-CEP, Connection End Point*]. All'interno di uno stesso (N)-SAP i punti terminali delle (N)-connessioni sono individuati dai cosiddetti *identificatori degli (N)-CEP*. Questi ultimi sono quindi distintivi per le singole connessioni che si attestano ad uno specifico (N)-SAP. Nella Fig. II.5.2 sono illustrate le definizioni ora fornite.

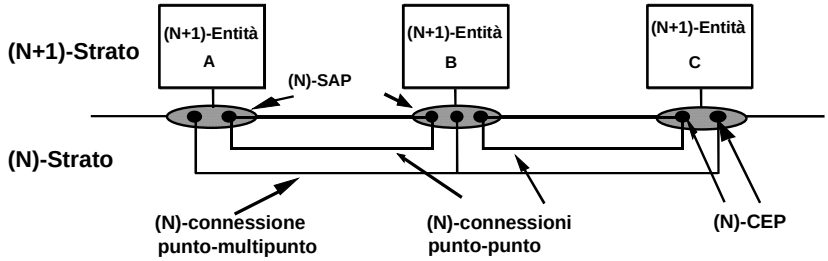


Figura II.5.2 - Connessioni di strato punto-punto e punto-multipunto

II.5.2 Gestione delle connessioni

Ci riferiamo dapprima alle funzioni di *instaurazione* e di *abbattimento* di una (N)-connessione nell'ipotesi che tutti gli strati dell'architettura operino nel modo con connessione.

Affinche' due (N+1)-entita' possano comunicare, una di esse deve richiedere all'(N)-strato l'*instaurazione* di una connessione attraverso gli (N)-SAP che le identificano. Le (N)-entita' raggiunte attraverso questi (N)-SAP provvedono a instaurare la connessione richiesta attraverso lo scambio (controllato da un opportuno protocollo) di (N)-PDU su una (N-1)-connessione che, a loro volta, richiedono all'(N-1)-strato. Conseguentemente la richiesta di connessione da parte delle (N+1)-entita' da' luogo verso il basso alla formazione di *N* connessioni attraverso la gerarchia degli strati fino a raggiungere quello di rango più basso, in cui le entità possono comunicare direttamente attraverso il mezzo trasmissivo.

Una (N)-connessione e' instaurata precisando, esplicitamente o implicitamente, un (N)-indirizzo per la (N+1)-entita' che ne fa richiesta e un (N)-indirizzo per ognuna delle (N+1)-entità che ne sono la destinazione. Una volta instaurata, una (N)-connessione ha due o più (N)-CEP. Ognuno di questi e' indiviso per ogni (N+1)-entita' e per ogni (N)-connessione. A esso viene quindi attribuito, come gia' detto, un *identificatore di riferimento* [identificatore di (N)-CEP] a uso esclusivo dell'(N+1)-entita' e dell'(N)-servizio.

L'instaurazione di una (N)-connessione presuppone che sia disponibile una (N-1)-connessione tra le (N)-entita' che agiscono di supporto e che queste ultime siano in uno stato nel quale possono svolgere lo scambio informativo protocollare per l'instaurazione della connessione. Se la prima condizione non e' verificata, deve essere preventivamente instaurata una (N-2)-connessione tra le entità alla pari dell'(N-1)-strato. Ciò richiede, per l'(N-1)-strato, il soddisfacimento delle stesse condizioni richieste per l'(N)-strato. Analoghe considerazioni si applicano per gli strati dell'architettura verso il basso fintantoché si trova una connessione già disponibile ovvero il mezzo trasmissivo.

L'*abbattimento* di una (N)-connessione e' normalmente richiesto al termine dello scambio informativo da una delle (N+1)-entità che l'hanno utilizzata. In presenza di condizioni anomale di funzionamento dell'(N)-strato (ad es. a causa di guasti o di errori procedurali), l'abbattimento di una (N)-connessione può essere iniziato anche da parte di una (N)-entita'. In questo caso saranno comunicati, alle (N+1)-entità impegnate nella comunicazione, l'avvenuto abbattimento della (N)-connessione e i motivi che hanno portato a tale provvedimento. Alle (N+1)-entita' spetta allora il compito di attivare le procedure necessarie per l'eventuale proseguimento della comunicazione.

Per quanto riguarda la corrispondenza tra connessioni appartenenti a strati adiacenti, possono essere individuati tre tipi di relazioni:

- una singola (N)-connessione corrisponde ad una singola (N-1)-connessione;
- una molteplicità di (N)-connessioni corrisponde ad una sola (N-1)-connessione (Fig. II.5.3);
- una singola (N)-connessione corrisponde ad una molteplicita' di (N-1)-connessioni (Fig.II.5.4).

La corrispondenza descritta nel secondo caso e' detta funzione di *multiplazione*, mentre nel terzo caso si parla di funzione di *suddivisione*; nella corrispondenza ipotizzata, ambedue queste funzioni sono svolte nell'(N)-strato.

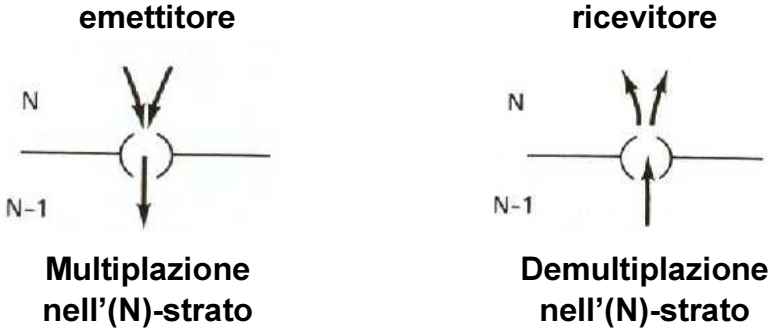


Figura II.5.3 – Multiplazione e demultiplazione

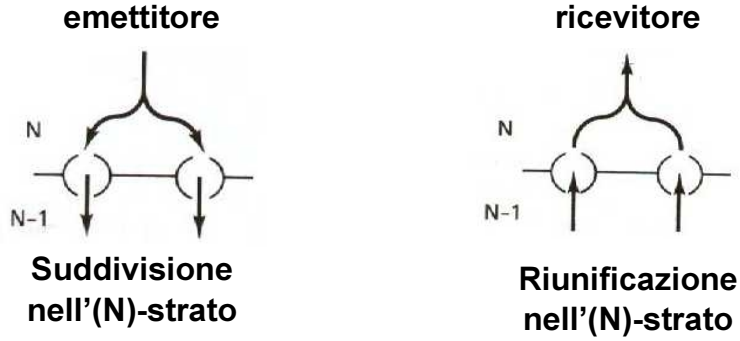


Figura II.5.4 – Suddivisione e riunificazione

L'esecuzione della funzione di multiplazione e' necessaria per rendere più economica ed efficiente l'utilizzazione di un (N-1)-servizio. La funzione di suddivisione e' invece impiegata per aumentare l'affidabilita' e le prestazioni di una (N)-connessione.

Le funzioni di multiplazione e di suddivisione comportano, all'interno dello strato in cui sono svolte, l'esecuzione di altre funzioni, che possono non essere necessarie per una corrispondenza uno ad

uno tra le connessioni di strato. In particolare, per quanto riguarda la multiplazione svolta nell'(N)-strato, occorre:

- *identificare l'(N)-connessione* per ogni (N)-PDU che viene trasferita sulla (N-1)-connessione; cio' al fine di assicurare che le informazioni di utenza, supportate da (N)-connessioni tra loro multiplate, non perdano la loro identita' e non possano quindi essere consegnate a una destinazione errata; l'identificazione della connessione e' effettuata mediante l'identificatore di (N)-CEP;
- *controllare il flusso* su ogni (N)-connessione (cfr. § II.5.3); ciò allo scopo di non compromettere le prestazioni della (N-1)-connessione con l'invio di unita' informative con ritmo superiore alla sua capacita' di trasferimento;
- *scadenzare* la prossima (N)-connessione da servire sulla (N-1)-connessione, quando piu' di una (N)-connessione e' pronta a inviare dati.

Nel caso invece della funzione di suddivisione, occorre provvedere a:

- *scadenzare* l'ordine di utilizzazione delle (N-1)-connessioni multiple, che sono utilizzate nella suddivisione di una singola (N)-connessione;
- effettuare, in ricezione, una eventuale *risequenziamento* delle (N)-PDU associate a una stessa (N)-connessione, dato che queste sono trasferite su (N-1)-connessioni diverse e quindi possono arrivare alla (N)-entita' di destinazione in una sequenza diversa da quella di emissione, anche se ogni (N-1)-connessione garantisce la sequenzialita' di consegna.

II.5.3 Trasferimento delle informazioni

Si distinguono trasferimenti di *dati normali* e di *dati veloci*. Nel caso di dati normali, si e' già visto che le informazioni di coordinamento e di utenza sono scambiate tra (N)-entita' sotto forma di (N)-PDU. Una (N)-SDU e' trasferita tra una (N+1)-entita' e una (N)-entita', attraverso un (N)-SAP, nella forma di una o più (N)-IDU; il suo trasferimento tra (N+1)-entita' alla pari avviene invece sotto forma di dati di utenza in una o più (N)-PDU. La (N)-SDU, in questo contesto, è la porzione dei dati inter-strato relativi all'(N)-interfaccia la cui identità è preservata da un estremo all'altro di una (N)-connessione. L'(N)-servizio garantisce cioè l'integrità di ogni (N)-SDU trasferita, nel senso che ne assicura la delimitazione e garantisce l'ordine dei dati al suo interno.

Una *unita' di dati veloce* (Expedited Data Unit) e' una SDU, che, richiedendo una particolare urgenza di consegna, e' trasferita e/o elaborata *con priorità* rispetto alle SDU normali; essa può essere utilizzata ad esempio per scopi di "interrupt". Concettualmente una connessione che supporta anche flussi di dati veloci può essere vista come composta di due vie, una per i dati normali, l'altra per i dati veloci; i dati inoltrati sulla via per i dati veloci hanno priorità sui dati normali. Conseguentemente l'(N)-strato assicura che una (N)-SDU veloce non sia consegnata a destinazione dopo una (N)-SDU veloce o normale, che sia inoltrata susseguentemente sulla stessa (N)-connessione.

Nel trasferimento di informazione possono essere richieste le funzioni di controllo di flusso, di sequenzializzazione e di controllo di errore, oltre a quelle di segmentazione, di aggregazione e di concatenazione, sulle quali si è già parlato nel par. II.4.

Il *controllo di flusso*, se effettuato, esercita una regolazione del ritmo di trasferimento dell'informazione e puo' operare, in alternativa, *su PDU* e *su IDU*. Nel *primo caso*, il controllo di flusso operato nell'(N)-strato regola il ritmo con il quale le (N)-PDU sono scambiate tra (N)- entità alla pari su una (N)-connessione. Nel *secondo caso*, tale funzione regola il ritmo di trasferimento con il quale le (N)-IDU sono scambiate tra una (N+1)-entita' e una (N)-entità attraverso un (N)-SAP. Si e' gia' visto che la multiplazione operata in uno strato puo' richiedere un controllo di flusso del primo tipo per ognuno dei flussi multiplati.

Consideriamo infine il caso in cui l'(N-1)-servizio non possa garantire la consegna delle (N-1)-SDU nello stesso ordine secondo cui queste sono state consegnate dall'(N)-strato. Se quest'ultimo richiede che venga preservato l'ordine delle (N-1)-SDU, l'(N)-strato deve prevedere una funzione di *sequenzializzazione*, e cioè di riordino delle (N-1)-SDU che sono state trasferite su una (N-1)-connessione.

II.5.4 Controllo di errore

Durante l'evoluzione della comunicazione all'interno di uno strato, possono verificarsi condizioni di errore. Queste possono essere causate da malfunzionamenti degli apparati, da errori procedurali o da altre cause accidentali, quali errori trasmissivi.

Per fronteggiare condizioni di questo tipo, un (N)-strato deve poter prevedere funzioni di riscontro, di rivelazione/notifica di errore e di reinizializzazione.

La funzione di *riscontro* puo' essere utilizzata da (N)-entita' alla pari, con l'impiego di un (N)-protocollo, per ottenere una piu' elevata probabilita' di rivelare la perdita di (N)-PDU di quanto non consenta l'(N-1)-servizio. A tale scopo ogni (N)-PDU trasferita tra (N)-entita' e' resa identificabile in modo univoco, in modo che il ricevitore possa informare l'emettitore circa la ricezione di ogni (N)-PDU. Una funzione di riscontro e' anche in grado di inferire la mancata ricezione di (N)-PDU e di attuare appropriate azioni di recupero.

Una (N)-entita' puo' essere in grado di rivelare eventuali condizioni di malfunzionamento (*errori*) dell'(N)-strato. Se cio' si verifica, la (N)-entita' deve notificare l'avvenuto errore alle entita' dello strato superiore. Questo insieme di funzioni, detto di *rivelazione e notifica di errore*, incrementa la qualita' del servizio offerto dallo strato inferiore, poiche' aumenta la probabilita' che un errore sia rivelato e quindi recuperato. Una volta che l'errore sia stato rivelato e notificato, spetta alle (N+1)-entita' adottare le opportune contromisure per recuperare le normali condizioni di funzionamento dello strato.

In alcuni casi, se il tipo di errore e' particolarmente grave, può essere necessario operare una *reinizializzazione* (reset) della (N)-connessione. Cio' si verifica, ad esempio, a seguito di una perdita di sincronizzazione tra le (N)-entita' in comunicazione. La funzione di reinizializzazione ha il compito di riportare la (N)-connessione a uno stato precedente, fino a cui si puo' assumere che lo scambio di informazioni sia avvenuto in modo esente da errori.

II.6 Modelli funzionali

Tra i modelli funzionali che sono stati finora definiti e che sono tuttora di attualità per applicazioni telematiche si menzionano i *modelli OSI* e *Internet*, ambedue basati sui principi delle architetture a strati e quindi trattabili con gli strumenti concettuali che sono stati introdotti nei precedenti paragrafi. Entrambi questi modelli sono stati concepiti per ambienti di *comunicazione di dati* e quindi con caratteristiche adeguate ad *applicazioni monomediali*. Non viene tuttavia esclusa, soprattutto per il modello Internet e con opportuni adattamenti, la possibilità di un allargamento ad applicazioni multimediali.

Il termine OSI, acronimo che fa riferimento alla "*interconnessione di sistemi aperti*" (Open System Interconnection), è utilizzato per definire la comunicazione tra sistemi di elaborazione che adottano un insieme di norme (le *norme OSI*), tali da consentire la cooperazione indipendentemente dalla natura dei sistemi e quindi con lo scopo di rendere possibile l'interconnessione e il colloquio tra *sistemi eterogenei* (cioè tra sistemi che hanno origine da costruttori diversi), oltre che tra sistemi omogenei. Il modello OSI è il risultato dell'attività di normalizzazione svolta dall'*ISO* e fatta propria dall'*ITU-T* (cfr. par. I.5). Gli obiettivi perseguiti in questa attività sono stati:

- disporre di norme la cui traduzione in prodotti fosse condizionata da vincoli tecnologici solo in modo debole;
- prevedere una estensibilità del modello e delle relative implicazioni normative verso evoluzioni future;
- comprendere una larga varietà di opzioni come compromesso tra esigenze spesso contrastanti, in quanto legate alla difesa di interessi industriali tra loro in competizione.

Il risultato è stato la produzione di norme, che sono ineccepibili dal punto di vista della generalità e della correttezza formale, ma che si sono dimostrate talvolta di difficile realizzazione pratica.

Il modello Internet ha invece avuto come origine l'organizzazione *ARPA* (*Advanced Research Project Agency*). Attualmente lo sviluppo, il coordinamento e la gestione del mondo Internet fanno capo all'*IAB* (cfr. par.I.5), il cui scopo principale è stato ed è tuttora l'adozione di soluzioni tecniche disponibili, che assolvano il loro compito e che consentano realizzazioni basate su posizioni consolidate. Ciò ha consentito di ottenere velocemente prodotti in linea con la normativa Internet e di adottare in modo tempestivo gli aggiornamenti necessari per rispondere a nuove esigenze.

La differenza tra i due modelli è dovuta sostanzialmente allo scopo diverso che le organizzazioni preposte si sono prefisso. Da un lato, con la modellistica OSI, gli organismi di normalizzazione formali (ISO e ITU-T) hanno puntato a definire un quadro normativo che fosse senza limitazioni circa la sua diffusione virtuale e che comprendesse ogni possibile accorgimento anche in vista degli aggiornamenti prevedibili. Dall'altro lato l'IAB ha cercato di rispondere in modo più rapido alle esigenze, anche a breve termine, dell'utente finale, fornendo quindi risposte concrete alle domande esistenti.

È il caso di sottolineare che, in un ambiente per comunicazioni di dati, le architetture del modello OSI e di quello Internet non sono le uniche che possono essere derivate dai criteri esposti nei precedenti paragrafi. A criteri analoghi sono infatti ispirate altre architetture sviluppate dai singoli costruttori (*architetture proprietarie*), quali per esempio la *System Network Architecture* (SNA) della IBM e la *Digital Network Architecture* (DNA) della DEC.

Come già detto, i modelli OSI e Internet sono stati concepiti per applicazioni monomediali, quali quelle orientate a scambi di dati. Per descrivere ambienti di comunicazione aperti ad *applicazioni multimediali* si è riconosciuta la necessità di modelli di riferimento che, pur adottando i principi-base delle architetture a strati e la terminologia ad esse associata, ne costituiscono una opportuna generalizzazione. Per sinteticità di riferimento chiameremo questi modelli con l'acronimo *PRM* (Protocol Reference Model) e indicheremo con *Multimed* l'ambiente che un PRM rappresenta.

II.6.1 Il modello OSI

L'identificazione degli strati che compongono l'architettura OSI è stata effettuata con l'applicazione dei criteri del raggruppamento e della gerarchizzazione (cfr. § II.2.1). Ciò ha portato alla definizione di sette strati funzionali: lo strato di *applicazione* (strato 7); lo strato di *presentazione* (strato 6); lo strato di *sessione* (strato 5); lo strato di *trasporto* (strato 4); lo strato di *rete* (strato 3); lo strato di *collegamento* (strato 2); lo strato *fisico* (strato 1). L'insieme degli strati ora citati e la loro struttura gerarchica sono mostrati in Fig. II.6.1.

Con l'eccezione dello strato fisico, che si interfaccia direttamente con il mezzo trasmissivo, i servizi di strato nel modello OSI sono stati definiti nelle versioni sia *con connessione* che *senza connessione*. Inoltre per gli strati di trasporto e di rete è previsto che un modo di servizio possa essere supportato da un differente modo nello strato contiguo inferiore (Fig. II.6.2): ad esempio un servizio senza connessione nello strato di trasporto può essere supportato da un servizio con connessione nello strato di rete. Qui di seguito vengono esposte in modo sintetico le finalità dei sette strati.

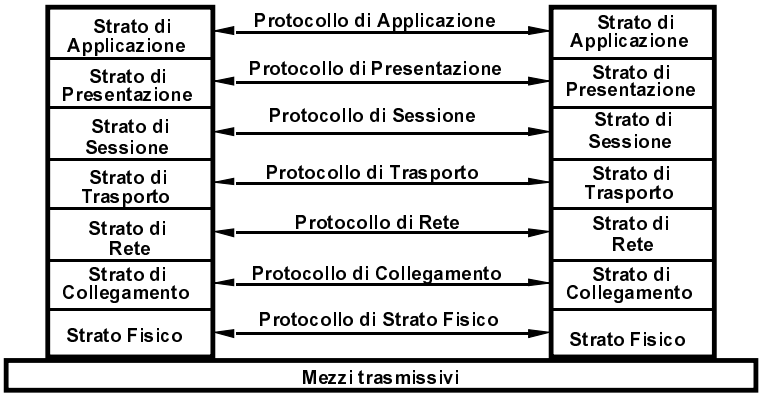


Fig. II.6.1- Architettura del modello OSI.

Lo *strato di applicazione* ha lo scopo di fornire ai processi applicativi residenti in un sistema i mezzi per accedere all'ambiente OSI, facendo sì che quest'ultimo costituisca una macchina virtuale in grado di associare un processo applicativo residente in un sistema con qualsiasi altro processo applicativo residente in un qualsiasi altro sistema remoto. Poiché gli utenti dei servizi offerti dallo strato di applicazione possono essere processi applicativi di qualsiasi natura, risulta arduo individuare un insieme di funzioni sufficientemente completo e generale che si adatti a tutte le possibili applicazioni. Per la descrizione delle funzioni interne allo strato di applicazione si segue quindi la strada di individuare, dapprima, elementi di servizio comuni a tutte le applicazioni e, in seguito, moduli disgiunti comprendenti elementi di servizio orientati a determinate classi di applicazioni.

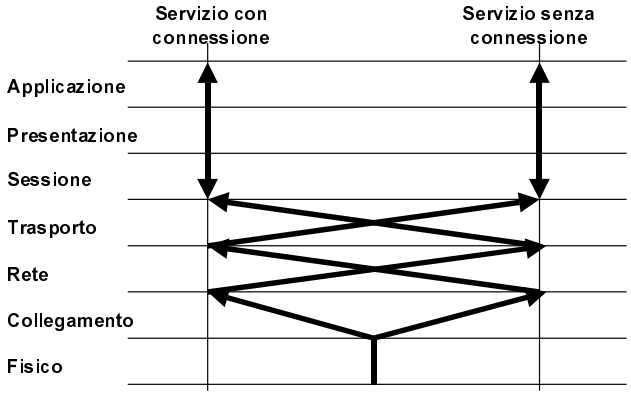


Figura II.6.2 - Supporto degli strati OSI per servizi con e senza connessione

Lo *strato di presentazione* ha due compiti fondamentali:

- rendere direttamente disponibili alle entità di applicazione i servizi dello strato di sessione relativi alla strutturazione, alla sincronizzazione e alla gestione del loro dialogo;
- risolvere i problemi di compatibilità tra due qualsiasi entità di applicazione per quanto riguarda la *rappresentazione dei dati* da trasferire e il modo di riferirsi a strutture di dati del sistema remoto.

In altre parole lo strato di presentazione ha lo scopo di sollevare le entità di applicazione da qualsiasi compito relativo alla *trasformazione della sintassi dei dati*. In generale, all'interno dello strato di presentazione esistono tre tipi di sintassi: due *sintassi locali*, utilizzate dalle singole entità di applicazione, e una *sintassi di trasferimento* utilizzata per lo scambio informativo vero e proprio. La

sintassi di trasferimento è negoziata dalle entità di applicazione, mentre la conversione tra sintassi locali e sintassi di trasferimento è effettuata dallo strato di presentazione.

Lo scopo dello *strato di sessione* è quello di assicurare alle entità di presentazione una risorsa logica per organizzarne il colloquio. Ciò ha lo scopo di strutturare e di sincronizzare lo scambio dati in modo da poterlo sospendere, riprendere e terminare ordinatamente. Lo strato di sessione presuppone che siano risolti tutti i problemi relativi al trasporto dell'informazione tra i sistemi. Il suo compito consiste invece nel mascherare, per quanto possibile, eventuali interruzioni del servizio di trasporto e nel mantenere una continuità logica nell'evoluzione del colloquio tra entità di presentazione.

Lo *strato di trasporto* fornisce alle entità di sessione una risorsa virtuale per il trasferimento trasparente delle unità di dati. Deve quindi, da un lato, colmare eventuali deficienze e fluttuazioni della qualità di servizio offerto dallo strato di rete e dall'altro, ottimizzare l'uso di questo servizio di rete in modo da poter garantire, al costo minimo, le prestazioni richieste. Procedendo dal basso verso l'alto nell'esame dell'architettura OSI, lo strato di trasporto è il primo che abbia esclusivamente un *significato da estremo a estremo*. In altre parole, mentre le funzionalità dei primi tre strati dell'architettura possono essere presenti sia nei sistemi terminali che nei nodi della rete di comunicazione, le procedure relative allo strato di trasporto sono eseguite solo dai sistemi terminali. Infatti, tutti i problemi relativi all'accesso dei sistemi terminali alla rete di comunicazione, all'instradamento e all'utilizzazione di eventuali reti di transito si considerano completamente risolti dallo strato di rete.

Lo *strato di rete* rende invisibile allo strato di trasporto il modo in cui sono utilizzate le risorse di rete, in particolare se sono utilizzate diverse sottoreti in cascata. L'unico elemento che è visibile all'interfaccia con lo strato di trasporto è la qualità di servizio complessiva offerta dallo strato di rete. Se il trasferimento interessa due sottoreti che offrono differenti qualità di servizio, la cooperazione tra le sottoreti può avvenire secondo due diverse modalità:

- le reti cooperano senza l'esecuzione di nessuna funzione addizionale; in questo caso la qualità di servizio risultante sarà sicuramente non superiore a quella corrispondente alla rete con minore qualità di servizio;
- sulla rete con qualità di servizio inferiore sono eseguite funzioni che hanno l'obiettivo di innalzare tale qualità di servizio allo stesso livello della seconda rete; in questo caso la qualità complessiva può essere uguale a quella della sottorete a più alta qualità.

La scelta tra queste due alternative dipende dalla differenza di prestazioni delle due sottoreti e da fattori di costo.

La funzione fondamentale eseguita dallo *strato di collegamento* è *rivelare e recuperare gli errori trasmissivi* verificatisi durante il trasferimento fisico. Tale funzione dovrebbe garantire allo strato di rete la fornitura di un servizio di rete dell'informazione esente da errori.

Lo *strato fisico* ha il compito principale di effettuare il trasferimento fisico delle cifre binarie scambiate tra le entità di collegamento. Le funzioni eseguite al suo interno e i servizi forniti allo strato superiore hanno lo scopo di assicurare l'indipendenza della comunicazione dalle caratteristiche del particolare mezzo trasmissivo utilizzato.

II.6.2 Il modello Internet

L'*architettura* del modello Internet, partendo dal livello gerarchicamente più basso, comprende quattro strati.

Un primo strato, detto di *accesso alla rete* (Network Access Layer), consente la utilizzazione di risorse infrastrutturali (sotto-reti), tra loro anche non omogenee per tecnica realizzativa. Lo strato di accesso alla rete include le funzioni che nel modello OSI sono comprese negli strati fisico, di collegamento e di rete, quest'ultimo almeno per ciò che riguarda gli aspetti connessi al funzionamento di ogni singola rete componente (*sottostrato di rete basso*).

Un secondo strato, che è denominato *IP* (*Internet Protocol*) dal nome del protocollo che gli è proprio, consente l'inter-funzionamento delle varie reti componenti con funzionalità che nel modello OSI sono collocabili in un *sottostrato di rete alto*. Il servizio di strato corrispondente è senza connessione.

Un terzo strato corrisponde allo strato di trasporto e a parte di quello di sessione nel modello OSI: viene denominato in base al protocollo che gli è proprio. Un primo tipo di protocollo di questo strato è il *TCP* (*Transmission Control Protocol*), nell'ambito del quale il servizio di strato è con connessione. Un protocollo alternativo è l'*UDP* (*User Datagram Protocol*), che a differenza di TCP è senza connessione;

Il quarto e ultimo strato del modello Internet è quello applicativo e corrisponde ai tre ultimi strati del modello OSI. Viene denominato strato dei *Servizi Applicativi* (Application Services).

In Fig. II.6.3 è effettuato un confronto fra gli strati dei due modelli OSI e Internet.

<table>
<tr><th>Modello OSI</th><th>Modello INTERNET</th></tr>
<tr><td>Applicazione</td><td rowspan="3">Applicativo</td></tr>
<tr><td>Presentazione</td></tr>
<tr><td>Sessione</td></tr>
<tr><td>Trasporto</td><td>Da Estremo
a Estremo</td></tr>
<tr><td>Rete</td><td>Internet</td></tr>
<tr><td>Collegamento</td><td rowspan="2">Accesso
di rete</td></tr>
<tr><td>Fisico</td></tr>
</table>

Figura II.6.3 – Confronto tra i modelli OSI e Internet

II.6.3 Il PRM per ambiente Multimed

La generalizzazione richiesta nella definizione di un PRM è legata all’esigenza di distinguere, in un ambiente di comunicazione multimediale, i tre tipi di flussi informativi già definiti in § I.3.1 e cioè quelli relativi alle *informazioni di utente*, di *controllo* e di *gestione*.

L'informazione di utente (*U-informazione*) può essere trasferita tra due o più utenti o tra utenti e centri di servizio e, nel trasferimento, puo' essere trattata dalla rete in modo trasparente oppure può essere elaborata, come accade nei processi di archiviazione e in quelli di codifica crittografica svolti all'interno della rete stessa.

L'informazione di controllo o di segnalazione (*C-informazione*) è necessaria affinchè possa avvenire il trasferimento della U-informazione. Essa e' quindi costituita dall'informazione, che, ad esempio, in una rete a circuito fissa è preposta a: (i) controllare una connessione di rete, per instaurarla o per abbatterla; (ii) controllare l'uso di una connessione di rete gia' instaurata, per modificarne, ad esempio, le caratteristiche durante una chiamata; (iii) controllare la fornitura di servizi supplementari. Nel caso di rete mobile, sempre ad esempio, è dedicata a: (1) controllare le risorse radio; (2) controllare la sessione; (3) gestire la mobilità degli utenti.

L'informazione di gestione (M-informazione) ha sostanzialmente due finalità principali. Da un lato, riguarda tutti gli aspetti di *gestione locale* relativi al trasferimento dei precedenti due tipi di informazione e al coordinamento tra le funzioni che vengono attivate a questo scopo. Dall'altra, ha significativita' in un contesto di *gestione globale*, che riguarda la gestione delle risorse all'interno della rete e lo scambio di flussi informativi preposti a compiti di esercizio e di manutenzione.

Lo scambio di ognuno di questi tre tipi di informazione attraverso la rete o nell'ambito di ogni elemento funzionale di questa e' descritto in un apposito piano. Si parla allora di *piano d'utente* (piano U), di *piano di controllo* (piano C) e di *piano di gestione* (piano M).

Mentre i primi due hanno una struttura a strati, il terzo e' a sua volta suddiviso in due parti corrispondenti alle due finalita' della M-informazione a cui si e' accennato in precedenza. La prima di queste parti, che comprende le *funzioni di gestione di piano*, riguarda il coordinamento locale tra tutti i piani e non e' stratificata. La seconda parte, che include le *funzioni di gestione di strato*, ha rilevanza attraverso tutti gli elementi funzionali della rete ed e' di tipo stratificato.

L'applicazione dei concetti di strato e di piano ha condotto alla definizione dell'elemento di base di un PRM, e cioè del *blocco di protocolli* (PB - Protocol Block). Un PB, che puo' essere considerato una generalizzazione del concetto di sistema introdotto in § II.2.5, puo' descrivere vari gruppi funzionali nelle sezioni di accesso e interna di una rete.

In Fig. II.6.4 viene mostrata la struttura semplificata di un PB, che adotta una stratificazione del tipo OSI. Si distinguono i tre piani, U, C e M e ognuno dei primi due e' potenzialmente strutturato in 7 strati. Nel piano M non vengono distinte le funzioni di gestione di piano e di strato; si tiene conto, invece, della suddivisione del piano C in due parti, chiamate *piano di controllo locale* e *piano di controllo globale*. Il primo di questi riguarda scambi di informazione di controllo tra entità adiacenti, mentre il secondo e' coinvolto quando tali scambi avvengono tra entità remote.

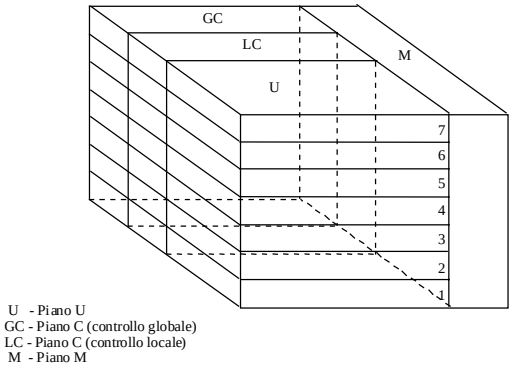


Fig. II.6.4 - Struttura semplificata di un PB.

All'interno di un PB sono possibili varie interazioni tra i raggruppamenti funzionali residenti in ognuno dei tre piani. Tra queste si possono citare le *interazioni tra i piani U e C*, che sono necessarie per sincronizzare le attivita' di questi piani. I piani U e C non interagiscono direttamente e possono comunicare solo con le entita' del piano M, utilizzando apposite *primitive di gestione*. Tali entita', che svolgono funzioni di gestione di piano, provvedono cosi' a coordinare le attivita' nei piani U e C. Nell'ambito poi di ognuno di questi, le *interazioni tra strati adiacenti* avvengono con primitive di servizio, secondo le usuali modalita' previste nel modello di Fig. II.2.6.

Se poi si considerano, come in Fig. II.6.5, le interazioni di un PB con l'esterno, si possono distinguere: 1) le *interazioni di un PB con i mezzi fisici*, e cioe' quelle che debbono assicurare il trasferimento fisico delle informazioni pertinenti a ogni piano tra due PB; 2) le *interazioni in corrispondenza della faccia superiore di un PB*, e cioe' quelle che riguardano la comunicazione tra ogni piano e vari processi applicativi; 3) le *interazioni tra entita' alla pari,* e cioe' tra entita' di un dato piano appartenenti allo stesso strato e residenti in PB diversi.

Circa le interazioni con i mezzi trasmissivi, il trasferimento fisico dei flussi informativi di piano puo' avere luogo, in alternativa, su mezzi fisici separati connessi ai singoli piani ovvero su un unico

mezzo trasmissivo condiviso. Questo secondo caso si verifica, ad esempio, per i piani U e C quando si impiega una segnalazione in banda ovvero quando la U-informazione e' trasferita sul canale D.

Relativamente poi alle interazioni con i processi applicativi, queste comprendono il trasferimento di informazione da/verso le applicazioni d'utente, quelle di controllo e quelle di gestione di sistema.

Infine le interazioni tra entità alla pari di PB diversi si svolgono secondo gli usuali principi dell'architettura stratificata e non vengono considerati in Fig. II.6.5 per motivi di semplicità grafica.

Per l'utilizzazione dei PB come elementi componenti di un PRM per rappresentare sistemi terminali o intermedi, e' opportuno distinguere questi due casi. Per rappresentare un sistema terminale e' sufficiente un singolo PB, mentre la rappresentazione di un sistema intermedio richiede l'impiego di due PB, di cui ognuno e' la versione speculare dell'altro.

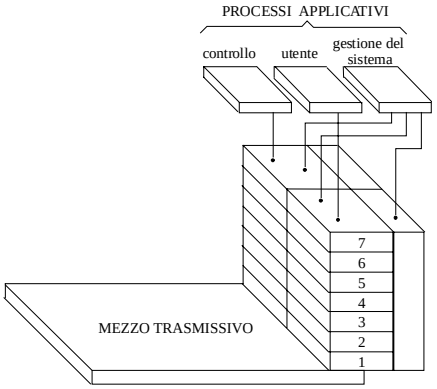


Fig. II.6.5 - Interazioni di un PB con l'esterno.

II.7 Modelli di riferimento

Scopo di questo paragrafo è introdurre modelli per la descrizione di un processo di comunicazione. Supponiamo che questi modelli applichino i principi delle architetture stratificate. Ipotizziamo inoltre che nei modelli considerati siano inclusi:

- due *sistemi terminali*, in corrispondenza modellistica con una coppia di *apparecchi terminali*;
- tre *sistemi intermedi*, in corrispondenza modellistica con due *nodi di* e un *nodo di transito* (cfr. § I.1.2).

Scopo di questi modelli è rappresentare le sezioni tipiche di una rete di telecomunicazione, e cioè le reti di accesso e di trasporto. Le due parti della rete di accesso consentono ad ogni apparecchio terminale di interagire con il pertinente nodo di accesso attraverso l'interfaccia utente-rete. La rete di trasporto interconnette i nodi di accesso attraverso il nodo di transito e le relative interfacce nodo-nodo.

Gli apparecchi terminali, in quanto modellabili tramite sistemi terminali, possono essere visti come organizzati in tanti strati funzionali quanti ne sono previsti nell'architettura di comunicazione considerata. I nodi di accesso e di transito, in quanto modellabili tramite sistemi intermedi, comprendono un sottoinsieme di detti strati; tale sottoinsieme è quello degli *strati di trasferimento* e include gli strati dell'architettura che, partendo dallo strato di rango più basso, arrivano fino allo strato che include una *funzione di rilegamento*. I rimanenti strati dell'architettura sono invece gli *strati di utilizzazione*.

L'architettura protocollare di un processo di comunicazione prevede sempre la definizione di *protocolli di accesso*, *di transito* e *di utilizzazione*. Nelle Figg. II.7.1 e II.7.2 sono presentati due esempi di queste pile protocollari.

I *protocolli di accesso* regolano le interazioni tra un apparecchio terminale e la relativa terminazione di rete; riguardano gli strati di trasferimento fino a quello di ordine più elevato, che è presente nell'interfaccia utente-rete. I *protocolli di transito* definiscono le regole di interazione tra un nodo di accesso e un nodo di transito, ovvero tra due nodi di transito, ovvero direttamente fra i due nodi di accesso; riguardano gli strati di trasferimento fino a quello di ordine più elevato, che è presente nelle interfacce nodo-nodo. I *protocolli di utilizzazione* riguardano gli strati omonimi; vengono gestiti da estremo a estremo, nel senso che le entità alla pari interagenti per il loro tramite risiedono negli apparecchi terminali.

Riferiamoci dapprima al modello di riferimento in Fig. II.7.1. È considerato il caso in cui gli strati di trasferimento includono gli stessi sotto-sistemi nei nodi di accesso e in quelli di transito. I nodi gestiscono i protocolli che sono pertinenti ad ognuna delle loro interfacce di interazione. La gestione di questi protocolli è effettuata *sezione per sezione*. Inoltre, sempre per ipotesi, un nodo di accesso opera una *conversione* tra i protocolli di accesso e quelli di transito, mentre un nodo di transito effettua una *rigenerazione* dei protocolli di transito. In ambedue i casi vengono attuate: una chiusura dei protocolli relativi all'interfaccia a monte e una apertura di quelli relativi all'interfaccia a valle.

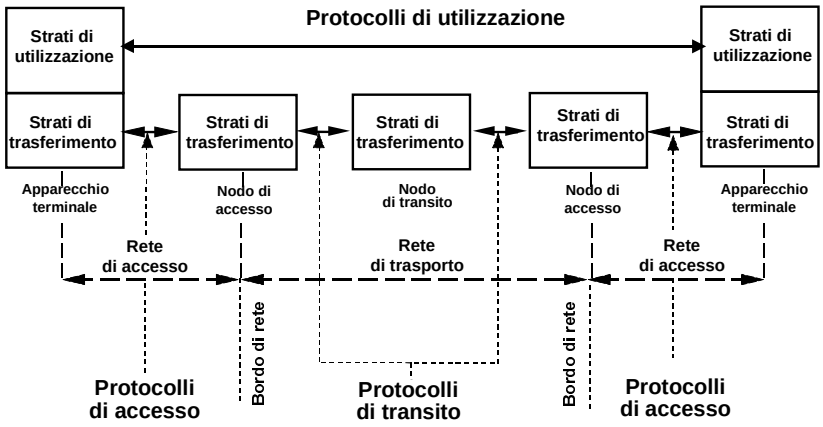


Figura II.7.1 – Primo modello di riferimento per un processo di comunicazione

Il secondo modello di riferimento (Fig. II.7.2) contempla il caso in cui uno o più tra gli strati di trasferimento di ordine gerarchico più elevato includono sotto-sistemi che sono presenti nei soli nodi di accesso e non in quelli di transito. In questo caso, i protocolli di transito sono di due tipi: quelli relativi agli strati di trasferimento di rango più basso, che sono presenti nei nodi di accesso e di transito, e quelli relativi agli strati di trasferimento di rango più alto, che sono presenti nei soli nodi di accesso. I primi rientrano nel caso già considerato dal primo modello di riferimento e sono quindi gestiti *sezione per sezione*. I secondi riguardano invece l'interazione diretta tra entità alla pari residenti nei *nodi di accesso* e sono quindi gestiti *da bordo a bordo.*

II.8 Architetture di interconnessione

Vengono qui esaminati i problemi connessi alla realizzazione di infrastrutture costituite dalla interconnessione di reti o di segmenti di rete tra loro *omogenei* o *non omogenei* . Nel caso in cui l'interconnessione riguardi reti tra loro non omogenee e quindi caratterizzate da architetture protocollari tra loro diverse, all'insieme degli elementi utilizzati nella interconnessione si attribuisce il termine *inter-rete*. Le reti individuali, omogenee al loro interno e componenti una inter-rete sono

denominate *sotto-reti*; a queste sono connessi gli apparecchi terminali (ad es. costituiti da elaboratori). Realizzare una inter-rete significa consentire scambi di informazione tra apparecchi terminali che sono connessi a sotto-reti *diverse*.

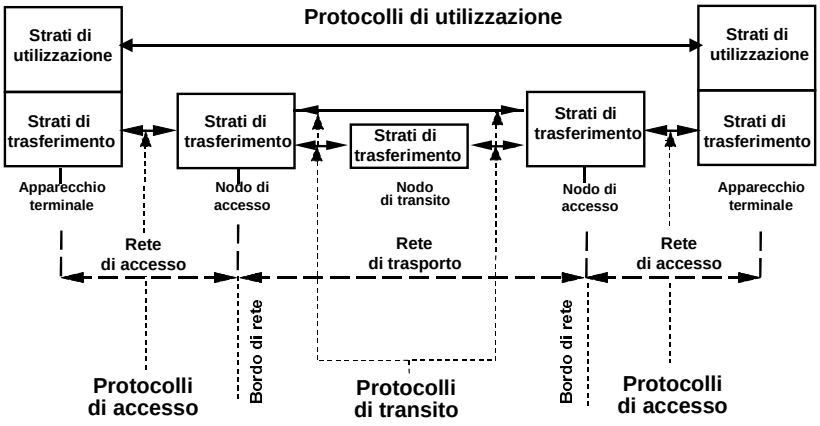


Figura II.7.2 - Secondo modello di riferimento per un processo di comunicazione

II.8.1 Funzioni in una inter-rete

Il caso delle inter-reti rientra nelle situazioni più generali in cui sia necessario interconnettere *segmenti di rete* per estenderne l'area di copertura. È quanto si verifica, ad esempio, nell'interconnessione di due o più segmenti di LAN (cfr. § I.3.2) che possono essere omogenei (ad es. segmenti di LAN Ethernet) ovvero disomogenei (ad es. due LAN di differente architettura protocollare). Altro esempio significativo è l'interconnessione di una LAN con una WAN. In ogni caso lo scambio di informazioni tra apparecchi terminali che sono connessi a segmenti di rete diversi è possibile se si soddisfano alcuni requisiti.

In primo luogo i segmenti di rete debbono essere *fisicamente connessi*; questa condizione non è però sufficiente. Occorre infatti interporre tra i segmenti di rete un *sistema di interconnessione*, che svolga una *funzione di rilegamento*. Questo sistema, nel seguito indicato con l'acronimo SINT, deve in generale svolgere funzioni di natura sia fisica che logica. Tra le funzioni fisiche vanno citati i trattamenti del segnale che supporta l'informazione (amplificazione, rigenerazione, ecc.); tra le funzioni logiche rientra invece l'instradamento da un segmento di rete ad un altro nel cammino che l'informazione deve compiere da un apparecchio terminale ad un altro.

Un SINT è un elaboratore dotato di un opportuno software e di due o più *interfacce fisiche* a seconda del numero di segmenti di rete che esso interconnette. Ogni interfaccia fisica deve essere compatibile con la sotto-rete verso cui si affaccia. La Fig.II.8.1 mostra un esempio in cui un SINT interconnette due o più sotto-reti. L'informazione avente origine in un apparecchio terminale connesso alla sotto-rete 1 e avente destinazione in un apparecchio terminale connesso alla sotto-rete 2 deve essere ricevuta da SINT sulla sua interfaccia con la sotto-rete 1 ed emessa sempre da SINT sulla sua interfaccia con la sotto-rete 2.

Le funzioni svolte da un SINT dipendono in generale dalle architetture protocollari degli apparecchi terminali da mettere in corrispondenza. A questo proposito supponiamo di voler interconnettere due apparecchi terminali (*host*) H1 e H2 che presentano architetture protocollari tra loro

diverse fino ad un certo strato di rango *N*, ma tra loro uguali per ciò che riguarda gli strati di rango superiore a *N*. In queste condizioni il dialogo tra H1 e H2 richiede che SINT contenga tutti i sottosistemi architetturali di rango minore o uguale a *N*. SINT deve realizzare una *conversione dei protocolli* di una sotto-rete nei corrispondenti protocolli di un'altra sotto-rete.

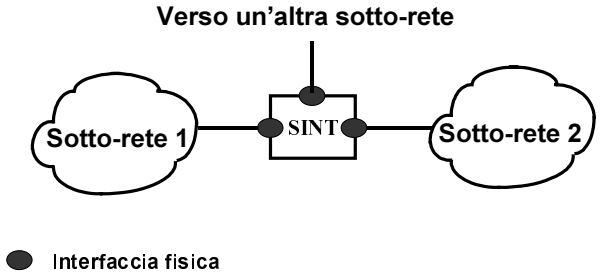


Figura II.8.1 – Interconnessione di sotto-reti con un SINT

Una possibile classificazione dei SINT è basata sul livello protocollare più alto a cui essi operano la conversione dei protocolli. La relativa nomenclatura (di uso corrente, ma con qualche variante nei prodotti commerciali) fa riferimento agli strati del modello OSI ed è la seguente:

- un *repeater* (ripetitore) agisce a livello di strato 1;
- un *bridge* (ponte) opera fino a un livello di strato 2;
- un *router* (instradatore) lavora fino a un livello di strato 3;
- un *gateway* (passerella) interviene su tutti gli strati dell'architettura.

Ritornando all'interconnessione in una inter-rete, le sotto-reti componenti possono presentare protocolli diversi *almeno fino allo strato di rango 3*, ove sono normalmente svolte le funzioni di indirizzamento e di instradamento. Infatti in una inter-rete deve essere possibile:

- indirizzare tutti gli apparecchi terminali che sono connessi alle sotto-reti componenti;
- instradare corrispondentemente l'informazione da trasferire;

ne segue che, secondo la nomenclatura ora introdotta, i SINT di una inter-rete sono *router* o *gateway*.

II.8.2 Architettura di una inter-rete

Per entrare nel merito delle funzioni di natura logica svolte da un SINT operante in una inter-rete, occorre prenderne in esame l'architettura protocollare. Consideriamo due apparecchi terminali, H1 e H2, rispettivamente connessi alle sotto-reti 1 e 2 e posti in corrispondenza per mezzo di un SINT (cfr. Fig.II.8.1). Supponiamo che le architetture dei due apparecchi terminali siano coerenti con quelle delle sotto-reti a cui sono connessi e che gli strati di trasferimento siano in numero di 3 come nel modello OSI; in particolare la suddivisione in sottosistemi dei modelli architetturali di H1 e di H2 sia tale che:

- i sottosistemi degli strati di rango *non inferiore a 4* trattano le stesse pile protocollari in H1 e in H2;
- i sottosistemi degli strati di rango *inferiore a 4* trattano pile protocollari diverse.

Una prima pila *{Px.1}, (x =1,2,3)*, corrisponde all'architettura della sotto-rete 1 e comprende i protocolli definiti per gli strati *Sx.1, (x =1,2,3)*. La seconda *{Px.2},(x =1,2,3)*, è invece in relazione all'architettura della sotto-rete 2 e include i protocolli definiti nell'ambito degli strati *Sx.2, (x =1,2,3)*.

I protocolli degli strati di rango non inferiore a 4 possono quindi essere gestiti *direttamente* (cioè da estremo a estremo) da H1 e H2. Nel caso invece degli strati di rango inferiore a 4, la gestione dei relativi protocolli deve essere effettuata sezione per sezione, e cioè tra H1 e SINT per la pila *{Px.1}* e tra H2 e SINT per la pila *{Px.2}*.

Denominiamo le due interfacce di SINT: con *interfaccia 1* quella verso la sotto-rete 1 e con *interfaccia 2* quella rivolta invece alla sotto-rete 2. Entrambe queste interfacce includono solo sottosistemi appartenenti agli strati di rango 1, 2 e 3. Le due interfacce differiscono però per la diversità delle architetture entro cui debbono operare. Nell'interfaccia 1 viene gestita la pila *{Px.1},* mentre l'interfaccia 2 tratta la pila *{Px.2}.* Chiamiamo poi *mezzo trasmissivo 1* o *2* il segmento della sotto-rete 1 o 2 che assicura il legame fisico nell'ambito di ciascuna delle coppie H1-SINT e SINT-H2.

Compito di SINT è convertire i protocolli della pila *{Px.1}* in quelli della pila *{Px.2}*. Le modalità di questa conversione sono ora illustrate con riferimento al modello architetturale di Fig.II.8.2. Si considerano le modalità di trattamento di un segmento informativo *IS* avente origine da un processo applicativo che risiede in H1 e avente destinazione in un processo applicativo che risiede in H2. Al riguardo si precisa che l'esposizione riguarda solo aspetti modellistici e prescinde da una qualunque considerazione realizzativa.

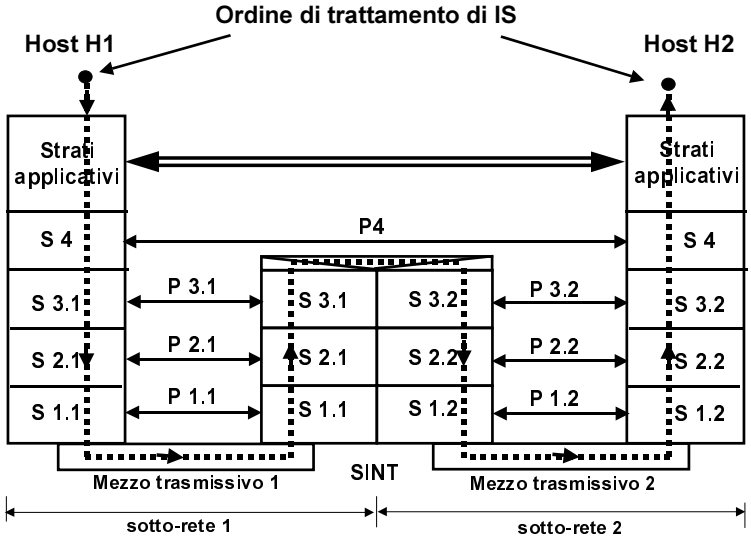


Figura II.8.2 – Architettura di sistema di interconnessione (SINT)

H1 agisce come sistema emittente. Conseguentemente l'ordine di trattamento del segmento *IS* in H1 è quello dei *ranghi decrescenti*. Si inizia dal sottosistema di rango più elevato che è presente in H1 e si scende via via attraverso i sottosistemi di rango intermedio fino al sottosistema di rango 1. Si conclude consegnando l'informazione al *mezzo trasmissivo 1*.

L'informazione associata al segmento *IS* perviene così a SINT che provvede a trattarla secondo le regole procedurali della sua interfaccia 1. Il ruolo di SINT in questo trattamento è quello di un sistema ricevente; quindi l'ordine di trattamento è quello dei *ranghi crescenti* e cioè dal sottosistema di rango 1 a quello di rango 3. Il trasferimento da H1 a SINT avviene quindi nel rispetto delle regole procedurali definite nella pila *{Px.1}*. In particolare si opera con incapsulamenti in H1 e con decapsulamenti in SINT; le unità di dati utilizzate (PDU, PCI, SDU) sono quelle della stessa pila *{Px.1}*.

La 3-SDU risultante dal trattamento sull'interfaccia 1 di SINT, viene poi consegnata al trattamento sull'interfaccia 2, ove diventa la SDU per il sottosistema di rango 3 in SINT. Questo sistema assume allora il ruolo di emittente. L'ordine di trattamento è quindi quello dei *ranghi decrescenti*.

L'informazione associata a *IS*, consegnata al *mezzo trasmissivo 2*, perviene così ad H2. In questo sistema, che agisce come ricevente, l'ordine di trattamento è quello dei *ranghi crescenti*. Si ha

quindi trasferimento da SINT a H2 nel rispetto delle regole procedurali della pila *{Px.2};* a questa stessa pila appartengono le unità di dati utilizzate da SINT per incapsulamenti e da H2 per decapsulamenti.

In conclusione, il segmento informativo *IS* viene consegnato alla sua destinazione. Nella Fig.II.8.2 l'ordine di trattamento di *IS* è rappresentato con linea tratteggiata. L'operazione di passaggio dalla pila *{Px.1}* alla *{Px.2}*, all'interno di SINT, è indicata graficamente in Fig.II.8.2 da un sottosistema con due segmenti diagonali.

Quanto detto non esaurisce però le funzioni che devono essere svolte dai sistemi coinvolti. Rimangono infatti da affrontare le questioni relative all'indirizzamento ed all'instradamento (cfr. par.V.5.).

III LE RISORSE DI RETE

Le risorse di una rete sono di natura varia; se tuttavia ci limitiamo all'insieme delle funzioni di tipo logico (cfr. § I.1.1), esse possono identificarsi in *risorse di trasferimento* e in *risorse di elaborazione*.

Le prime risiedono nei rami e nei nodi della rete di trasporto, oltre che nelle potenzialità della rete di accesso. Le seconde interessano principalmente i nodi della rete di trasporto, ma si incontrano anche nella rete di accesso al livello di interfaccia utente-rete; per esse possono individuarsi almeno tre finalità principali, e cioè il trattamento di chiamata nel caso di comunicazioni su base chiamata, il controllo dello scambio dell'informazione nel rispetto di opportune regole di procedura e di protezione e, infine, la gestione della rete.

Più in generale si può affermare, che, per l'espletamento delle loro funzioni, una rete e le sue parti componenti (apparecchiature nodali, organi di interfaccia, ecc.) devono svolgere *attività* e devono utilizzare *risorse*.

Una attività è un insieme coerente di azioni elementari che perseguono uno scopo definito e che utilizzano le risorse all'uopo necessarie. Una attività è quindi l'associazione evolutiva di risorse che concorrono al conseguimento di uno scopo comune.

Con riferimento all'interazione tra attività e risorse, questo capitolo chiarisce dapprima (par. III.1) la distinzione tra risorse indivise e risorse condivise: a conferma del ruolo fondamentale che queste ultime hanno nello sviluppo di una infrastruttura per telecomunicazioni, vengono esaminati i problemi posti dalla condivisione e alcuni dei criteri gestionali che consentono di risolverli. Il capitolo si conclude (par III.2) con la introduzione di alcuni concetti che sono di base per la comprensione della varietà di scelte sistemistiche da considerarsi per la realizzazione di un ambiente di comunicazione.

III.1 La condivisione delle risorse

Le risorse di una rete di telecomunicazione debbono essere commisurate in termini di quantità e di potenzialità (*vincolo di costo*) alle esigenze di un servizio qualitativamente accettabile (*vincolo di qualità di servizio*). Sorgono allora i problemi di:

- valutare le prestazioni conseguibili con date risorse (*problema di analisi*);
- dimensionare, qualitativamente e quantitativamente, le risorse quando siano fissate le prestazioni desiderate (*problema di sintesi*).

A questo riguardo, un criterio fondamentale, che trova la sua giustificazione nel rispetto del vincolo di costo, suggerisce di limitare allo stretto indispensabile l'impiego di *risorse indivise*, e cioè di risorse assegnate, in modo permanente o semi-permanente, allo svolgimento di una specifica attività.

E' invece preferibile, ove possibile, prevedere l'impiego di *risorse condivise*, che costituiscono un insieme, i cui elementi sono utilizzabili da più attività, anche se queste perseguono scopi distinti. Dato pero' che ogni risorsa può essere al servizio di una sola attività alla volta, una sua utilizzazione da parte di un insieme di attività deve avvenire, per ciascuna di queste, in intervalli di tempo distinti.

III.1.1 Domande e risposte

Nella loro evoluzione temporale, le attività devono poter utilizzare risorse (indivise o condivise) e quindi interagire con queste. L'interazione tra attività e risorse si manifesta con una *domanda* e con una *risposta*.

Le domande sono presentate dalle attività e sono costituite da *richieste di servizio* che mirano a una utilizzazione delle risorse; sono descritte da due componenti:

- la"*intensità" di richiesta* che misura, su una fissata base dei tempi, il numero di richieste di servizio presentate da attività interessate a una utilizzazione di risorse;
- la "*quantità" di lavoro* che una risorsa deve svolgere a seguito di una richiesta di servizio.

Il lavoro richiesto e l'unita' assunta per misurarne la quantità variano in relazione alla natura della risorsa e del suo utilizzatore. Ad esempio, nel caso in cui si consideri una risorsa di trasferimento e l'utilizzatore sia un messaggio di dati, l'unita' di lavoro corrisponde al trasferimento di una singola cifra binaria o a quello di una sequenza strutturata di cifre binarie che compongono il messaggio. Quando invece si tratta di una risorsa di elaborazione e l'utilizzatore e' un programma, come unita' di lavoro può essere assunto il trattamento di una singola istruzione tra quelle che compongono il programma.

Le risposte riguardano il *modo di utilizzazione* delle risorse come risultato della loro interazione con le attività; sono qualificate mediante numerosi parametri tra i quali si citano quelli relativi a:

- la"*efficienza*" *di utilizzazione*, con validità per risorse indivise o condivise;
- la "*facilità*" *di accesso*, con significatività solo per risorse condivise.

Nell'evoluzione temporale di una attività si distinguono in generale *intervalli di vitalità*, che si alternano ad altri *di latenza* nel corso dell'intera *durata dell'attività*, e cioè nell'intervallo tra l'inizio e la conclusione di detta evoluzione (Fig. III.1.1). In ogni intervallo di vitalità l'attività ha bisogno di ricevere un lavoro da parte di una risorsa.

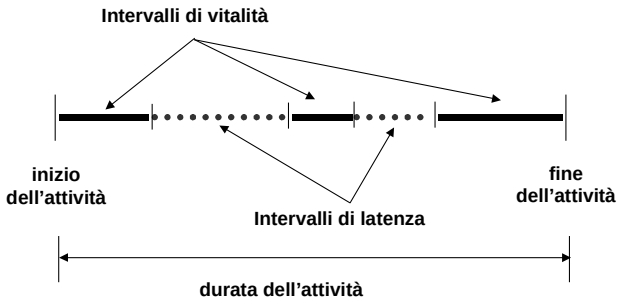


Figura III.1.1 – Evoluzione temporale di una attività

Inoltre le domande da parte di una attività si manifestano come fenomeni di *natura aleatoria*; in particolare:

- la presentazione degli intervalli di vitalità o di latenza, cosi' come quella delle durate di ogni attività, e' in generale di tipo aleatorio;
- altrettanto aleatoria e', sempre in generale, la quantità di lavoro che le risorse sono chiamate a svolgere a seguito delle richieste di servizio.

Riferiamoci ora, per facilità di riferimento, all'interazione in cui è coinvolta una singola risorsa che può essere indivisa o condivisa. L'estensione al caso di molteplici risorse non introduce particolari difficoltà.

La domanda è descrivibile con una grandezza che riassume le due componenti di intensità di richiesta e di quantità di lavoro. Detta grandezza è il *carico* (load) della risorsa, che, per definizione e con riferimento all'istante di osservazione *t,* è il *numero* $\lambda_g(t)$ *di unità di lavoro che la risorsa dovrebbe svolgere nell'unità di tempo per soddisfare la domanda;* cioè:

$$\textit{carico di una risorsa all'istante } t \equiv \lambda_g(t) ,$$

ove $\lambda_g(t)$ è una quantità aleatoria con distribuzione del primo ordine in generale dipendente dall'istante di osservazione *t*.

Alla domanda rappresentata dal carico, una risorsa risponde entro quanto consentito dalla sua *capacità*, che per definizione è *il numero massimo* γ *di unità lavoro che la risorsa è in grado di svolgere nell'unità di tempo;* cioè

$$\textit{capacità di una risorsa} \equiv \gamma \ .$$

La capacità riguarda esclusivamente la *potenzialità della risorsa* a svolgere i compiti che le sono propri; è quindi il "*dato di targa*" che qualifica la risorsa.

La risposta di una risorsa (indivisa o condivisa) al carico che la interessa *a seguito delle richieste di servizio che le sono presentate*, è costituita dalla *portata* (throughput), che, per definizione e con riferimento all'istante di osservazione *t*, è *il numero* $\lambda_s(t)$ *di unità di lavoro che la risorsa svolge nell'unità di tempo*; cioè

$$\textit{portata di una risorsa all'istante } t \equiv \lambda_s(t) \ ,$$

ove $\lambda_s(t)$, come $\lambda_g(t)$, è una quantità aleatoria con distribuzione del primo ordine in generale dipendente dall'istante di osservazione di *t*. Poiché, in qualunque istante *t*, deve essere

$$\lambda_s(t) \leq \gamma \ ,$$

la portata di una risorsa non può essere superiore alla capacità.

In generale, secondo la teoria dei processi aleatori, le distribuzioni (del primo ordine) delle variabili aleatorie che caratterizzano l'interazione tra attività e risorse dipendono dall'istante di osservazione o dalle condizioni di inizializzazione relative all'evoluzione temporale dell'interazione (*condizioni iniziali*). Esistono tuttavia casi in cui, in detta evoluzione temporale, si possono distinguere due fasi:

- una fase iniziale, che è chiamata *regime transitorio*, in cui l'interazione ha il comportamento generale sopra descritto;
- una fase successiva, che è chiamata *regime permanente* o *condizione* di *equilibrio statistico*, in cui la distribuzione (del primo ordine) delle variabili aleatorie associate diventa *indipendente* sia dall'instante di osservazione (diventa cioè *stazionaria*), che *dalle condizioni iniziali*.

Questa seconda fase è raggiungibile teoricamente solo dopo un tempo infinitamente grande rispetto all'istante iniziale dell'evoluzione, ma in pratica può attuarsi in intervalli di tempo finiti. Casi di questo tipo sono di prevalente interesse nell'analisi dei fenomeni qui considerati. Pertanto nel seguito ci riferiremo costantemente a condizioni di equilibrio statistico, nell'ipotesi che queste sussistano.

Una prima conseguenza di questa posizione è la possibilità di caratterizzare domanda e risposta attraverso *valori medi* di carico e di portata che sono indipendenti dall'istante di osservazione e dalle condizioni iniziali dell'evoluzione. Conseguentemente per il *carico medio* Λ_g e per la *portata media* Λ_s in condizioni di equilibrio statistico e indicando con E [•] l'operatore "valore atteso" della variabile aleatoria racchiusa tra le parentesi, valgono le seguenti definizioni:

$$\Lambda_g = E[\lambda_g(t)]$$

$$\Lambda_s = E[\lambda_s(t)]$$

che sono indipendenti dall'istante di osservazione *t.*

Sempre in condizioni di equilibrio statistico e come ulteriore parametro prestazionale della risposta risultante dall'interazione tra una risorsa e le attività interessate al suo uso, si può definire il *rendimento di utilizzazione* U di una risorsa, che è una qualificazione dell'efficienza di utilizzazione della risorsa e che corrisponde all'esigenza economica di limitare la quantità o la qualità delle risorse da rendere globalmente disponibili e di affrontare quindi un costo adeguato al benefico ottenibile. Il rendimento *U*,

che ha significatività per risorse sia indivise che condivise, è definito operativamente dal *rapporto tra la portata media Λ_s e la capacità della risorsa* γ:

$$U = \Lambda_s \ / \ \gamma ;$$

esprime quindi la *quota parte media del tempo* in cui la risorsa è utilizzata in base alla domanda esistente. Conseguentemente l'obiettivo prestazionale è assicurare per ogni risorsa un elevato rendimento di utilizzazione.

Quanto ora detto vale per risorse sia indivise che condivise. Nel seguito del paragrafo verranno chiariti i problemi connessi a una condivisione di risorse.

III.1.2 Le attività di gestione

L'accesso a un insieme di una o più risorse condivise deve essere opportunamente controllato. Per questo scopo, accanto alle *attività di utilizzazione* a cui ci si e' riferiti finora, devono essere presenti anche *attività di gestione*.

Le attività di utilizzazione perseguono gli scopi che sono a loro propri senza curarsi dei problemi posti dalla condivisione delle risorse con altre attività di utilizzazione. Per ciò che riguarda la loro interazione con le attività di gestione, esse si limitano a formulare a queste ultime le richieste del servizio che e' loro necessario.

Le attività di gestione debbono invece provvedere a controllare gli accessi alle risorse condivise a cui sono preposte. Il loro obiettivo e' quello di soddisfare la domanda in modo da rispettare i vincoli di costo e quelli di qualità di servizio. Con questi obiettivi, compiti fondamentali delle attività di gestione sono:

- risolvere le *condizioni di contesa*, dovute alla concorrenza delle richieste di servizio da parte delle attività di utilizzazione;
- minimizzare il rischio di *situazioni di stallo*, che si possono manifestare nell'evoluzione di due o più attività di utilizzazione aventi possibilità di accesso allo stesso insieme di risorse condivise;
- attuare le *strategie di assegnazione* delle risorse alle attività di utilizzazione che ne fanno richiesta;
- evitare, nei limiti del possibile, l'insorgere di *fenomeni di sovraccarico* delle risorse, in quanto questi determinano un drastico peggioramento delle prestazioni di qualità di servizio.

Per i primi due compiti verranno forniti qui di seguito (§ III.1.3 e III.1.4) alcuni dettagli, mentre sugli altri due compiti si ritornerà in § III.2.2 e III.2.3.

III.1.3 Condizioni di contesa

La concorrenza delle richieste di servizio presentate dalle attività di utilizzazione può determinare *condizioni di contesa*. Queste si verificano quando tutte le risorse disponibili sono occupate a favore di altrettante attività di utilizzazione in corso di evoluzione e vengono presentate nuove richieste di servizio. Le condizioni di contesa debbono essere risolte assicurando *equità di trattamento* nei confronti delle attività di utilizzazione. Come già si e' detto, questo e' uno dei compiti delle attività di gestione.

Questo compito può essere assolto con varie modalità. Qui se ne presentano alcune senza entrare nel merito delle soluzioni tecniche che possono essere adottate per la loro realizzazione. Si preferisce invece effettuarne una descrizione su un piano astratto.

Con quest'ultima precisazione, le principali modalità per risolvere condizioni di contesa sono quella *orientata al ritardo* e quella *orientata alla perdita*. Se ne chiarisce il significato facendo riferimento, per semplicità, alla disponibilità di una sola risorsa condivisa.

Nella *modalità orientata al ritardo*, l'accesso alla risorsa e' attuato tramite una fila d'attesa, che si assume di capienza illimitata. Si pongono allora in fila d'attesa tutte le attività di utilizzazione che presentano una richiesta di servizio quando la risorsa e' già occupata a favore di un'altra attività di utilizzazione (*stato di blocco*). Non appena la risorsa si libera da questo impegno, una delle attività di utilizzazione in fila d'attesa, scelta con un opportuno ordine di accesso, viene ammessa alla risorsa.

L'evento che caratterizza questa modalità e' allora quello di *ritardo*, che si presenta quando la risorsa e' in uno stato di blocco con il condizionamento che arrivi una nuova richiesta di servizio; cioè con riferimento all'istante di osservazione *t*

$$\textit{evento di ritardo nell'istante } t \equiv$$
$$\{\textit{stato di blocco in } t \mid \textit{arriva una nuova richiesta di servizio in } (t,\ t+dt)\}\ .$$

In tali situazioni, detta richiesta può essere accolta solo dopo lo stazionamento della attività di utilizzazione richiedente in fila d'attesa; ciò comporta un *ritardo di attesa*, corrispondente all'intervallo di tempo tra l'istante di arrivo della nuova richiesta e quello di ammissione alla risorsa.

Nella *modalità orientata alla perdita* l'accesso alla risorsa e' attuato senza fila d'attesa. Conseguentemente viene rifiutata ogni richiesta di servizio che sia presentata quando la risorsa e' in uno stato di blocco. L'attività di utilizzazione che ha subito questo trattamento dovra' ripresentare la sua richiesta di servizio in tempi successivi senza alcuna garanzia a priori di vederla accolta.

Pertanto, in questa modalità l'evento caratterizzante e' quello di *rifiuto*. Si tratta ancora di un evento condizionato, le cui condizioni di presentazione sono analoghe a quelle dell'evento di ritardo, ma con la differenza sostanziale, già sottolineata, che riguarda il trattamento delle attività di utilizzazione che ne subiscono gli effetti; cioè, sempre con riferimento all'istante di osservazione *t*, si definisce

$$\textit{evento di rifiuto nell'istante } t \equiv$$
$$\{\textit{stato di blocco in } t \mid \textit{arriva una nuova richiesta di servizio in } (t,\ t+dt)\}\ .$$

E' ovviamente anche possibile adottare una modalità intermedia tra le due ora descritte; si può cioè procedere con una *modalità orientata al ritardo con perdita*. A tale scopo e' sufficiente che l'accesso alla risorsa avvenga tramite una fila d'attesa avente capienza limitata. In tal modo si possono verificare ambedue gli eventi di rifiuto e di ritardo. In particolare, una attività di utilizzazione, che presenta la sua richiesta di servizio quando, oltre allo stato di blocco della risorsa desiderata, si verifica anche la completa occupazione della fila d'attesa, subirà lo stesso trattamento che si ha nella modalità orientata alla perdita. Se, invece, la fila d'attesa non e' completamente occupata, il trattamento e' quello della modalità orientata al ritardo.

Per qualificare la "*facilità*" *di accesso* a una risorsa condivisa e come ulteriore parametro prestazionale della risposta risultante dall'interazione tra la risorsa e la attività interessante al suo uso in presenza di condizioni di contesa, si fa riferimento al *grado di accessibilità* alla risorsa, e cioè a un parametro che:

- misura, in una opportuna metrica, *la qualità di servizio* di cui può godere una attività di utilizzazione per una sua evoluzione senza limitazioni percepibili;
- mette in evidenza i limiti di operatività connessi a un ambiente basato sull'impiego di risorse condivise.

Il grado di accessibilità è misurabile, in condizioni di equilibrio statistico, con quantità dipendenti dal modo di risoluzione delle condizioni di contesa. In ogni caso, tuttavia, l'obiettivo prestazionale è

assicurare, per ogni risorsa, un grado di accessibilità commisurato alle esigenze di qualità di servizio da parte delle attività di utilizzazione potenzialmente interessate a quella risorsa.

Nel caso in cui le contese sono risolte con modalità orientate al ritardo e in condizioni di equilibrio statistico, il grado di accessibilità a una risorsa condivisa può essere misurato, ad esempio, con la *probabilità di ritardo* Π_r definita da

$$\Pi_r = \wp\{\text{evento di ritardo in } t\},$$

ove con $\wp\{\bullet\}$ si indica la probabilità dell'evento descritto tra parentesi. Questa probabilità è esprimibile anche con la *quota parte media delle richieste di servizio che sono accolte solo con ritardo*.

Nel caso invece in cui le contese sono risolte con modalità orientate alla perdita e in condizioni di equilibrio statistico, il grado di accessibilità a una risorsa condivisa può essere misurato, ad esempio, con la *probabilità di rifiuto* Π_p definita da

$$\Pi_p = \wp\{\text{evento di rifiuto in } t\};$$

questa probabilità è esprimibile anche con la *quota parte media delle richieste di servizio che vengono rifiutate*.

Quando le contese sono risolte con modo orientato al ritardo con perdita, il grado di accessibilità ad una risorsa condivisa è misurabile, sempre in condizioni di equilibrio statistico con entrambe le probabilità di ritardo e di rifiuto.

Si sottolinea che l'ipotesi dell'equilibrio statistico assicura che le probabilità dei due eventi di ritardo e di rifiuto sono indipendenti dall'istante di osservazione e dalle condizioni iniziali dell'interazione tra attività e risorse.

III.1.4 Stati di stallo

Gli stalli sono un fenomeno tipico in un ambiente di risorse condivise e possono compromettere l'evoluzione di due o più attività di utilizzazione. Esaminiamo un esempio dell'insorgere di stati di questo tipo con riferimento a due sole attività di utilizzazione. Se l'attività di utilizzazione *A* conserva il possesso di una risorsa α in attesa della risorsa β che è in possesso dell'attività di utilizzazione *B* e se quest'ultima ha necessità di accedere alla risorsa α per proseguire la sua evoluzione (Fig. III.1.2), le due attività si bloccano mutuamente in un "*abbraccio mortale*" (dead lock).

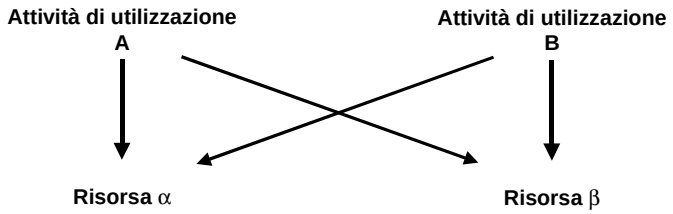


Figura III.1.2 – Condizione in cui due attività di utilizzazione A *e* B *si bloccano mutuamente*

Per minimizzare il rischio di stati di stallo tra le attività di utilizzazione in una rete di telecomunicazione, possono essere adottati vari accorgimenti, che ben si differenziano da quelli largamente studiati in sistemi con controllo centralizzato: in una rete, infatti, l'interazione tra attività di utilizzazione e risorse condivise avviene solitamente nell'ambito di *controlli distribuiti*, che sono compito delle singole attività di gestione.

Tra tali accorgimenti uno dei più frequentemente utilizzati, nel caso di risoluzione delle condizioni di contesa orientata al ritardo, consiste nell'adozione sistematica di *temporizzatori* per il ritardo di attesa nell'accesso ad ogni risorsa. In queste condizioni, quando tale ritardo relativo ad una risorsa specifica

ha superato un limite opportunamente fissato (*tempo massimo*), si suppone che si sia verificato uno stato di stallo. Si libera allora quella risorsa, interrompendo l'attività di utilizzazione che ne era in possesso, e si effettua un recupero dell'errore procedurale che conseguentemente si determina.

III.1.5 Campi di impiego delle risorse

E' naturale domandarsi a questo punto quali siano i campi di conveniente impiego delle risorse indivise e di quelle condivise. Per rispondere a questa domanda si farà riferimento ai parametri che sono stati definiti in § III.1.1 e III.1.3 per qualificare la domanda e le prestazioni derivanti dall'interazione tra attività di utilizzazione e risorse indivise o condivise.

Se una risorsa e' indivisa, il carico medio che la riguarda e' determinato dalle sole richieste dell'attivita' di utilizzazione che ne ha disponibilita' esclusiva. Invece, il carico medio su una risorsa condivisa e' la risultante delle richieste presentate da tutte le attività di utilizzazione che hanno accesso alla risorsa.

Circa la portata media, questa, nel caso di risorsa indivisa, può risultare decisamente inferiore alla relativa capacita' se, come talvolta si verifica, la domanda comporta un basso valore del carico medio. Inoltre non possono presentarsi condizioni di sovraccarico della risorsa, se l'attività che ne ha disponibilita' esclusiva contiene la sua domanda entro limiti compatibili con la capacita' della risorsa. Una situazione contraria si presenta invece nel caso di risorse condivise; qui, ancora in relazione alla modalita' di risoluzione delle condizioni di contesa, puo' intervenire l'esigenza di controllare i fenomeni di sovraccarico.

Nel caso di soluzione con accesso indiviso, il rendimento di utilizzazione di una risorsa puo' essere di piccolo valore in relazione alle caratteristiche della domanda dell'unica attività avente diritto di accesso. Quindi, dato che e' invece senza limitazioni il grado di accessibilita', la soluzione con risorse indivise dovrà essere ristretta a quei casi in cui sia soddisfatta almeno una delle condizioni seguenti:

- sia assicurato un elevato carico medio da parte dell'attività avente accesso indiviso;
- risulti fortemente prevalente l'esigenza di conseguire un accesso senza contese e, quindi, senza ritardi o senza rifiuti.

Nel caso invece di soluzione con accesso condiviso, il rendimento di utilizzazione di una risorsa può essere di valore più soddisfacente rispetto alla soluzione precedente. Ma, corrispondentemente, il grado di accessibilita' si riduce e in misura tanto maggiore quanto più elevato e' il rendimento di utilizzazione della risorsa. Questa soluzione e' quindi, in generale, preferibile alla precedente per il rispetto del vincolo di costo, ma richiede, per il rispetto del vincolo di qualità di servizio, una attenta soluzione dei relativi problemi di dimensionamento e di gestione delle risorse.

III.2 Traffico e prestazioni

Viene completata la trattazione iniziata nel precedente par. III.1. Più in dettaglio si chiarisce dapprima (§ III.2.1) il concetto di *risorsa virtuale*, che è di particolare importanza per la trattazione dei successivi capitoli. Con riferimento poi ad un ambiente di risorse condivise, vengono presentati due ulteriori compiti delle attività di gestione citati in § III.1.2, e cioè l'attuazione delle *strategie di assegnazione* (§ III.2.2) e il *controllo dei sovraccarichi* (§ III.2.3).

Elemento comune di questi argomenti e' il legame tra fenomeni di traffico nell'interazione tra attività e risorse da un lato e le conseguenti prestazioni di efficienza e di accessibilità dall'altro.

III.2.1 Risorse virtuali

In un ambiente di condivisione, una *risorsa fisica R* condivisa, sotto il controllo di una attivita' di gestione, può essere impegnata, in intervalli di tempo distinti, da parte di più attività di utilizzazione. Si supponga che la gestione di *R* sia tale che

- la domanda di ognuna di queste attivita' sia pienamente soddisfatta;
- il grado di accessibilità sia rispondente alle esigenze connesse all'evoluzione di ogni attivita' di utilizzazione.

In queste condizioni, ciascuna delle attivita' di utilizzazione, ad esempio, la *i*-esima, dato che non percepisce limitazioni al soddisfacimento della sua domanda e alle sue esigenze di grado di accessibilita', utilizza la risorsa fisica *R* come se questa fosse a lei riservata, e cioè come se fosse una risorsa R_i a sua disposizione esclusiva.

Tale tipo di disponibilita' e' pero' solo apparente, in quanto la risorsa R_i è solo l'immagine di *R* vista dalla *i*-esima attività di utilizzazione con il condizionamento determinato dal suo carico (domanda di servizio) e dalle sue esigenze di grado di accessibilita' (prestazioni necessarie).

Per sottolineare tale distinzione, l'immagine R_i di *R* è chiamata *risorsa virtuale associata alla i-esima attività di utilizzazione*. In Fig. III.2.1 e' illustrata la corrispondenza tra la risorsa fisica *R* e le risorse virtuali associate alle *n* attività di utilizzazione che accedono a *R*.

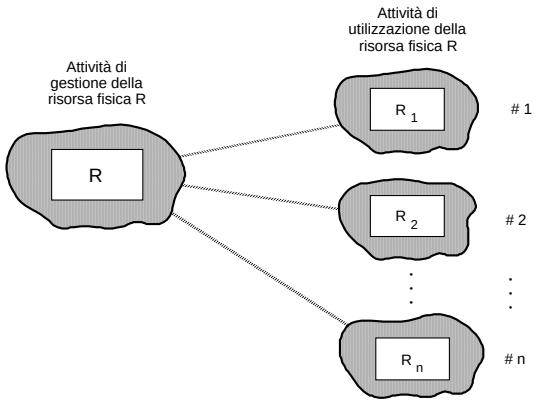


Fig. III.2.1 - Corrispondenza tra la risorsa condivisa R e le relative risorse virtuali R_i (i = 1, 2, ..., n)

Si rimarca che l'immagine di cui si parla e' di tipo condizionato dalla domanda offerta e dalle prestazioni desiderate da parte di una particolare attivita' di utilizzazione. Pertanto, e' possibile parlare di risorsa virtuale per detta attivita' solo con la precisazione a priori di questi dati di richiesta e solo dopo la verifica a posteriori della possibilita' di soddisfare quanto richiesto nell'interazione tra la risorsa fisica *R* e le attivita' che accedono ad essa.

III.2.2 Strategie di assegnazione di risorse condivise

Le strategie di assegnazione definiscono le modalità di accesso a un insieme di risorse condivise da parte di un gruppo di attività di utilizzazione sotto il controllo delle attività di gestione. Possono attuarsi secondo le alternative schematizzate in Fig. III.2.2.

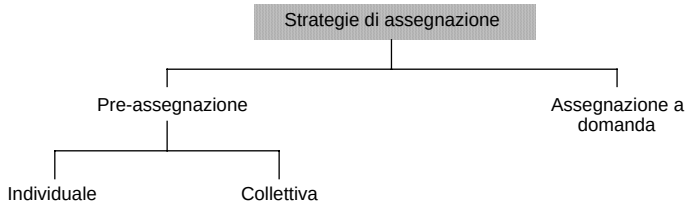


Fig. III.2.2 - Alternative di attuazione per le strategie di assegnazione delle risorse.

Nella strategia di *assegnazione a domanda*, una risorsa viene assegnata ad una attività di utilizzazione solo a seguito di una richiesta di servizio effettuata immediatamente a monte di un intervallo di vitalita' e viene restituita non appena questo intervallo è terminato. Cioè, nel corso della durata di una attività di utilizzazione, questa può impegnare la risorsa più volte per intervalli di tempo minori di detta durata e, negli intervalli residui, la risorsa può essere impegnata da altre attività. Ciò può assicurare, per la risorsa in questione, un rendimento di utilizzazione anche elevato, a spese però di un grado di accessibilità che può risultare insufficiente.

Quando i requisiti prestazionali relativi al grado di accessibilità sono più stringenti è preferibile adottare differenti strategie di assegnazione, e cioè quelle di *pre-assegnazione*. In questo caso una attività di utilizzazione presenta una richiesta di assegnazione immediatamente a monte della sua evoluzione temporale. Se la risposta a tale richiesta è favorevole, l'assegnazione è mantenuta per tutta la durata dell'attività. In questo ambito si possono però distinguere almeno due alternative, che si distinguono in base all'oggetto dell'operazione di pre-assegnazione.

Nella prima di queste, chiamata *pre-assegnazione individuale*, la risorsa fisica viene assegnata in modo indiviso ad una sola attività per tutta la sua durata. Il vantaggio di tale alternativa è quello di prospettare una probabilita' di ritardo o una di rifiuto non nulle solo all'inizio dell'evoluzione dell'attività di utilizzazione e di assicurare successivamente, per la durata di questa, una accessibilità senza ritardi o senza rifiuti. Lo svantaggio è quello di monopolizzare una risorsa a vantaggio di una sola attività di utilizzazione per volta, seppure limitatamente alla durata di questa. Il risultato possibile è allora quello di un insufficiente rendimento di utilizzazione.

Una differente alternativa di pre-assegnazione, che cerca di combinare i vantaggi offerti dalle due strategie precedentemente descritte, è la *pre-assegnazione collettiva*. In questo caso la risorsa fisica può essere impegnata da più attività di utilizzazione, seppure in intervalli di tempo distinti, con un meccanismo di accesso del tutto analogo a quello della assegnazione a domanda. Ciò è però possibile solo per un *numero controllato* di attività di utilizzazione, che sono contemporaneamente in possesso di una *autorizzazione di accesso*.

Per ottenere quest'ultima, una attivita' di utilizzazione interessata presenta, immediatamente a monte della sua evoluzione, una esplicita richiesta in cui siano precisate le caratteristiche della sua domanda e le sue esigenze di grado di accessibilita'. Detta richiesta, che è trattata dalle attività di gestione, viene accolta o meno in base all'applicazione di una opportuna *regola di decisione*. Il risultato favorevole è la pre-assegnazione di una risorsa virtuale.

La regola di decisione per una pre-assegnazione collettiva si basa su dati forniti dalle attivita' di utilizzazione richiedenti e su una loro elaborazione effettuata dall'attivita' di gestione preposta alla decisione. I dati in questione sono:

- le caratteristiche della domanda (*dati di traffico preesistente*) e le esigenze di grado di accessibilita' (*requisiti prestazionali preesistenti)* per ognuna delle attivita' di utilizzazione le cui richieste di accesso sono gia' state accolte in precedenza;
- dati analoghi relativi all'attivita' di utilizzazione la cui richiesta e' oggetto della decisione (*dati di traffico richiesto* e *relativi requisiti prestazionali*).

L'elaborazione riguarda in primo luogo i dati di traffico preesistente, in modo da valutare lo stato di carico della risorsa, risultante dalle richieste gia' accolte (*stato di carico di riferimento*). Vengono poi utilizzati i dati di traffico richiesto per valutare come verrebbe modificato lo stato di carico di riferimento se la nuova richiesta fosse accolta (*stato di carico di riferimento modificato*). Questo secondo stato costituisce successivamente l'ingresso per una valutazione delle prestazioni. Se i gradi di accessibilita' per le attivita' di utilizzazione gia' accolte e per quella presentante la nuova richiesta, quali risultano da questa valutazione, rispettano i requisiti prestazionali preesistenti e quelli richiesti, l'autorizzazione di accesso viene concessa; in caso contrario, viene negata.

In tutti i tre tipi di strategie di assegnazione considerate in Fig. III.2.2 le risorse fisiche o virtuali, che sono assegnabili, appartengono a un insieme di dimensioni limitate. Le attività di gestione debbono allora risolvere le *condizioni di contesa*, che si verificano quando tutte le risorse disponibili sono state assegnate e vengono presentate nuove richieste di assegnazione.

In particolare, condizioni di contesa si possono presentare in due casi:

- in una operazione di pre-assegnazione individuale o collettiva (*contesa di pre-assegnazione*) all'inizio dell'evoluzione di una attività di utilizzazione;
- nel corso dell'evoluzione di varie attività di utilizzazione, se queste agiscono nell'ambito di strategie con assegnazione a domanda o con pre-assegnazione collettiva (*contesa di utilizzazione*).

III.2.3 Controllo dei sovraccarichi

Quando si fa crescere il carico medio su un insieme di risorse condivise, si può osservare che, al di sopra di un carico-limite (*soglia di sovraccarico*), le prestazioni dell'insieme peggiorano rapidamente. In particolare, diminuisce bruscamente il grado di accessibilita' per le singole attività di utilizzazione e diminuisce, altrettanto bruscamente, la portata media dell'insieme.

Per evitare condizioni di questo tipo (*condizioni di congestione*), che possono condurre rapidamente al completo collasso dell'insieme, occorre affidare alle attività di gestione compiti di controllo che consentano di "*filtrare*" le domande di accesso in modo che:

- l'insieme operi con un carico medio che sia sempre inferiore alla soglia di sovraccarico;
- le risorse componenti siano utilizzate in modo efficiente.

Si tratta, come e' evidente, di obiettivi tra loro contrastanti, dato che questa seconda condizione indurrebbe a lavorare in prossimità della soglia. La soluzione deve essere quindi di compromesso.

Il criterio di filtraggio delle domande d'accesso è normalmente basato sulla individuazione di risorse critiche e consiste nel limitare, per queste, il rendimento di utilizzazione che si conseguirebbe nel caso in cui una nuova domanda fosse accolta.

IV I SERVIZI DI RETE

Le esigenze di comunicazione in un ambiente di servizi integrati a banda stretta o larga, con o senza mobilità, possono essere soddisfatte tenendo conto delle caratteristiche dei flussi informativi di utente che devono essere trattati. Questi flussi hanno origine da sorgenti numeriche caratterizzate da ritmi binari di picco compresi in una dinamica di almeno sei decadi, e cioè tra un limite inferiore minore di 1 kbit/s e un limite superiore di circa 1 Gbit/s. Le sorgenti possono essere CBR o VBR e, in questo secondo caso, il grado di intermittenza può assumere valori che vanno da qualche unità ad alcune decine.

Il trattamento di flussi informativi di utente con queste caratteristiche richiede che la rete renda disponibile, una potenzialità di trasferimento ad alta capacità in corrispondenza con l'interfaccia utente-rete e con le apparecchiature della rete di trasporto. Tale capacità, a cui nel seguito verrà talvolta sostituito, come sinonimo, il termine *banda disponibile*, dovrebbe infatti essere superiore di almeno due ordini di grandezza a quella fornita in un ambiente a banda stretta.

Tra gli svariati problemi che sono stati affrontati per perseguire questi obiettivi il presente capitolo ne focalizza alcuni di particolare rilievo, e cioè quelli riguardanti le tecniche di trasferimento dell'informazione con prevalente orientamento ad un ambiente di comunicazioni fisse. In questo quadro

- si parte dalla presentazione dei *servizi di rete*, precisandone le caratteristiche (par. IV.1);
- si passa poi all'analisi dei *modi di trasferimento* (par. IV.2 – IV.5);
- si considerano i principi di funzionamento dei *nodi di rete* (IV.6-IV.7);
- si conclude con i problemi posti dall'*interconnessione di segmenti di rete* (par. IV.8);

IV.1 Caratteristiche di servizio

Si è già visto (cfr. cap. I) come la fornitura di un servizio di rete sia richiesta per lo svolgimento di una comunicazione tra due o più utenti. Compito di questo servizio è provvedere a trattare l'informazione emessa nell'ambito della comunicazione, gestendo:

- l'attuazione di una specifica strategia di assegnazione delle risorse (di trasferimento e/o elaborazione) che sono condivise;
- la risoluzione delle contese di pre-assegnazione e di utilizzazione, ove queste si presentino.

Le caratteristiche e le prestazioni di un servizio di rete debbono essere adattate a quelle richieste nella fornitura del *servizio applicativo*, che fruisce di quanto messo a sua disposizione dal servizio di rete.

Il modello di un servizio di rete è il seguente:

- i suoi *utenti* sono le *sorgenti* che emettono informazioni e i *collettori* a cui le informazioni debbono essere inoltrate;
- suo *fornitore* è l'insieme delle funzioni e dei protocolli che concorrono a trasferire l'informazione da una sorgente ad uno o più collettori, nel rispetto di definiti requisiti prestazionali.

La base per la fornitura di un servizio di rete è un *modo di trasferimento*, e cioè la modalità operativa per trasferire informazione attraverso la rete logica.

Ogni sorgente emette informazione sotto forma di *stringhe di cifre binarie*. Queste stringhe possono essere sostanzialmente di due tipi, a seconda della loro appartenenza a:

- un *flusso intermittente*, che è organizzato *a messaggi* (message type);

- un *flusso continuo* (senza soluzione di continuità), che si presenta sotto forma di un *getto ininterrotto* (stream type) di cifre binarie con ritmo di emissione (cfr. § I.2.2) costante (CBR) o variabile (VBR).

Indipendentemente dal tipo di flusso emesso dalle sorgenti, l'informazione da trattare in un servizio di rete è presentata al fornitore sotto forma di stringhe di cifre binarie, che sono sezionate in RCB *(raggruppamenti di cifre binarie)*. In alcuni casi gli RCB relativi a un servizio di rete debbono essere completati con una *intestazione*. In questi casi, che verranno motivati nel seguito, l'insieme di uno o più RCB (*testo*) e di una intestazione (Fig. IV.1.1) verrà chiamato *Unità Informativa* (UI). Si parlerà infine di *segmento informativo* quando si vuole alludere senza differenziazione a una UI o a un RCB.

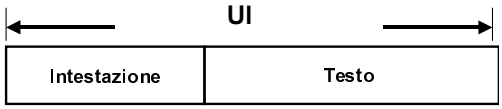


Figura IV.1.1 – Struttura di una Unità Informativa

IV.1.1 Prestazioni

Le prestazioni più significative di un servizio di rete riguardano la *flessibilità di accesso*, l'*integrità informativa* e la *trasparenza temporale*.

Il *grado di flessibilità di accesso* misura l'adattabilità del servizio di rete:

- nel trattare flussi informativi aventi origine da sorgenti con *ritmo di emissione* e con caratteristiche di *attività* (burstiness) tra loro anche molto diverse;
- nell'assicurare, *in modo indipendente* dalle caratteristiche dei flussi informativi, un accettabile *rendimento di utilizzazione* delle risorse condivise.

Per definire l'integrità informativa e la trasparenza temporale consideriamo poi le diversità che si possono manifestare, in relazione a una particolare modalità di trasferimento, tra

- la sequenza di cifre binarie emessa dalla sorgente nell'ambito del servizio considerato (*sequenza di emissione*);
- la corrispondente sequenza di cifre binarie ricevute, e cioè a valle dell'operazione di trasferimento dell'informazione dalla sua origine alla sua destinazione (*sequenza di ricezione*).

L'integrità informativa riguarda le diversità, che si possono manifestare in modo aleatorio,

- tra le singole cifre binarie emesse e quelle corrispondenti ricevute (*integrità di cifra binaria*);
- tra le intere sequenze di emissione e di ricezione o tra segmenti corrispondenti di queste (*integrità di sequenza di cifre binarie*).

Nel primo caso le diversità sono dovute a *errori*, normalmente di tipo trasmissivo; nel secondo sono addebitabili a errori procedurali o a specifiche modalità di trasferimento. Il *grado di integrità informativa* qualifica in termini prestazionali l'integrità di cifra binaria o quella di sequenza binaria. È tanto più elevato quanto minore è la diversità (*distanza*) tra le due sequenze di emissione e di ricezione, misurata, ad esempio nel primo caso, dal *tasso di errore binario*.

Un elevato grado di integrità informativa è un requisito particolarmente importante quando l'informazione scambiata è costituita da dati. Conseguentemente, nei servizi di comunicazione di dati, occorre normalmente mettere in atto procedure protettive, che siano in grado di recuperare gli errori quando questi vengano rivelati ovvero che riescano a correggere gli errori quando questi si manifestano. L'obiettivo di queste procedure protettive è contenere il tasso di errore binario *residuo* a valori orientativamente inferiori a 10^{-10}.

Nel caso invece in cui si tratti di informazione audio, il grado di integrità informativa che è richiesto dipende dal mezzo di rappresentazione che si è scelto. Se a tale mezzo è associata una forma di codifica senza riduzione di ridondanza, detto grado può essere di valore decisamente inferiore al caso dell'informazione di dati. Deve invece essere via via aumentato man mano che si riduce la ridondanza dell'informazione sottoposta a codifica.

La trasparenza temporale riguarda invece i *ritardi di trasferimento* che differenti segmenti della sequenza di ricezione possono presentare rispetto ai corrispondenti segmenti della sequenza di emissione. Questi ritardi sono dovuti a cause di natura varia agenti, in generale, in modo aleatorio. Il *grado di trasparenza temporale* può essere quindi valutato quantitativamente con un parametro che qualifichi la variabilità dei ritardi di trasferimento e che sia di valore tanto più elevato quanto minore è tale variabilità. Questo parametro ha allora valore massimo quando la variabilità dei ritardi di trasferimento è nulla o di entità trascurabile, cioè quando i ritardi di trasferimento sono di valore praticamente costante. Non lo ha quando tale condizione non è soddisfatta. La distinzione tra queste due condizioni è chiarita graficamente in Fig. IV.1.2.

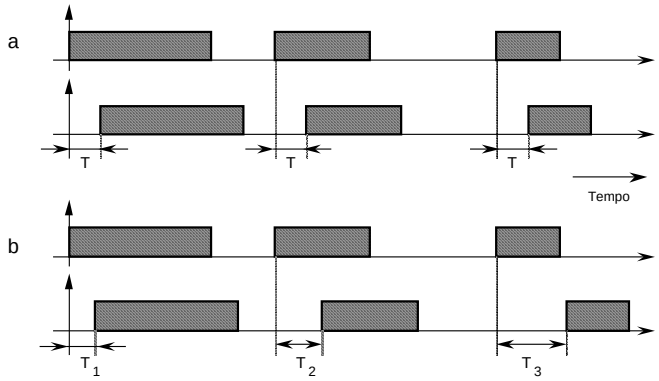


Figura IV.1.2 - Sequenze di emissione e di ricezione nei due casi di servizio di rete (a) temporalmente trasparente e (b) non temporalmente trasparente

Esistono servizi, per i quali una corretta interpretazione dell'informazione in ricezione richiede che il trasferimento avvenga con un massimo grado di trasparenza temporale. Ne esistono invece altri che non hanno esigenze analoghe. Esempi significativi di servizi che richiedono un elevato grado di trasparenza temporale si incontrano in tutti quei casi in cui l'informazione da trasferire sia il risultato di una operazione di conversione analogico-numerica, e cioè in cui dal segnale analogico originario si estrae una sequenza di campioni, che occorre restituire in ricezione con la stessa periodicità dell'operazione di campionamento. Pertanto casi tipici in questo ambito sono offerti da servizi audio e video. Non rientrano invece in queste esigenze i servizi di dati.

Un servizio per il quale è richiesto un elevato grado di trasparenza temporale può utilizzare un trasferimento anche non temporalmente trasparente. Ciò può avvenire solo se la distribuzione dei ritardi di trasferimento è tale da presentare un valore medio sufficientemente basso e deviazioni intorno a questo sufficientemente contenute. In tal caso, in ricezione, si può operare una *equalizzazione dei ritardi*, che eleva il grado di trasparenza temporale senza pregiudicare quello di integrità informativa.

Ai fini della qualità del servizio è di interesse anche il valore di picco dei ritardi di trasferimento. Questo deve essere inferiore a un limite che dipende dal servizio considerato e che è tanto più basso quanto maggiori sono le esigenze di interattività della comunicazione, come nel caso dei servizi di conversazione. Nel caso della telefonia questo limite deve essere inferiore a circa 300 ms.

I gradi di flessibilità di accesso, d'integrità informativa e di trasparenza temporale hanno valori dipendenti dallo stato di condivisione delle risorse impegnate dal servizio di rete. In particolare il

valore conseguito dal grado di flessibilità di accesso dipende dalla scelta della *strategia di assegnazione*; infatti, come è facile convincersi, diminuisce per pre-assegnazioni individuali e aumenta per assegnazioni a domanda o per pre-assegnazioni collettive.

I valori dei gradi di integrità informativa e di trasparenza temporale dipendono invece dal modo di *risoluzione delle contese di utilizzazione*. Più in particolare il grado di integrità informativa diminuisce per contese risolte a perdita e aumenta per contese risolte a ritardo. Inoltre il grado di trasparenza temporale aumenta per contese risolte a perdita e diminuisce per contese risolte a ritardo.

In Fig. IV.1.3 sono riassunte le prestazioni di un servizio di rete e l'influenza che su di esse hanno le strategie di assegnazione e le risoluzioni delle contese di utilizzazione.

PARAMETRO PRESTAZIONALE	SENSIBILITA'
Grado di flessibilità di accesso	alla *strategia di assegnazione* •diminuisce per pre-assegnazione individuale •aumenta per assegnazione a domanda o per pre-assegnazione collettiva
Grado di integrità informativa	al *modo di risoluzione delle contese di utilizzazione* •diminuisce per contese risolte a perdita •aumenta per contese risolte a ritardo
Grado di trasparenza temporale	al *modo di risoluzione delle contese di utilizzazione* •diminuisce per contese risolte a ritardo •aumenta per contese risolte a perdita

Figura IV.1.3 – Prestazioni di un servizio di rete e loro sensibilità alle strategie di assegnazione e alla risoluzione delle contese di utilizzazione

IV.1.2 Relazioni tra le parti in comunicazione

Un servizio di rete può essere attuato con due modalità alternative, e cioè nei modi *con connessione* e *senza connessione*. Le opportunità offerte da ciascuno di questi modi sono già state illustrate con riferimento ai modi di servizio di strato (cfr. § II.3.5); qui vengono riproposte con la precisazione che trattasi di attuazioni riguardanti un particolare strato che appartiene all'insieme di trasferimento e nel quale sono svolte le funzioni di multiplazione e di commutazione. Questo strato, che è chiamato "*di modo di trasferimento*" (o in breve MT), verrà definito più in dettaglio nel par. IV.4. La scelta del modo di servizio nello strato MT non condiziona la scelta relativa agli strati di rango più alto o più basso. Inoltre le UI che sono oggetto di trattamento da parte delle funzioni dello strato MT sono le PDU (cfr. par. II.4) definite nell'ambito dei protocolli ivi operanti. Nel tempo, in relazione a scelte diverse sullo strato MT e sulle architetture protocollari che lo definiscono, sono state utilizzate per le UI denominazioni diverse, legate allo specifico protocollo che ne ha specificato il formato. I nomi adottati sono stati, ad esempio, *pacchetto*, *datagramma*, *cella* e *trama*.

Nel caso di un servizio di rete *con connessione* e di una comunicazione *punto-punto*, il fornitore deve provvedere a instaurare un *percorso da estremo a estremo* e cioè tra gli utenti del servizio. Questo percorso è un'associazione temporanea di connessioni da nodo a nodo ed è identificato dalla sequenza di rami e di nodi della rete logica che occorre attraversare per passare da un estremo all'altro della rete. Una volta che detto percorso sia stato instaurato, il meccanismo di controllo tratta i segmenti informativi conservandone l'identità tra gli estremi del percorso. Quando il percorso deve essere

abbattuto, vengono liberate le risorse fisiche o virtuali che sono state utilizzate per connettere gli utenti del servizio di rete.

Il caso di comunicazione punto-punto è immediatamente generalizzabile a quello multi-punto, ove le connessioni debbono essere *punto-multipunto* o *multipunto-punto* o *multipunto-multipunto* a seconda della configurazione della comunicazione (cfr. § I.3.3).

In un servizio di rete *senza connessione*, una comunicazione si svolge senza l'instaurazione preventiva di un percorso di rete da estremo a estremo. Gli utenti consegnano al fornitore i loro segmenti informativi e questi vengono trattati come unità indipendenti, e cioè senza alcun legame con la comunicazione nell'ambito della quale vengono scambiati. Il servizio di rete garantisce l'accettazione di tali segmenti nel punto di accesso di origine e, entro certi limiti, la loro consegna nel punto di accesso di destinazione.

I due servizi con e senza connessione sono in stretta relazione con le strategie di assegnazione delle risorse adottate per realizzare un trasferimento di informazione. Il caso con connessione implica che siano attuate le operazioni di instaurazione e di abbattimento e che nell'intervallo temporale tra queste avvenga la fruizione del servizio di rete: quando la connessione viene instaurata, il fornitore del servizio assegna risorse fisiche o virtuali che rimangono a disposizione degli utenti per tutta la durata del servizio e che vengono liberate quando la connessione viene abbattuta. Ne consegue che un servizio di rete con connessione è attuabile con una strategia di *pre-assegnazione*, che è individuale o collettiva a seconda che le risorse assegnate siano fisiche o virtuali, rispettivamente.

Passando poi al caso senza-connessione, l'assegnazione delle risorse si verifica solo a seguito di esigenze di una loro utilizzazione e viene mantenuta solo fino a quando queste esigenze vengano a mancare, senza tenere conto di eventuali loro presentazioni successive. Se ne deduce che un servizio di rete senza connessione è attuabile con una strategia di *assegnazione a domanda*.

Le relazioni tra modi di un servizio di rete e strategie di assegnazione per una loro attuazione sono riassunte in Fig. IV.1.4.

MODO DI SERVIZIO	STRATEGIA DI ASSEGNAZIONE
Con connessione fisica	Pre-assegnazione individuale
Con connessione virtuale	Pre-assegnazione collettiva
Senza connessione	Assegnazione a domanda

Figura IV.1.4 – Modi di un servizio di rete e associate strategie di assegnazione delle risorse

Il caso di servizio di rete con connessione prevede due ulteriori modalità alternative e cioè una connessione *commutata* e una connessione *semi-permanente* (Fig. IV.1.5). Una terza modalità è rappresentata da una connessione *permanente*, che non coinvolge però funzionalità del servizio di rete. Non verrà quindi ulteriormente considerata.

Una connessione commutata può essere instaurata in un qualunque istante a seguito di una richiesta da parte di un utente del servizio di rete. L'abbattimento si verifica quando la connessione non è più necessaria. Il caso di connessione commutata è quello usuale per una *comunicazione su base chiamata*. Un nodo di rete che tratta connessioni commutate opera secondo tre fasi temporalmente distinte:

- l'*instaurazione* del percorso da estremo a estremo;

- il *trasferimento* dell'informazione;
- l'*abbattimento* del percorso.

L'instaurazione avviene per iniziativa dell'utente chiamante; il trasferimento dell'informazione ha luogo secondo le esigenze del servizio applicativo; l'abbattimento si attua su richiesta di uno qualunque degli utenti del servizio.

MODO DI SERVIZIO	TIPO DI CONNESSIONE
Con connessione fisica o virtuale	commutata
	semi-permanente

Figura IV.1.5 – Tipi di connessione in relazione ai modi di servizio di rete

Per una comunicazione *su base chiamata*, l'instaurazione di una connessione commutata è tradizionalmente inclusa tra le funzioni svolte da un nodo di rete per *trattare la chiamata*: è quanto si verifica, ad esempio, nei nodi della rete telefonica. Oggi la tendenza è prevedere l'instaurazione di una connessione *indipendentemente* dal trattamento di chiamata: infatti, in un ambiente di comunicazione multimediale, è rilevante poter instaurare connessioni con caratteristiche prestazionali differenziate in relazione alla possibile varietà di esigenze di servizio da parte dei flussi informativi da trasferire. In ogni caso l'instaurazione di una connessione commutata è richiesta al nodo interessato tramite *messaggi di segnalazione*. Conseguentemente, un nodo in grado di instaurare una connessione commutata deve avere la capacità di *trattare l'informazione di segnalazione*.

Una *connessione semi-permanente* è instaurata a seguito di un accordo contrattuale tra cliente e fornitore del servizio; ha durata che, in relazione a tale accordo, può essere indefinita ovvero definita nell'ambito di un opportuno intervallo di tempo (un giorno, una settimana, ecc.). Nel caso di connessione semi-permanente, il coinvolgimento dei nodi è del tutto analogo a quello che si verifica per le connessioni commutate, almeno per ciò che riguarda la fase di trasferimento dell'informazione; non si verifica invece per ciò che riguarda le operazioni di instaurazione e di abbattimento della connessione. L'instaurazione di una connessione semi-permanente è richiesta al nodo interessato tramite *messaggi gestionali*. Ne segue che un nodo in grado di gestire solo connessioni semi-permanenti non ha funzionalità di trattamento di chiamata. Un nodo con queste caratteristiche è chiamato "*cross-connect*".

IV.1.3 Componenti di un modo di trasferimento

Le componenti di un servizio di rete e del modo di trasferimento ad esso associato sono:

- la *multiplazione*;
- la *commutazione*;
- l'*architettura protocollare*.

Lo schema di multiplazione identifica le modalità logiche adottate per utilizzare la capacità di trasferimento dei *rami delle reti* di accesso e di trasporto, e cioè i modi in cui la banda disponibile di questi rami viene condivisa logicamente dai flussi informativi che li attraversano.

Il principio di commutazione riguarda i concetti generali sui quali è basato il funzionamento logico dei *nodi di rete*, e cioè i modi secondo cui l'informazione è trattata in un nodo per essere guidata verso

la destinazione desiderata. In particolare questo principio descrive le modalità fisiche o logiche adottate per attraversare i nodi e per utilizzarne la relativa capacità di elaborazione.

Infine, l'architettura protocollare definisce la stratificazione delle funzioni di trasferimento, sia nell'ambito degli apparecchi terminali che in quello delle apparecchiature di rete (nodi di accesso o di transito). In particolare questa architettura individua le funzioni che ogni nodo deve svolgere sull'informazione in esso entrante e da esso uscente. Si tratta quindi dell'organizzazione delle funzioni che sono espletate, nelle reti di accesso e di trasporto, per assicurare il trasferimento entro fissati obiettivi prestazionali.

Un modo di trasferimento è completamente definito solo se le sue tre componenti sono adeguatamente descritte. In questo senso ogni componente può essere considerata una variabile indipendente, il cui valore può essere opportunamente stabilito ai fini del conseguimento di un dato scopo (ad esempio, di fissati obiettivi prestazionali).

IV.2 La multiplazione

Per un *ramo* della rete logica, la multiplazione definisce il modo secondo cui i segmenti informativi emessi da una molteplicità di sorgenti (dette *tributarie*) condividono logicamente la capacità di trasferimento di quel ramo (*canale multiplato*).

Lo scopo principale di questa funzione e dello schema che la attua è rendere efficiente, e quindi economicamente conveniente l'uso del canale multiplato tra la postazione di utente e un nodo di accesso ovvero tra due nodi della rete di trasporto. Il conseguimento dello scopo può essere tuttavia subordinato alla esigenza di soddisfare requisiti prestazionali specifici, come, ad esempio, un elevato grado di trasparenza temporale per i servizi che lo richiedono.

Uno schema di multiplazione, in vista della realizzazione di un ambiente di comunicazione per servizi integrati, deve essere in grado di *gestire la banda* del canale multiplato con adatte strategie di assegnazione delle risorse e con opportuni criteri di risoluzione delle contese. Occorre però anche considerare le differenti modalità di multiplazione, a seconda che l'accesso al canale multiplato sia *centralizzato* o *distribuito*.

Nel caso di *accesso centralizzato*, esiste un organo (*multiplatore*), che riceve, dalle sorgenti tributarie, i segmenti informativi da multiplare e che provvede direttamente a formare il flusso multiplato (Fig. IV.2.1). Ciò avviene assegnando alle sorgenti, in modo statico o dinamico, la banda richiesta e risolvendo le eventuali situazioni di contesa. È questa la soluzione che normalmente si adotta per utilizzare la capacità di trasferimento dei rami nella rete di trasporto.

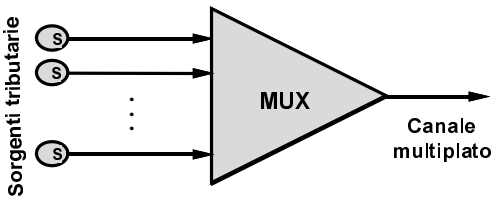


Figura IV.2.1 – Accesso centralizzato

Nel caso, invece, di accesso distribuito (chiamato solitamente *accesso multiplo*), sono le stesse sorgenti tributarie a dover gestire collettivamente la formazione del flusso multiplato. Questa gestione dovrà avvenire nel rispetto di regole di comportamento, che consentano una efficiente utilizzazione della banda disponibile e un adeguato grado di accessibilità al canale multiplato. Le regole a cui si allude sono dedicate al *controllo di accesso al mezzo* (MAC - Medium Access Control): sono cioè finalizzate allo svolgimento delle funzioni di assegnazione di banda e di risoluzione delle eventuali

situazioni di contesa, con l'obiettivo di assicurare equità di accesso per tutti i richiedenti (Fig. IV.2.2). Tali regole sono definite nei *protocolli MAC*. Nel seguito di questo paragrafo si tratterà solo l'accesso centralizzato.

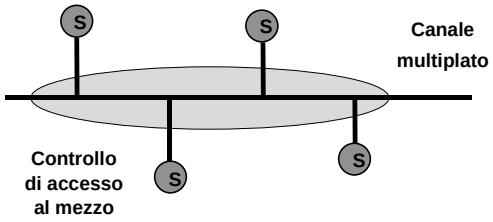


Figura IV.2.2 – Accesso multiplo

In base agli attuali orientamenti della tecnica, che adottano le tecnologie numeriche come mezzo per il trasporto delle informazioni di utente e di segnalazione, saranno qui considerati solo schemi di *multiplazione a divisione di tempo* (TDM, Time Division Multiplexing). In passato sono stati utilizzati (e in parte lo sono anche attualmente) schemi di *multiplazione a divisione di frequenza* (FDM, Frequency Division Multiplexing). Una riproposizione degli schemi FDM è oggi di piena attualità con l'impiego delle frequenze ottiche ed è denominata *multiplazione a divisione di lunghezza d'onda* (WDM, Wavelength Division Multiplexing). Sono oggi oggetto di particolare attenzione anche schemi di *multiplazione a divisione di codice* (CDM, Code Division Multiplexing).

Per sottolineare la distinzione tra TDM e FDM, è significativo fare riferimento alla occupazione del canale multiplato da parte di segmenti informativi aventi origine da una sorgente tributaria. Tale occupazione:

- nel caso *TDM*, avviene in *intervalli di tempo* che non si sovrappongono con quelli riguardanti i segmenti emessi da altre sorgenti tributarie;
- nel caso *FDM*, avviene in una *banda di frequenze* (all'interno della banda passante del canale multiplato) che è disgiunta rispetto alle bande utilizzate dai segmenti emessi da altre sorgenti tributarie.

In un *accesso centralizzato* a un ramo della rete logica, la funzione di multiplazione è svolta, in modo cooperativo, dal *multiplatore* e dal *demultiplatore* operanti alle due estremità del *canale multiplato.* Compito di questi dispositivi di rete è la condivisione del canale multiplato; più in particolare:

- il *multiplator*e deve provvedere affinché i segmenti informativi emessi da ogni sorgente tributaria accedano al canale multiplato nel rispetto della strategia di assegnazione che si è prescelta e della modalità di risoluzione delle eventuali contese di pre-assegnazione e di utilizzazione;
- il *demultiplatore* deve essere in grado di identificare i segmenti informativi che gli pervengono da ogni sorgente tributaria (funzione di *identificazione*) e riconoscere quale sia il collettore a cui ogni segmento deve essere inoltrato (funzione di *indirizzamento*).

In Fig. IV.2.3 è mostrata una possibile classificazione degli schemi TDM. I criteri seguiti per tale classificazione sono due. Il primo considera il modo secondo cui le entità emittente e ricevente trattano l'asse dei tempi nello svolgimento delle loro funzioni (*gestione dei tempi*). Il secondo riguarda le strategie di assegnazione di banda adottate dal multiplatore e i conseguenti compiti del demultiplatore (*gestione della capacità multiplata*).

Secondo il primo criterio di classificazione, uno schema TDM utilizza il flusso continuo di cifre binarie emesse dal multiplatore come un vettore di trasporto in cui i segmenti informativi provenienti da sorgenti tributarie diverse vengono caricate in un opportuno ordine temporale. Ciò equivale a dire

che la multiplazione viene effettuata seguendo un principio di separazione a divisione di tempo, che prevede la emissione di segmenti informativi diversi in intervalli di tempo distinti.

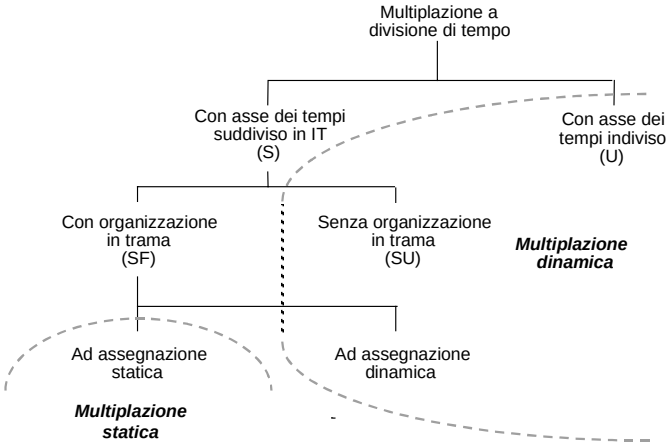


Figura IV.2.3 - Classificazione degli schemi di multiplazione a divisione di tempo.

In uno schema TDM lo svolgimento della funzione di identificazione richiede al demultiplatore, nel flusso continuo di cifre binarie ricevute, di individuare ogni singolo segmento informativo, distinguendo la stringa di cifre binarie che lo compongono dalle altre cifre binarie che la precedono e la seguono. Si tratta cioè di operare, nel processo di multiplazione, una *delimitazione*, *implicita* o *esplicita*, di ogni singolo segmento informativo, in modo che la individuazione di questo in demultiplazione possa avvenire in modo univoco. La delimitazione è *implicita* se ogni segmento informativo è identificabile solo sulla base della sua collocazione temporale all'ingresso del demultiplatore; è invece *esplicita* se ogni segmento informativo è identificabile mediante l'uso di opportune modalità di delimitazione (ad es. con l'aggiunta di delimitatori o con violazioni di codifica).

Circa la funzione di indirizzamento in uno schema TDM, occorre fare in modo che ogni singolo segmento informativo, quando perviene al demultiplatore (ad esempio, all'ingresso di un nodo di rete), possa essere inoltrato verso la corretta destinazione. Anche questa operazione di *indirizzamento* può essere effettuata in modo *implicito* o *esplicito* e deve essere armonizzata con il principio di commutazione che regola il funzionamento della rete. Un indirizzamento è *implicito* se ogni segmento informativo è indirizzabile solo sulla base della sua collocazione temporale all'ingresso del demultiplatore; è invece *esplicito* se ogni segmento informativo è indirizzabile mediante l'aggiunta di una "*etichetta"* (label), contenente l'informazione di indirizzo che consente al demultiplatore (in modo *diretto* o *indiretto*) di conoscere la destinazione e l'origine di quel segmento. L'etichetta è aggiunta nella intestazione di una unità informativa (cfr. Fig. IV.1.1).

IV.2.1 Gestione dei tempi

Circa il modo in cui multiplatore e demultiplatore trattano l'asse dei tempi, possono essere identificate due alternative di base (Fig. IV.2.4): nella prima, chiamata qui *alternativa S* (slotted), l'asse dei tempi è *suddiviso* in *intervalli temporali* (IT) tutti di ugual durata, mentre nella seconda, denominata *alternativa U* (unslotted), tale asse è *indiviso*.

L'alternativa *S* tratta segmenti informativi aventi lunghezza, che è *costante* e commisurata alla *durata di un singolo IT* (Fig. IV.2.5). Quest'ultimo è quindi utilizzato per contenere e trasportare un singolo segmento informativo, che risulta così delimitato in *modo implicito* se esiste una

sincronizzazione di IT tra multiplatore e demultiplatore. Cioè la sequenza degli istanti di inizio di ogni IT deve essere a disposizione di ambedue tali apparati terminali. Inoltre, per fronteggiare i casi di perdita di questa condizione di sincronizzazione, deve essere prevista una *procedura di recupero* veloce. Ciò ha lo scopo di ridurre nei limiti del possibile il numero segmenti informativi che, durante la condizione di fuori-sincronismo, risultano inevitabilmente perduti, in quanto non delimitabili e quindi non riconoscibili.

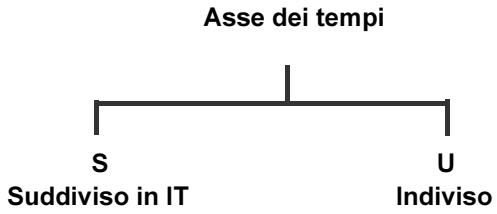


Figura IV.2.4 – Alternative S *e* U *per la gestione dei tempi*

Nell'alternativa *U*, i segmenti informativi possono essere di lunghezza variabile e vengono emessi senza vincoli sul loro istante di inizio, compatibilmente con le situazioni di contesa che si presentano nell'accesso al canale multiplato. Deve allora essere adottata una *delimitazione esplicita* dei singoli segmenti. Tale operazione può essere effettuata con gruppi di cifre binarie, chiamate *delimitatori* (flags), posti all'inizio e alla fine di ogni segmento, e distinguibili in modo univoco da parte del demultiplatore, nel senso che un delimitatore non deve poter essere simulato da analoghi gruppi di cifre binarie appartenenti ad altri segmenti (Fig. IV.2.5).

Un delimitatore non simulabile è, ad esempio, quello impiegato per delimitare le trame dei protocolli dello strato di collegamento orientati al bit, del tipo HDLC. Tale delimitatore è costituito da un ottetto di cifre binarie, comprendente uno ZERO, seguito da sei UNO e concluso da uno ZERO. Per evitare la possibilità di simulazione, nel contenuto di un segmento informativo in emissione si inserisce uno ZERO (detto di *riempimento*) dopo una sequenza di cinque UNO consecutivi. Il bit di riempimento viene rimosso in ricezione.

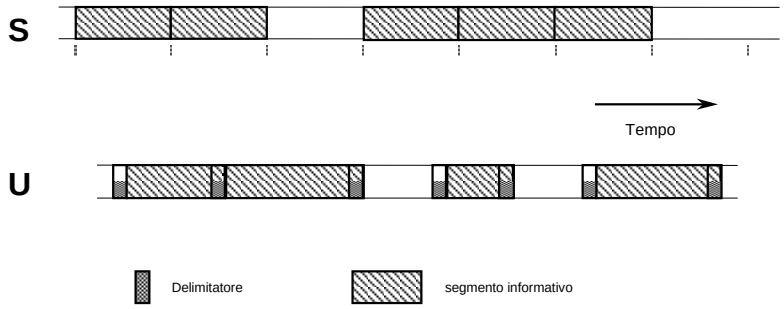


Fig. IV.2.5 – Attuazione delle alternative S *e* U.

L'alternativa *S* è realizzabile con due possibili soluzioni (Fig. IV.2.6): una *con organizzazione in trama* (framed), chiamata qui *alternativa SF* e l'altra *senza organizzazione in trama* (unframed), denominata *alternativa SU*.

Nell'alternativa *SF*, gli IT sono strutturati in *trame*, e cioè in intervalli, tutti di ugual durata, che comprendono un numero intero di IT consecutivi e che contengono, in posizione fissa (ad esempio, all'inizio della trama), anche una configurazione di cifre binarie, chiamata *parola di allineamento* (Fig. IV.2.7). È proprio questa che controlla la condizione di *sincronizzazione di trama* e che consente quindi

di ottenere implicitamente anche la sincronizzazione di IT. Per verificare la condizione di sincronizzazione di trama e per il suo recupero una volta che sia stata persa, viene seguita una opportuna strategia (*strategia di allineamento*), che consente di operare anche in presenza di interruzione del canale multiplato e di errori trasmissivi nel riconoscimento delle cifre della parola di allineamento.

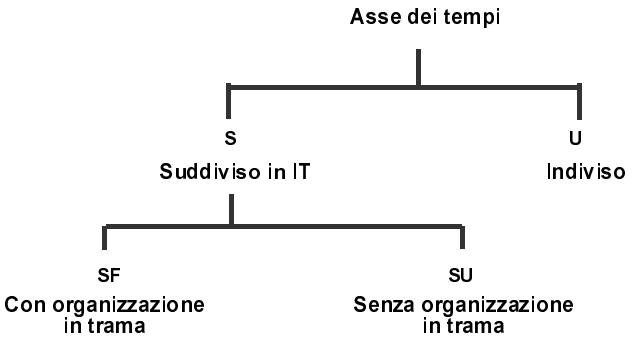


Figura IV.2.6 – Alternative SU *e* SF *per la gestione dei tempi*

Nell'alternativa *SU*, gli IT si susseguono senza alcuna struttura sovrapposta. Di conseguenza deve essere previsto un opportuno meccanismo di sincronizzazione di IT, quale quello che prevede di inserire, in IT non utilizzati (*IT vuoti*), particolari gruppi di cifre binarie, chiamati *unità di sincronizzazione* (Fig. IV.2.7). Il demultiplatore riconosce tali unità e può quindi verificare lo stato di allineamento o anche recuperarlo alla prima opportunità offerta dalla ricezione di una unità di sincronizzazione. Bisogna tuttavia osservare che il successo di questo meccanismo dipende dalla possibilità di emettere unità di sincronizzazione con frequenza media sufficientemente elevata. La tecnica è quindi tanto più efficiente quanto meno il canale multiplato è caricato. Infatti, man mano che il carico medio del canale aumenta, cresce l'intervallo medio tra due unità di sincronizzazione consecutive e, conseguentemente, aumentano la probabilità di perdere il sincronismo e il tempo di recupero dello stesso.

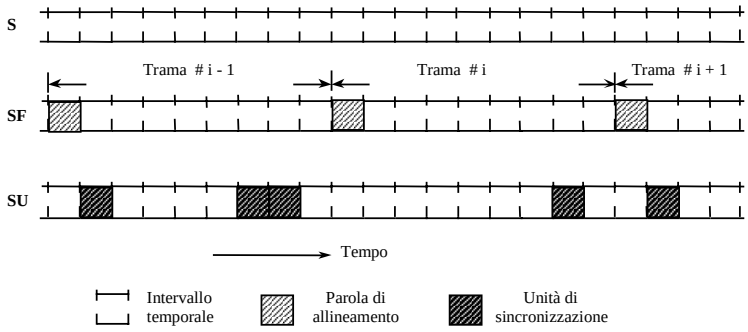


Fig. IV.2.7 - Attuazione delle alternative SU *e* SF

Un differente meccanismo di sincronizzazione di IT, che non presenta questo inconveniente, è basato su una *auto-sincronizzazione*, ottenuta, ad esempio, con una accorta utilizzazione delle intestazioni di UI.

IV.2.2 Gestione della capacità multiplata

Applicando poi il secondo criterio di classificazione utilizzato in Fig. IV.2.3, la assegnazione della capacità del canale multiplato può essere *statica* o *dinamica*. Le relative strategie di attuazione seguono le modalità introdotte in § III.2.2.

La assegnazione statica prevede che la capacità complessiva del canale sia suddivisa in *sub-canali fisici*, aventi capacità che assumono valori in un insieme discreto. All'inizializzazione di una comunicazione, un sub-canale è *pre-assegnato individualmente* alle relative necessità di trasferimento. La durata dell'assegnazione è normalmente uguale a quella della comunicazione.

Nell'assegnazione dinamica la capacità globale del canale multiplato è trattata come una risorsa singola, a cui le sorgenti tributarie possono accedere secondo le loro necessità di trasferimento e nel rispetto di una opportuna procedura di controllo. La banda è *assegnata a domanda* o *pre-assegnata collettivamente* a seconda che l'accesso sia consentito

- a qualunque sorgente tributaria presenti propri segmenti informativi al multiplatore;
- solo alle sorgenti tributarie che, presentando propri segmenti al multiplatore, siano in possesso di una *autorizzazione di accesso* ottenuta all'inizio dell'evoluzione della comunicazione.

In questo secondo caso le autorizzazioni di accesso hanno normalmente validità per l'intera durata di una comunicazione.

In base alle considerazioni svolte sui contenuti di Fig. IV.1.4, si può poi stabilire una relazione tra le tecniche di assegnazione statica o dinamica e i modi di servizio di rete in cui ciascuna di queste tecniche trova impiego (Fig. IV.2.8). Una multiplazione che opera con *assegnazione statica* è componente di un servizio di rete operante nel modo *con connessione fisica*. Invece una multiplazione che opera con assegnazione dinamica può essere componente di un servizio di rete operante in entrambi i modi con o senza connessione: il modo è *con connessione virtuale* se si adotta una strategia di *pre-assegnazione collettiva*; è invece senza connessione se la strategia adottata è quella di *assegnazione a domanda*.

Mentre un'assegnazione dinamica può essere realizzata con tutte le tre alternative SF, SU, e U per la gestione dei tempi, una assegnazione statica può essere effettuata solo se viene adottata l'alternativa SF (Fig. IV.2.9). Infatti la strutturazione in trame crea una associazione logica tra gli IT che sono collocati nella stessa posizione entro trame differenti. Ciò significa che una sequenza di IT, scelti uno per trama in modo da rispettare tale associazione logica, costituisce un *sub-canale di base*, la cui capacità C_s è facilmente valutabile. Infatti, indichiamo con L_s e L_f le lunghezze (in cifre binarie) di un IT e di una trama rispettivamente e con C_m la capacità del canale multiplato. Allora

$$C_s = C_m L_s / L_f \quad .$$

GESTIONE DELLA CAPACITA' MULTIPLATA	MODO DI SERVIZIO DI RETE
Assegnazione statica	con connessione fisica
Assegnazione dinamica	con connessione virtuale
	senza connessione

Figura IV.2.8 – Relazione tra le tecniche di gestione della capacità multiplata e i corrispondenti modi di servizio di rete

In altre parole la capacità C_m può essere considerata suddivisa in tanti sub-canali di base quanto è il numero N di IT in una trama. Si noti però che, in generale,

$$L_f > NL_s$$

in quanto una trama comprende, oltre agli N IT, anche altri gruppi di cifre binarie per scopi di servizio.

GESTIONE DELLA CAPACITA' MULTIPLATA	GESTIONE DEI TEMPI
Assegnazione statica	Modalità SF
Assegnazione dinamica	Modalità U Modalità SU Modalità SF

Figura IV.2.9 – Alternative di attuazione di uno schema di multiplazione a divisione di tempo

Nell'assegnazione statica, si possono effettuare assegnazioni *a singolo IT* o a *IT multiplo* (Fig. IV.2.10). Nel primo caso (*multiplazione di base*) i segmenti informativi generati nell'ambito di una singola comunicazione sono trasferiti su un singolo sub-canale di base. Nel secondo caso (*sovra-multiplazione*), a una comunicazione viene reso disponibile un sub-canale con capacità, che è un multiplo di quella del sub-canale di base; l'ordine di molteplicità è uguale al numero di IT assegnati in una trama e non può, ovviamente, essere maggiore di N. Ad esempio, se la capacità del sub-canale di base è di 64 kbit/s e se la capacità disponibile del canale multiplato è di 1.920 kbit/s, la sovra-multiplazione consente di affasciare cinque sub-canali di capacità uguale a 384 kbit/s (6 x 64 kbit/s) ovvero un insieme di sub-canali con questa capacità e di sub-canali di base, in numero tale che la somma delle relative capacità non sia superiore alla capacità disponibile del canale multiplato.

Per il dimensionamento della lunghezza L_s di un IT, assumiamo, che la capacità C_s del sub-canale di base sia utilizzata per il trasferimento di un flusso tributario continuo con caratteristiche CBR, risultante da un'operazione di conversione analogico-numerica di un segnale analogico (ad esempio, di voce), che è campionato con periodo T_c e che è codificato con un numero b fisso di cifre binarie per ogni campione; conseguentemente

$$C_s = b / T_c \quad .$$

In queste condizioni, è conveniente assumere la durata T_f della trama uguale a un multiplo di ordine m (m = 1,2,...) di T_c . Con questa scelta, dato che $L_s/C_s = T_f$, la lunghezza di un IT ha un valore uguale a m volte il numero b

$$L_s = mb \, .$$

Ad esempio, nella multiplazione PCM il sub-canale di base deve trasferire voce di qualità telefonica con codifica PCM; si assume allora $T_f = T_c$ = 125 μs, C_s = 64 kbit/s e b = 8 bit; conseguentemente L_s= 8 bit. Se, invece, si assume T_f = 250 μs, il ritmo binario risultante da una codifica PCM può essere trasferito con un IT di lunghezza uguale a due ottetti di cifre binarie.

Oltre alla multiplazione di base e alla sovra-multiplazione si può anche effettuare una *sotto-multiplazione*, che consente di ottenere una capacità di sub-canale, che è sottomultipla di quella del

sub-canale di base. Si possono a questo scopo adottare due modalità, che chiameremo *a trama singola* o *a multi-trama* (Fig. IV.2.10).

TIPO DI MULTIPLAZIONE	TIPO DI ASSEGNAZIONE	CAPACITA' ASSEGNATA
Multiplazione di base	a singolo IT	C_s
Sovra-multiplazione	a IT multiplo	multipla di C_s
Sotto-multiplazione		
•a trama singola	a frazione di IT	frazione di C_s
•a multitrama	a singolo IT	frazione di C_s

Figura IV.2.10 – Alternative di assegnazione statica e relative capacità di trasferimento assegnate

Nella *modalità a trama singola*, il minimo sotto-multiplo ottenibile della capacità del sub-canale di base è uguale a C_s/L_s e corrisponde ad assegnare, a un singolo flusso tributario, una sola cifra binaria tra le L_s che compongono l'IT. Conseguentemente, un singolo IT, assegnato con periodicità di trama, può essere il supporto di (al massimo) L_s sub-canali di capacità uguale a C_s/L_s ovvero di un opportuno insieme di sub-canali con capacità multiple di C_s/L_s fino a saturare la capacità C_s. Ad esempio, se C_s = 64 kbit/s e L_s = 8 bit, la minima capacità assegnabile è uguale a 8 kbit/s e il sub-canale di base può ospitare un numero di flussi tributari continui con ritmi binari di 8, 16 o 32 kbit/s, con il vincolo che la somma di questi ritmi non superi il valore di 64 kbit/s.

Per *adattare* poi un flusso tributario continuo con ritmo binario minore di C_s/L_s (o di un suo multiplo) ad una capacità uguale a C_s/L_s (o a un suo multiplo), si può ricorrere al *riempimento* di tale flusso con cifre addizionali e alla *ripetizione* di ogni cifra binaria del tributario. Un flusso continuo con ritmo di 4,8 kbit/s può, ad esempio, essere adattato ad una capacità di 8 kbit/s, riempiendo con 32 cifre binarie addizionali (con scopi di servizio) una sequenza di 48 cifre binarie del tributario; in tal modo la capacità di 8 kbit/s viene utilizzata per 48/80 della sua piena potenzialità, secondo quanto necessario. Invece, un corrispondente adattamento di flussi tributari continui con ritmi di 0,6 o 1,2 o 2,4 kbit/s richiede, oltre al provvedimento precedente, anche la ripetizione di ogni cifra binaria di questi tributari per 8 o 4 o 2 volte, rispettivamente. Ciò però comporta, per la banda disponibile, un rendimento di utilizzazione che è tanto più basso quanto minore è il ritmo binario da adattare.

Per ciò che riguarda poi la *modalità a multi-trama*, se si richiede che il minimo sotto-multiplo della capacità C_s sia uguale a C_s/K (K = 2, 3...), occorre definire una multi-trama di lunghezza (in cifre binarie) uguale a KL_f e operare su questa come nella multiplazione di base o nella sovra-multiplazione.

Osserviamo che l'ordine K di sotto-multiplazione non può essere aumentato al di sopra di un opportuno limite, che è legato all'esigenza di contenere il *ritardo di riempimento D* di un IT nel caso di un flusso tributario continuo. Tale ritardo, che è il tempo necessario a riempire un IT con un flusso CBR con ritmo binario uguale a C_s/K, è infatti uguale a

$$D = KL_s / C_s \ .$$

Esso deve essere limitato, in quanto costituisce una componente additiva del ritardo di trasferimento nell'operazione di trasferimento, di un flusso tributario continuo. Supponiamo che detto flusso risulti da un'operazione di conversione analogico-numerica di un segnale analogico, che è campionato con un periodo T_c uguale alla durata T_f della trama di base (cioè $T_c = L_s/C_s$) e che è codificato con $b=L_s/K$ cifre binarie per campione. Ad esempio se la capacità C_s è di 2,048 Mbit/s con una durata di trama di base uguale a 125 μsec, la lunghezza L_s di un IT è uguale a 256 bit. In queste condizioni, se il tributario CBR a minor ritmo binario fosse l'uscita di un codificatore PCM e richiedesse quindi una capacità di 64 kbit/s (cioè 1/32 di C_s), la multi-trama di multiplazione dovrebbe avere una lunghezza uguale a 32 trame di base e il ritardo di riempimento D sarebbe uguale a 32×125 μs =4ms.

In conclusione, l'assegnazione statica riguarda i sub-canali nei quali è stata suddivisa la capacità del canale multiplato, e cioè uno o più tra

- i sub-canali di base, ognuno con capacità C_s (*multiplazione di base*);
- i sub-canali con capacità multipla di quella C_s (*sovra-multiplazione*);
- i sub-canali con capacità che è una frazione di C_s (*sotto-multiplazione*).

Poiché ogni sub-canale fisico, una volta attribuito al servizio di una comunicazione, diventa componente di una connessione tra una origine e una destinazione, la durata della pre-assegnazione è uguale a quella della connessione di cui il sub-canale è divenuto componente.

Passando poi all'assegnazione dinamica associata ad un servizio di rete con connessione, la relativa operazione di pre-assegnazione riguarda i cosiddetti *canali logici* . Questi rappresentano le risorse virtuali viste dalle comunicazioni concorrenti come *immagine* del canale multiplato: il condizionamento di ogni comunicazione è determinato dalle sue esigenze di trasferimento (flusso intermittente o costante di tipo CBR o VBR) e di prestazione (trasparenza, integrità, ecc.). La durata della pre-assegnazione è uguale a quella della connessione virtuale di cui il canale logico è componente.

Nel caso infine di assegnazione dinamica, associata a un servizio di rete senza connessione, le singole comunicazioni non dispongono di canali fisici o logici per trasferire il loro flusso informativo in modo separato da altri flussi. La separazione di segmenti informativi emessi e trasferiti nell'ambito di comunicazioni diverse è attuabile solo attraverso l'identificazione e l'indirizzamento dell'informazione in essi contenuta.

In relazione alle due modalità di gestione della capacità multiplata si distinguono i due principali tipi di multiplazione: la multiplazione *statica* e quella *dinamica*. Queste due tecniche sono oggetto della trattazione che segue.

Prima di procedere poniamo però a confronto la definizione di funzione di multiplazione che è stata data in questo paragrafo con quella che è stata fornita in § II.5.2 relativamente ad una *multiplazione di strato*. Concentriamo l'attenzione su una multiplazione statica o su una dinamica con canali logici. Dato che esiste una evidente corrispondenza tra le connessioni di strato da un lato e i sub-canali fisici o i canali logici che sono gestiti nelle assegnazioni statica o dinamica, si può riconoscere che in entrambe le definizioni la multiplazione è la funzione preposta a trasferire flussi informativi distinti su altrettante connessioni e con il supporto di una singola risorsa rappresentata dal canale multiplato. È altrettanto evidente che una multiplazione dinamica operante in assenza di connessioni non rientra in questa definizione.

IV.2.3 Multiplazione statica

Dato l'uso di una gestione statica delle risorse, questa tecnica di multiplazione, in presenza di concorrenza di richieste, comporta solo *contese di pre-assegnazione.* Una contesa di questo tipo si presenta quando, a seguito di uno stato di occupazione del canale multiplato, non è disponibile il sub-

canale fisico che viene richiesto da uno degli utenti del servizio. La risoluzione di queste contese, nel trattamento delle richieste che non possono essere subito soddisfatte, può essere effettuata con meccanismi *a perdita* o *a ritardo* a seconda che le richieste di assegnazione non soddisfacibili vengano rifiutate o vengano poste in uno stato di attesa (cfr. § III.1.2). In questo secondo caso il trattamento è sempre di tipo misto a perdita e a ritardo, dato che è necessario introdurre una limitazione al numero di richieste in attesa.

Poiché in una multiplazione statica la gestione dei tempi è secondo la modalità SF, multiplatore e demultiplatore debbono operare in condizioni di *sincronizzazione di trama*, cioè gli orologi delle due apparecchiature debbono avere uguale frequenza di bit e debbono essere in opportuna relazione di fase, tale da assicurare l'individuazione dell'inizio di ogni trama.

La condizione di sincronizzazione di trama rende impliciti l'*indirizzamento* e la *delimitazione* dei segmenti informativi trasferiti. Infatti i segmenti emessi da una sorgente tributaria sono contenuti in posizioni di trama, che sono definite dal multiplatore (in accordo con il demultiplatore) quando viene attuata l'assegnazione a favore della comunicazione in cui opera quella sorgente tributaria: i segmenti informativi sono quindi implicitamente identificati e quindi indirizzati per mezzo di quelle posizioni di trama. Non è quindi necessario un indirizzamento esplicito e ciò giustifica la denominazione di *multiplazione senza etichetta* attribuita alla multiplazione statica. Tuttavia l'associazione logica tra una comunicazione e una posizione di trama richiede che, durante la fase di assegnazione, tale associazione (decisa dal multiplatore) sia notificata al demultiplatore tramite *informazione di segnalazione*. Circa la delimitazione, è ancora la condizione di sincronizzazione di trama ad assicurare in modo implicito la delimitazione di ogni segmento informativo.

Nella multiplazione statica, i segmenti informativi emessi da una singola sorgente tributaria sono trasferiti come una *sequenza periodica*. Ogni periodo è uguale alla durata della trama (o della multi-trama) e può includere uno o più segmenti in accordo al tipo di sub-canale che viene pre-assegnato alla comunicazione considerata. In relazione a questa periodicità di trasferimento, la multiplazione statica è anche chiamata *multiplazione a divisione di tempo sincrona* (S-TDM, Synchronous TDM).

Per effetto di una multiplazione statica si ottengono le seguenti *caratteristiche prestazionali*:

- il grado di *trasparenza temporale* non subisce variazioni legate all'operazione di multiplazione;
- il grado di *integrità informativa* rimane anch'esso inalterato;
- il grado di *flessibilità di accesso*, nel caso di multiplazione di base, è limitato dalla granulosità della capacità C_s e, nel caso di sovra- o di sotto-multiplazione, è limitato dalla complessità del controllo.

IV.2.4 Multiplazione dinamica

In questa tecnica di multiplazione l'*indirizzamento* è *sempre esplicito*, mentre la *delimitazione* dipende dalla modalità di gestione dei tempi: è *esplicita* nella modalità U, mentre è *implicita* quando si adotta una delle modalità SU e SF. Circa l'uso di indirizzi espliciti, è da rilevare come questa è l'unica possibilità a disposizione del demultiplatore ai fini di un corretto inoltro dell'informazione ricevuta. L'uso di un indirizzamento esplicito giustifica la denominazione di *multiplazione con etichetta* e implica che l'informazione trattata sia sempre strutturata in *Unità Informative* (UI) (cfr. Fig. IV.1.1). La delimitazione implicita nelle modalità SU e SF è poi possibile se il demultiplatore opera in condizione di *sincronizzazione di IT* con il multiplatore; questa condizione, nella modalità SF, può essere conseguita anche con una *sincronizzazione di trama*.

Dato l'uso di una assegnazione dinamica delle risorse, questa modalità di multiplazione comporta in ogni caso *contese di utilizzazione* . Si hanno anche contese di pre-assegnazione nei casi in cui il servizio di rete che si realizza è con connessione.

Contese di utilizzazione si presentano quando UI consegnate al multiplatore non possono essere inoltrate senza ritardo verso il canale multiplato, in quanto questo è già impegnato nel trasferimento di altre UI. La risoluzione di queste contese può avvenire con le modalità già illustrate in § III.1.3. In un trattamento *a perdita*, le UI sono scartate. In un trattamento *a ritardo*, le UI sono memorizzate in un *buffer di multiplazione* ove sostano fino alla loro emissione. Si nota che l'utilizzazione di un buffer di capienza necessariamente finita fa sì che, in generale, nella risoluzione di contese di utilizzazione si possano presentare entrambi gli eventi di rifiuto e di ritardo: si ha trattamento *puramente a ritardo* solo nel caso ideale di buffer di multiplazione avente capienza illimitata; si ha invece trattamento *puramente a perdita* quando non esiste un buffer di multiplazione.

Contese di pre-assegnazione si presentano quando, a seguito dello stato di carico del canale multiplato, non è possibile soddisfare subito la richiesta di assegnazione di un canale logico. Il criterio per pre-assegnare un canale logico o per negarne la assegnabilità segue le linee-guida della regola di decisione illustrata in § III.2.2; su questo tema di tornerà in § IV.7.3 con riferimento ad una comunicazione inizializzata su base chiamata. Per ciò che riguarda poi la risoluzione di contese di pre-assegnazione in questo contesto, valgono ancora le modalità illustrate in generale in § III.1.2: anche in questo caso specifico possono essere considerati meccanismi *a perdita* o *a ritardo,* in cui la prevalenza degli eventi di rifiuto o di ritardo dipende dal numero delle richieste non soddisfacibili che possono essere poste in uno stato di attesa.

Nella multiplazione dinamica l'intervallo temporale che si può osservare sul canale multiplato tra due UI consecutive, emesse da una singola sorgente e trasferite nell'ambito della stessa comunicazione, è variabile. Tale intervallo è unicamente legato alle caratteristiche di emissione della associata sorgente tributaria e alle condizioni di carico del canale multiplato. In accordo a tale caratteristica, questo schema è anche chiamato *multiplazione a divisione di tempo asincrona* (A-TDM, Asynchronous TDM).

Da un punto di vista prestazionale una multiplazione dinamica presenta caratteristiche dipendenti dalle modalità di risoluzione delle contese di utilizzazione. Se le contese di utilizzazione sono risolte con meccanismo *puramente a ritardo*:

- il grado di *trasparenza temporale* subisce un peggioramento, in quanto il tempo di sosta di una UI nel buffer di multiplazione è una quantità variabile aleatoriamente;
- il grado di *integrità informativa* non subisce variazioni legate all'operazione di multiplazione.

Se le contese di utilizzazione sono invece risolte con meccanismo *puramente a perdita*:

- il grado di *trasparenza temporale* non subisce deterioramenti;
- il grado di *integrità informativa* subisce un *peggioramento*, in quanto si scartano le UI che incontrano contesa.

Indipendentemente dalle modalità di risoluzione delle contese di utilizzazione, il *grado di flessibilità di accesso* è il *massimo possibile*, dato che la capacità di trasferimento del canale multiplato è utilizzata, di volta in volta, secondo le necessità delle sorgenti tributarie.

IV.2.5 Multiplazione ibrida

Per completare il quadro delle tecniche di multiplazione a divisione di tempo, deve essere menzionata una ulteriore alternativa, che è chiamata *multiplazione ibrida*.

Questa utilizza una struttura a trama, divisa in due *regioni*: una sempre suddivisa in IT, sulla quale si applicano schemi di multiplazione statica; l'altra, che può essere suddivisa in IT oppure indivisa, la cui capacità può essere utilizzata applicando schemi di multiplazione dinamica. La *frontiera* tra le due parti della trama può essere *fissa* o *mobile*; il caso di frontiera mobile si applica per adattare le capacità delle due regioni alle richieste di trattamenti secondo le due tecniche S-TDM e A-TDM.

IV.3 La commutazione

Per un nodo della rete logica, la commutazione definisce il modo secondo cui un qualunque ingresso del nodo (*ramo di ingresso*) viene *associato logicamente* con una qualunque uscita (*ramo di uscita*). Lo scopo è attuare uno scambio, tra ingresso e uscita del nodo, operato sul flusso di informazione che perviene al nodo nell'ambito dell'espletamento di un servizio di rete. La definizione riguarda comunicazioni *punto-punto*, ma può agevolmente essere generalizzata al caso di comunicazioni *multipunto*. Una commutazione è attuata per mezzo delle funzioni di *instradamento* e di *attraversamento*.

L'*instradamento* è la funzione *decisionale*, svolta da un nodo della rete logica con lo scopo di stabilire il ramo di uscita verso cui deve essere inoltrato un segmento informativo che perviene da un ramo d'ingresso. Perciò la scelta risultante dalla funzione di instradamento è il mezzo che consente ai nodi di rete di guidare i segmenti informativi verso la loro corretta destinazione. Tale scelta non è in generale univoca e deve quindi essere operata in base a opportune regole (*algoritmi di instradamento*). Per fare in modo che la scelta rispetti i requisiti di qualità del servizio, gli algoritmi di instradamento più evoluti prevedono che questa funzione sia espletata da ogni nodo in modo coordinato con gli altri nodi della rete. In tal modo si può tenere conto di possibili guasti di nodo o di ramo e modificare le scelte quando certe parti di rete entrano in uno stato di congestione.

In un nodo della rete logica, l'*attraversamento* è la funzione *attuativa*, che ha lo scopo di trasferire, attraverso quel nodo, un segmento informativo da un ramo d'ingresso ad uno di uscita. L'attraversamento pertanto realizza quanto deciso dalla funzione di instradamento e perciò può attuarsi solo se la funzione di instradamento è stata già espletata. La funzione di attraversamento si effettua, individualmente, su ogni segmento informativo che perviene al nodo. D'altra parte ogni segmento entra nel nodo e ne esce all'interno di flussi multiplati su un ramo entrante e su uno uscente, rispettivamente. Conseguentemente, prima di svolgere la funzione di attraversamento, occorre, per ogni ramo entrante, *demultiplare il relativo flusso*, estraendo cosi' i segmenti che quel ramo trasferisce al nodo. Dopo aver svolto la funzione di attraversamento, i segmenti inoltrati verso il pertinente ramo uscente debbono essere *multiplati per costituire il flusso uscente* su questo secondo ramo.

L'attraversamento di un nodo può avvenire con due differenti modi:

- con il modo *diretto*, in cui i flussi informativi all'ingresso e all'uscita del nodo sono *multiplati staticamente* e in cui il percorso interno ingresso-uscita è *temporalmente trasparente*;
- con il modo *ad immagazzinamento* e *rilancio*, in cui i flussi informativi all'ingresso e all'uscita del nodo sono *multiplati dinamicamente* e in cui ogni segmento informativo attraversante il nodo viene *memorizzato*, completamente o parzialmente, prima di essere rilanciato verso l'uscita.

Entrambi i modi di attraversamento operano su una informazione che deve avere *formati coerenti* con quelli degli schemi di multiplazione agenti sui rami di ingresso e di uscita. Conseguentemente:

- un nodo che attua un attraversamento nel *modo diretto* e che quindi mette in corrispondenza uno schema di multiplazione statica in ingresso con uno analogo in uscita trasferisce i segmenti informativi associati ad una comunicazione con un *indirizzo* che è *implicito*; oggetto della funzione di scambio operata dal nodo è quindi un RCB;
- un nodo che attua un attraversamento nel *modo ad immagazzinamento e rilancio* e che quindi mette in corrispondenza uno schema di multiplazione dinamica in ingresso con uno analogo in uscita trasferisce i segmenti informativi associati ad una comunicazione con un *indirizzo* che è *esplicito (etichetta)*; oggetto della funzione di scambio operata dal nodo è quindi una *unità informativa.*

Da un punto di vista prestazionale, i due modi di attraversamento si distinguono in base al *ritardo di attraversamento* che viene imposto ai segmenti informativi che transitano attraverso un nodo di rete da un suo ramo di ingresso a uno di uscita. Tale ritardo coincide con l'intervallo di tempo che intercorre tra l'istante in cui un segmento entra completamente nel nodo e quello in cui inizia ad uscirne: la definizione è illustrata graficamente in Fig.IV.3.1 con riferimento ad un segmento informativo *IS*. Nel caso del modo *diretto*, il ritardo di attraversamento è di valore costante per tutti i segmenti originati dallo stesso utente. Nel caso invece del modo *ad immagazzinamento e rilancio*, tale ritardo assume valori che variano aleatoriamente per ogni segmento.

A chiarimento di queste ultime affermazioni:

- nel modo diretto si stabilisce tra l'ingresso e l'uscita del nodo una connessione fisica che introduce un *ritardo costante*, con un valore principalmente dipendente dalla tecnica (divisione di spazio o di tempo) utilizzata per la realizzazione di questa connessione;
- nel modo ad immagazzinamento e rilancio, ogni UI entrante nel nodo deve essere memorizzata per poterne elaborare l'intestazione, in particolare per leggerne l'etichetta e per poter quindi decidere quale sia il ramo di uscita verso cui inoltrarla; se poi, come normalmente si opera, le contese di utilizzazione nell'accesso alla multiplazione in uscita sono risolte a ritardo, interviene un secondo motivo di memorizzazione, attuata a monte di ogni ramo di uscita dal nodo.

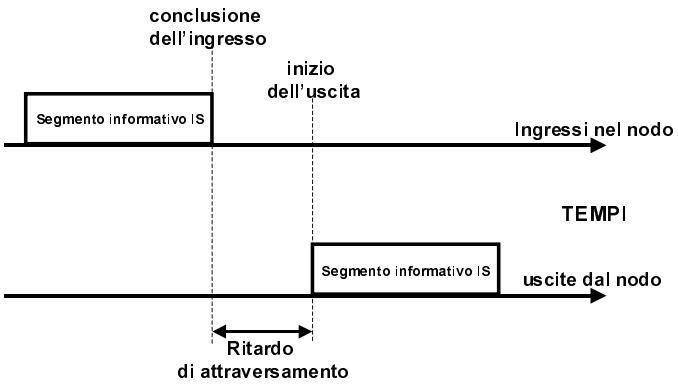


Figura IV.3.1 – Per la definizione di ritardo di attraversamento in un nodo

Pertanto, se l'attraversamento è attuato nel modo ad immagazzinamento e rilancio con le contese di utilizzazione in uscita risolte a ritardo, il ritardo di attraversamento è costituito da due componenti

- una è *costante*, in quanto data dalla somma della durata di elaborazione di una UI e del ritardo di propagazione (sempre trascurabile) tra l'ingresso e l'uscita del nodo;
- l'altra è *variabile*, in quanto determinata dalla durata di memorizzazione di ogni UI all'interno del nodo.

Conseguentemente, nelle condizioni ora dette, il ritardo di attraversamento di un nodo dipende dal carico richiesto sia all'unità di elaborazione del nodo, che al multiplatore dinamico di uscita. Poiché questo carico è variabile aleatoriamente, tale risulta anche il ritardo di attraversamento.

Attraversamenti nel modo ad immagazzinamento e rilancio sono normalmente attuati con *memorizzazioni complete* delle UI in ogni nodo prima che questo proceda a un loro rilancio verso il nodo successivo. Con lo scopo di ridurre il ritardo di attraversamento si possono anche effettuare *memorizzazioni parziali* di ogni UI. Ciò può riguardare sia l'elaborazione della UI, sia la risoluzione delle contese per l'accesso al ramo di uscita. Per la prima funzione è sufficiente memorizzare la sola intestazione purché la capacità elaborativa del nodo sia così elevata (rispetto alla capacità di

trasferimento del ramo uscente) da concludere il trattamento prima che inizi l'arrivo della parte residua della UI.

Per la memorizzazione nei buffer di multiplazione si può procedere in modo differenti a seconda dello stato di occupazione del ramo di uscita. Se questo è libero, l'intestazione è inoltrata senza ritardi verso il nodo successivo anche se il resto della UI non ha ancora raggiunto il nodo in esame; occorre però provvedere affinchè l'intestazione, per ogni ramo attraversato, prenoti il trasferimento della restante parte di UI in modo da conservare l'unitarietà dell'unità informativa. Se invece il ramo di uscita è occupato, la memorizzazione è inevitabilmente completa. Questo metodo, noto come "cut-through" e applicabile in servizi di rete con o senza connessione, tende ad avvicinare il modo ad immagazzinamento e rilancio a quello diretto, emulando da quest'ultimo un elevato grado di trasparenza temporale. Il presupposto è però che i nodi operino in condizioni di basso carico.

Le possibili alternative per attuare la funzione di commutazione sono descritte nel diagramma ad albero di Fig. IV.3.2, ove si assumono come riferimenti: i possibili servizi di rete in accordo alle relazioni tra le parti in comunicazione e le possibili strategie di assegnazione delle risorse. Tra queste ultime si distinguono due tecniche, quella *a circuito* e quella *a pacchetto*. Alla presentazione di queste due tecniche è dedicata la trattazione in § IV.3.1 e IV.3.2.

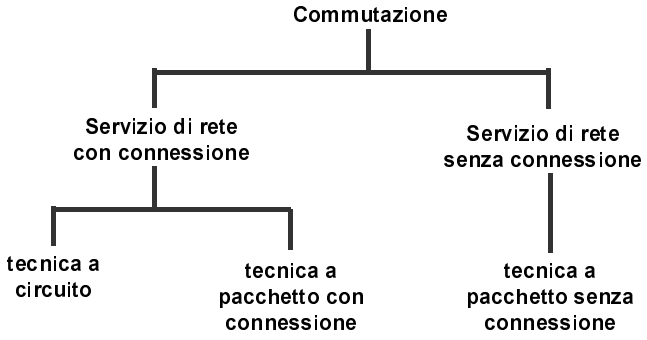


Figura IV.3.2 – Alternative per attuare la funzione di commutazione

Gli stessi contenuti di Fig. IV.3.2 sono riportati anche in Fig. IV.3.3, ove per ogni servizio di rete e per ogni tecnica di assegnazione delle risorse sono precisati il *modo di attraversamento* e lo *schema di multiplazione* utilizzato sui rami all'ingresso e all'uscita del nodo.

SERVIZIO DI RETE	ASSEGNAZIONE DELLE RISORSE	MODO DI ATTRAVERSAMENTO	MULTIPLAZIONE
con connessione	tecnica a circuito	diretto	statica
con connessione	tecnica a pacchetto con connessione	a immagazzinamento e rilancio	dinamica
senza connessione	tecnica a pacchetto senza connessione	a immagazzinamento e rilancio	dinamica

Figura IV.3.3 – Tecniche di attraversamento in relazione con i possibili servizi di rete e con le strategie di assegnazione delle risorse

IV.3.1 Tecnica a circuito

Si opera con tecnica a circuito quando si vuole realizzare un servizio di rete con connessione fisica e quindi con piena trasparenza temporale. Con questa tecnica, in ogni nodo di rete viene gestita una multiplazione statica sui rami entranti o uscenti e viene attuato un attraversamento con il modo diretto. In altre parole caratteristica distintiva della tecnica a circuito è la *gestione statica* delle risorse di trasferimento disponibili da parte di ogni organo di rete.

Nella tecnica a circuito ogni nodo, nella fornitura di un servizio di rete, impegna/rilascia risorse di trasferimento al suo interno e sui rami uscenti. Lo scopo è concorrere con altri nodi della rete a instaurare/abbattere un *percorso* di rete tra sorgente e collettore di una comunicazione. La durata di questo impegno è quella del servizio. La tecnica a circuito è quindi l'applicazione di una pre-assegnazione individuale con durata uguale alla vita della connessione posta al servizio di ogni comunicazione. In questa applicazione un nodo di rete prende la decisione relativa all'instradamento solo in corrispondenza dell'instaurazione del percorso che lo coinvolge.

Nella tecnica a circuito la concorrenza delle richieste di servizio determina contese che possono essere solo di *pre-assegnazione*. Queste contese si verificano ogni qualvolta l'impegno delle risorse di trasferimento *interne* ed *esterne* a uno o più nodi coinvolti nel servizio impedisce ulteriori assegnazioni. Al riguardo si sottolinea che un nodo può rendere disponibile una connessione tra un suo ramo di ingresso e uno di uscita solo se sono trovati contemporaneamente liberi un percorso interno ingresso-uscita e una via di trasferimento sul ramo uscente.

La risoluzione di queste situazioni è stata già considerata per le risorse esterne quando si è trattata la multiplazione statica. Per le risorse interne, la struttura di nodo che realizza la funzione di attraversamento e che è chiamata *rete di connessione* può trovarsi in stati di occupazione che impediscono l'instaurazione di un ulteriore percorso interno ingresso-uscita in aggiunta ai percorsi già instaurati e indipendentemente dallo stato di occupazione delle uscite. Anche per le contese relative all'impegno delle risorse interne possono essere utilizzati criteri di risoluzione analoghi a quelli per le risorse esterne, con i quali peraltro debbono essere convenientemente armonizzati.

Come avviene per ogni servizio di rete con connessione, i percorsi tra sorgenti e collettori possono essere in alternativa dei tipi *commutato* e *semipermanente.* Nel caso di percorso commutato, l'instaurazione e l'abbattimento della connessione avvengono a seguito di scambi di informazioni di segnalazione tra i nodi che sono attraversati dal percorso e con ritardi di attuazione il più possibile contenuti rispetto all'evoluzione della comunicazione che ne fa richiesta. Ogni comunicazione deve essere quindi trattata "*in tempo reale*". Una situazione tipica è quella delle comunicazioni inizializzate su base chiamata e con connessione non modificata nel corso della comunicazione: in questo caso il percorso è instaurato quando la chiamata ha inizio e viene abbattuto quando la chiamata si conclude.

Pertanto il nodo che gestisce connessioni commutate e che applica la tecnica a circuito deve essere in grado di trattare l'informazione di segnalazione e deve essere dotato di *risorse di elaborazione* in grado di svolgere la funzione di instradamento: tali risorse sono coinvolte solo durante le fasi di instaurazione/abbattimento di ogni connessione, mentre rimangono praticamente inoperose durante la fase di trasferimento dell'informazione.

Invece nei nodi che gestiscono connessioni semi-permanenti, l'instaurazione/abbattimento di un percorso è un provvedimento di natura gestionale, senza quindi scambi di informazione di segnalazione e senza esigenza di trattamenti in tempo reale.

La commutazione che impiega la tecnica a circuito come strategia di assegnazione delle risorse di trasferimento è detta "*a circuito*".

IV.3.2 Tecnica a pacchetto

La tecnica a pacchetto può avere due diverse attuazioni:

- una prima consente di realizzare servizi di rete *senza connessione*;
- una seconda ha come finalità servizi di rete *con connessione*, ma con la specificità (rispetto alla tecnica a circuito) che si tratta di connessione avente natura esclusivamente logica.

Secondo questa tecnica e con riferimento alle risorse di trasferimento, in un nodo di rete viene gestita una multiplazione dinamica sui rami entranti o uscenti e viene attuato un attraversamento con il modo ad immagazzinamento e rilancio. Ne consegue la *gestione dinamica* delle risorse di trasferimento che sono a disposizione di ogni organo di rete. Inoltre le risorse di elaborazione in ogni nodo sono sempre coinvolte nella fase di trasferimento dell'informazione e possono esserlo anche nelle fasi di instaurazione/abbattimento quando il nodo operi nell'ambito di un servizio con connessione.

In presenza di una concorrenza di richieste di servizio, questa gestione dinamica implica la possibile presentazione di *contese di utilizzazione* qualunque sia il servizio di rete che viene attuato. Oggetti di condivisione in ogni nodo sono il processore per l'elaborazione delle UI e l'accesso a ciascuno dei rami di uscita. Se questo accesso avviene attraverso un buffer e se si tiene conto della memorizzazione a monte della elaborazione, le contese di utilizzazione sono risolte prevalentemente a ritardo. Inoltre il parametro prestazionale che descrive in ogni nodo gli eventi di contesa è il ritardo di attraversamento, che subisce ciascuna UI transitante attraverso il nodo. Il valor medio di questo ritardo dipende ovviamente dalla capacità di elaborazione del processore di nodo e dalla capacità di trasferimento dei rami di uscita. Nei nodi a pacchetto di ormai passata generazione era la capacità di trasferimento dei rami a costituire il "collo di bottiglia" maggiormente condizionante. Nelle realizzazioni più recenti il deciso aumento dei ritmi di trasferimento sulla rete fisica ha consentito di ridurre fortemente i valori medi del ritardo di attraversamento.

Nel caso di servizio con connessione, ad ogni ramo uscente da un nodo sono associabili i *canali logici* della multiplazione dinamica che agisce su quel ramo. Il nodo deve provvedere ad assegnare tali canali quando ne viene fatta richiesta da parte di una comunicazione. Il criterio adottato segue le già citate regole di accettazione (cfr. § IV.7.3). La concatenazione di canali logici lungo il percorso di rete tra sorgente e collettore di una comunicazione dà luogo ad una connessione logica che è detta *circuito virtuale*. Per il caso di connessione commutata o semi-permanente, la decisione relativa all'instradamento è presa solo quando è richiesta l'instaurazione della connessione; durante la fase di trasferimento dell'informazione, lo scambio delle UI avviene sul circuito virtuale messogli a disposizione e, per questo scopo, utilizza la memoria di quanto deciso nella fase di instaurazione. Questa attuazione della tecnica a pacchetto, basata sull'assegnazione di risorse virtuali, è quindi l'applicazione di una pre-assegnazione collettiva con durata uguale alla vita della connessione posta al servizio di una comunicazione; verrà pertanto indicata nel seguito come *"tecnica a pacchetto con connessione*".

Un servizio di rete senza connessione può invece essere realizzato con una assegnazione a domanda delle risorse necessarie. Ogni UI viene trattata in modo indipendente dalle altre UI che la precedono o la seguono nell'attraversamento del nodo. In particolare la decisione di instradamento è presa "*UI per UI*". In tal modo due UI successive, anche se scambiate nell'ambito della stessa comunicazione possono essere trasferite lungo percorsi di rete diversi. Questo tipo di assegnazione verrà chiamato "*tecnica a pacchetto senza connessione*".

I contenuti dell'etichetta contenuta nelle UI uscenti da un nodo dipendono dal tipo di servizio di rete. Se questo è senza connessione, l'etichetta deve contenere gli indirizzi della destinazione e dell'origine. Se invece il servizio è con connessione e se si considerano le esigenze di indirizzamento nelle fasi di trasferimento e di abbattimento, è sufficiente, come si è visto in § II.5.1, utilizzare un *identificatore di connessione*, dato che la connessione individua univocamente l'origine e la

destinazione dell'informazione uscente dal nodo. Con riferimento ad una specifica comunicazione, i flussi entranti in un nodo o uscenti da questo sono allora identificati da *due etichette*: una, per il flusso entrante, identifica il canale logico sul ramo di ingresso, l'altra, per il flusso uscente, è invece identificatore del canale logico sul ramo di uscita. Ciò implica che, nell'attraversamento del nodo, occorre trasformare l'etichetta che identifica il flusso entrante in quella che identifica il flusso uscente: questa operazione è denominata *"traduzione di etichetta*".

La corrispondenza tra le due etichette è stabilita in sede di instaurazione della connessione e conservata in una memoria, che è chiamata *tabella di attraversamento*. La memorizzazione ha la durata del servizio. La tabella è consultata ogni qualvolta una UI entra nel nodo e la corrispondenza ivi disponibile è utilizzata per effettuare la traduzione di etichetta.

Rimanendo nel caso di nodi che attuano un servizio di rete con connessione, la concorrenza delle richieste di instaurazione determina anche *contese di pre-assegnazione*, che si presentano quando lo stato di impegno delle risorse interne ed esterne in uno o più nodi coinvolti nel servizio impedisce ulteriori instaurazioni di circuito virtuale. Le risorse di cui si parla hanno significatività puramente virtuale: le risorse esterne sono rappresentate dai canali logici dello schema di multiplazione dinamica, mentre quelle interne sono identificabili nelle potenzialità di relazione ingresso-uscita residenti nella tabella di attraversamento. Con questo elemento distintivo sulla significatività delle risorse, per le contese di pre-assegnazione nel caso di tecnica a pacchetto possono ripetersi le stesse considerazioni svolte a proposito delle contese di pre-assegnazione nella tecnica a circuito; ciò vale in particolare anche per le relative modalità di risoluzione.

La commutazione che impiega la tecnica a pacchetto come strategia di assegnazione delle risorse di trasferimento e di elaborazione è detta "*a pacchetto*" con o senza connessione, a seconda del modo di servizio di rete che si intende realizzare.

IV.3.3 Ritardi nella tecnica a pacchetto

Svolgiamo alcune considerazioni sul *ritardo di trasferimento* che deve subire un flusso informativo *intermittente* o *continuo* nel suo transito attraverso una rete i cui nodi operano secondo la tecnica a pacchetto.

Cominciamo considerando il caso di un *flusso intermittente* e determiniamo il ritardo di trasferimento *D* di una stringa di cifre binarie, che chiameremo "*messaggio*". Il ritardo *D* è, per definizione, l'intervallo temporale tra l'emissione del primo bit del messaggio e la ricezione dell'ultimo bit, come viene illustrato in Fig. IV.3.4 che fa riferimento ad un messaggio *X*. Il trasferimento del messaggio o di una sua parte strutturata avviene attraverso una successione di *salti* che consentono al messaggio o a un suo segmento di passare da un nodo a quello successivo dall'origine alla destinazione. Conseguentemente, nell'ipotesi di un servizio di rete con connessione, il ritardo *D* dipende da:

- le *capacità di trasferimento* dei rami di rete attraversati dal circuito virtuale;
- i *ritardi di propagazione* su questi rami;
- le *lunghezze* (in cifre binarie) del messaggio e delle eventuali parti in cui il messaggio viene segmentato;
- il numero di *salti intermedi*;
- il *ritardo di attraversamento* di ogni nodo coinvolto nel trasferimento.

Ogni salto contribuisce alla determinazione del ritardo di trasferimento di un segmento informativo con la somma del *ritardo di attraversamento* del nodo a monte e del ritardo di transito subito dal segmento sul ramo uscente da questo nodo.

Il primo contributo ha le caratteristiche di variabilità precedentemente sottolineate; il secondo è la somma del *ritardo di propagazione* e del *tempo di emissione* del segmento sul ramo in questione.

Quest'ultimo contributo è uguale al rapporto tra la lunghezza del segmento e la capacità di trasferimento del ramo considerato.

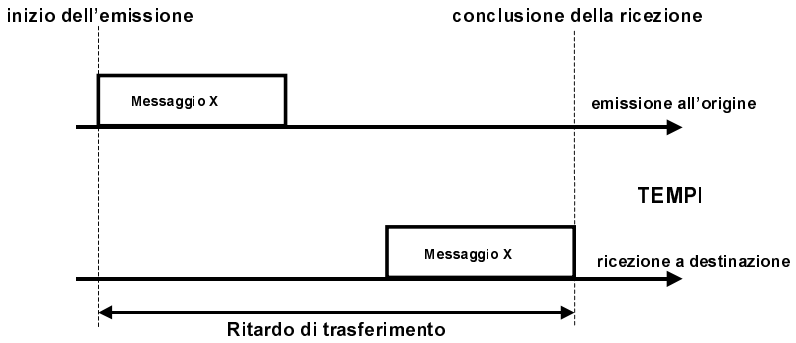


Figura IV.3.4 – Per la definizione di ritardo di trasferimento di un messaggio

Esaminiamo come questi varii contributi si combinano per determinare il ritardo *D*. Il ritardo di propagazione complessivo tra origine e destinazione dipende unicamente dalla lunghezza del percorso di rete associato al circuito virtuale. I ritardi di attraversamento dei nodi coinvolti dipendono dal carico di lavoro che interessa la rete e quindi anche questi nodi. Infine i tempi di emissione sui rami attraversati forniscono un contributo a *D* che è determinato da due fattori: l'effetto *"bitdotto"* (pipelining) e quello "*extra-informazione*" (overhead).

Per valutarne l'incidenza supponiamo che ogni nodo operi in modo che un segmento informativo sia completamente memorizzato prima di essere rilanciato verso il nodo successivo lungo il percorso di rete. Allora è immediato convincersi che il contributo dei tempi di emissione può essere ridotto segmentando il messaggio in UI più corte. Ad esempio, come in Fig. IV.3.5, consideriamo il caso di un percorso di rete a due salti. Se il messaggio non è segmentato, questo contributo è uguale a 2 volte il tempo di emissione dell'intero messaggio, dato che questo, una volta effettuato il primo salto, può essere riemesso solo dopo essere stato completamente memorizzato. Se invece il messaggio è segmentato in due UI e se si trascura l'aggiunta dei bit di intestazione, il contributo è uguale a 1,5 volte il tempo di emissione dell'intero messaggio. Questa riduzione è legata al fatto che il secondo nodo può rilanciare la prima UI non appena ne ha completata la memorizzazione e mentre sta ancora completando la memorizzazione della seconda UI. Questo parallelismo tra le emissioni successive su rami consecutivi di un percorso di rete è reso possibile dalla segmentazione del messaggio. Si tratta dell'effetto bitdotto.

Quanto ora detto potrebbe indurre alla conclusione che convenga incrementare la segmentazione e quindi operare con UI più corte a parità di lunghezza del messaggio. Aumentando però il numero di UI in cui viene segmentato il messaggio, aumenta anche il numero di bit di extra-informazione, da aggiungere come intestazione al testo di ogni UI. Ad esempio, se si segmenta il messaggio in *P* UI, se *H* (bit) è la lunghezza (supposta costante) dell'intestazione per ogni UI e se infine *M* (bit) è la lunghezza del messaggio (comprensiva della sua intestazione), allora il numero totale di cifre binarie da trasferire aumenta da *M* a $M+P\,H$. Questo incremento, che chiameremo "effetto extra-informazione", comporta un corrispondente aumento del contributo legato al tempo di emissione.

Calcoliamo ora il ritardo di trasferimento *D* del messaggio di lunghezza *M*, che ipotizziamo variabile aleatoriamente. Indichiamo con *L* la lunghezza *massima* (in bit) del testo di una delle UI in cui viene segmentato il messaggio. Il messaggio è segmentato in testi di lunghezza *L*, con l'ultimo testo costituito dallo sfrido della segmentazione. Allora, se $\lceil x \rceil$ denota il più piccolo valore intero "maggiore di" o "uguale a" *x*, il numero *P* è uguale a

$$P = \left\lceil \frac{M}{L} \right\rceil \quad .$$

Di queste UI le prime *P-1* hanno un testo che è costituito da *L* bit, mentre il testo dell'UI finale ha lunghezza compresa tra 1 e *L*. Il numero totale *B* di cifre binarie dopo la formazione delle UI con le loro intestazioni è quindi dato da

$$B = M + \left\lceil \frac{M}{L} \right\rceil H \quad .$$

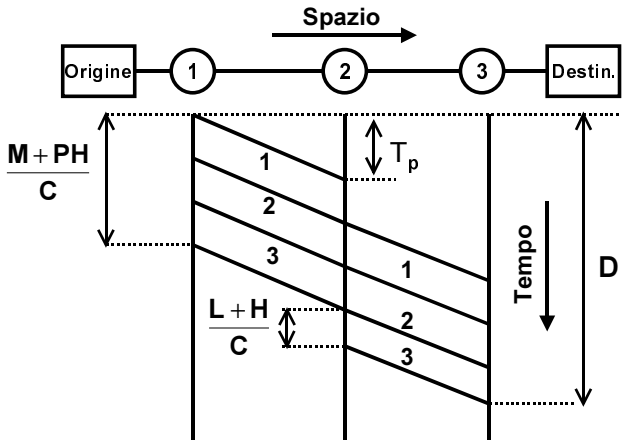


Figura IV.3.5 – Effetto bitdotto nel trasferimento di un messaggio segmentato in UI

Dalle espressioni di *B* e di *P* si può valutare come variano l'*efficienza di trasferimento* e il *carico di elaborazione* in funzione della lunghezza *L*. Se *M* tende all'infinito, la quota parte di extra-informazione è uguale a

$$\frac{H}{H+L}$$

e quindi aumenta al diminuire di *L*. Al decrescere di *L* a parità di *M*, diminuisce allora l'efficienza di trasferimento. D'altra parte il carico di elaborazione aumenta al crescere del numero *P* di UI per un fissato *M*. Ma *P* aumenta se si diminuisce *L*. Quindi al decrescere di *L* a parità di *M* aumenta il carico di elaborazione. In sintesi l'efficienza di trasferimento e il carico di elaborazione consigliano di operare con elevati valori di *L*.

Per mettere in evidenza come agisce *L* sul ritardo *D* e con riferimento ai rami del circuito virtuale, indichiamo con

- Δ il ritardo di propagazione su questi rami, supposti di lunghezza costante;
- C la capacità di trasferimento degli stessi rami supposti di uguale capacità.

Con queste notazioni CD e $C\Delta$ sono i valori dei ritardi D e Δ normalizzati rispetto al tempo di trasmissione di un bit. Supponiamo che

- il circuito virtuale sia costituito da *S* rami in cascata;
- la rete sia debolmente caricata in modo che possa essere trascurato il ritardo di attraversamento dei nodi appartenenti al circuito virtuale;
- siano trascurabili gli eventi di errore su ogni ramo;
- risulti $M \geq L$.

Come di deduce dalla Fig. IV.3.5, si ha

$$CD = C\Delta S + (L+H)(S-1) + M + \left\lceil \frac{M}{L} \right\rceil H \quad .$$

Il primo addendo, legato al ritardo di propagazione, è indipendente da L. Il secondo, derivante dall'effetto extra-informazione, cresce linearmente con L; il terzo, coincidente con il numero B di cifre binarie dopo la segmentazione e l'aggiunta nell'intestazione, è inversamente proporzionale a L.

È interessante determinare il valore L_{opt} per il quale risulta minimo il valore atteso E[CD] di CD. Facendo l'approssimazione

$$E\left[\left\lceil \frac{M}{L} \right\rceil\right] = E\left[\frac{M}{L}\right] + \frac{1}{2}$$

che è ragionevole se la distribuzione di M è uniforme su intervalli di L bit, si ottiene

$$E[CD] = C\Delta S + (L+H)(S-1) + E[M]\left(1 + \frac{H}{L}\right) + \frac{H}{2} \quad .$$

Da questa espressione per differenziazione risulta infine

$$L_{opt} = \sqrt{\frac{E[M]H}{S-1}} \quad .$$

Conseguentemente, quando H cresce, aumenta anche L_{opt} . Quando invece cresce la lunghezza del percorso (rappresentata dal numero S di rami), si ha una diminuzione di L_{opt} . Quando infine aumenta la lunghezza media del messaggio, cresce anche L_{opt} .

Per concludere passiamo al caso di *flusso continuo CBR,* per il quale il ritardo di trasferimento D che interessa è l'intervallo di tempo tra l'istante in cui un dato bit entra nella rete e l'istante in cui lo stesso bit ne esce. Utilizziamo, oltre all'intestazione H già definita, le seguenti ulteriori notazioni:

- R il ritmo binario di sorgente, che è per ipotesi costante;
- L la lunghezza, supposta costante, del testo di una UI;
- C_i la capacità di trasferimento dell'i-esimo ramo.

Supponiamo che

- per ogni ramo appartenente al percorso del flusso informativo, risulti

$$\frac{L+H}{C_i} \le \frac{L}{R} \quad ;$$

  cioè le UI siano trasferite con intervallo temporale imposto dal tempo di riempimento del pacchetto (*ritardo di pacchettizzazione*) e subiscano su ogni ramo un ritardo (*tempo di emissione*) che è sempre non superiore a quello di pacchettizzazione;
- il ritardo di propagazione sia trascurabile;
- la rete sia debolmente caricata in modo che possa essere trascurato il ritardo di attraversamento dei nodi.

Con queste ipotesi, si ha allora

$$D = \frac{L}{R} + (L+H)\sum_i \frac{1}{C_i} \quad ,$$

ove il primo addendo è il ritardo di pacchettizzazione, mentre il secondo è il tempo di emissione di una UI sull'insieme dei vari rami che costituiscono il percorso del flusso informativo. Si vede che:

- D diminuisce quando L diminuisce, finché per uno o più rami risulti

$$\frac{L+H}{C_i} = \frac{L}{R} \quad ;$$

  questa condizione fornisce il *minimo* ritardo di trasferimento;

- diminuendo ulteriormente L, il ritardo di trasferimento diventa infinito, in quanto si ha accumulo indefinito di UI sul ramo per cui

$$\frac{L}{R} < \frac{L+H}{C_i} \quad ;$$

- all'aumentare della capacità di trasferimento C_i, l'addendo dominante nell'espressione di D è L/R, termine che non è influenzato dalla presenza di altro traffico.

IV.4 Architetture protocollari

L'architettura protocollare riguarda il modello di riferimento che viene adottato per realizzare uno specifico modo di trasferimento. Essa descrive la stratificazione delle funzioni di comunicazione e stabilisce l'attribuzione di tali funzioni alle apparecchiature terminali e di rete. In questo paragrafo ci riferiremo, per semplicità, ai modi di trasferimento riguardanti le informazioni di utente; il caso dell'informazione di segnalazione verrà considerato nel par. IV.8.

Consideriamo i modelli di riferimento descritti nel par. II. 8 e, in particolare nelle Figg. II.8.1 e II.8.2.

Lo strato più alto tra quelli di trasferimento verrà indicato nel seguito come strato di *"modo di trasferimento"* (per brevità *strato MT*). L'identificazione degli strati di trasferimento e, in particolare, dello strato MT qualifica il modo di trasferimento utilizzato per trasportare l'informazione; in particolare il relativo servizio di strato può essere *con* o *senza* connessione.

Per fissare le idee, assumiamo che la stratificazione delle funzioni di comunicazione sia quella a sette strati adottata nel modello OSI. Inoltre ci limitiamo dapprima a considerare particolarizzazioni del primo modello di riferimento. Individuiamo lo strato di modo di trasferimento. Questo è lo strato gerarchicamente più elevato che interessa, oltre agli apparecchi terminali, anche i nodi di rete; include una funzione di rilegamento. Almeno in linea di principio e nei limiti della stratificazione adottata dal modello OSI, lo strato MT può in alternativa essere quello fisico, quello di collegamento o quello di rete. La individuazione dello strato MT è legata a considerazioni varie; ne presentiamo qui alcune per chiarire, seppure sommariamente, i termini della questione.

Innanzitutto, la funzione di rilegamento svolta da un nodo ha, tra le sue componenti essenziali, la funzione di commutazione. Questa è il risultato di una demultiplazione dei flussi informativi entranti in un nodo e di una loro successiva multiplazione in accordo agli indirizzi (impliciti o espliciti) dei segmenti informativi componenti questi flussi. Può essere espletata sulla base delle stesse informazioni di protocollo di strato, che sono utilizzate per identificare i segmenti informativi in un flusso multiplato. Si può quindi concludere che lo strato di modo di trasferimento è quello in cui sono svolte le *funzioni di commutazione e di multiplazione*.

D'altra parte, la individuazione dello strato MT ha implicazioni dirette sulle prestazioni del servizio di rete e, in particolare, sui gradi di trasparenza temporale e di integrità informativa. Ciò è chiarito nei due successivi esempi, ove viene messo in evidenza che gli effetti di una scelta dello strato MT sono tra loro contrastanti per ciò che riguarda il conseguimento di fissati obiettivi prestazionali.

Riferiamoci al trattamento protocollare del flusso informativo entrante in un nodo di rete; lo svolgimento di una funzione appartenente a un dato strato richiede, da parte del nodo, la preventiva chiusura dei protocolli degli strati inferiori. Se la scelta del livello dello strato MT fosse indirizzata verso un rango alto, aumenterebbe la complessità delle procedure che debbono essere gestite da un nodo di rete e quindi peggiorerebbe la prestazione di grado trasparenza temporale. Ovviamente si otterrebbe un effetto opposto indirizzando la scelta verso uno strato MT di rango basso.

Se lo strato MT fosse ad esempio quello di rete o quello di collegamento, il protocollo dello strato di collegamento (che è tra l'altro preposto alla funzione di recupero degli errori trasmissivi) potrebbe

essere gestito sezione per sezione, in accordo al modello di riferimento di Fig. IV.4.1. Ciò consentirebbe di fronteggiare con maggior successo i casi di canali trasmissivi con tasso di errore binario relativamente elevato. Conseguentemente la scelta di uno strato MT di ordine più elevato (rispetto allo strato fisico) comporta migliori prestazioni di integrità informativa.

Alla luce delle considerazioni ora svolte, si possono commentare le due alternative di architettura protocollare nelle Figg. IV.4.1 e IV.4.2, che costituiscono una particolarizzazione del modello di riferimento di Fig. II.8.1.

Nell'alternativa di Fig. IV.4.1 lo strato MT è quello *fisico* e quindi in questo strato sono svolte le funzioni di commutazione e di multiplazione. Inoltre tutte le funzioni logiche preposte al trattamento protocollare dell'informazione (a cominciare da quelle tipiche dello strato di collegamento) sono espletate con interazioni da estremo a estremo. Conseguentemente i nodi di accesso e di transito sono completamente trasparenti nei confronti dei flussi informativi che li attraversano. Questa architettura protocollare è particolarmente adatta per la fornitura di un servizio di rete con spiccate esigenze di un elevato grado di trasparenza temporale e con requisiti meno stringenti di integrità informativa.

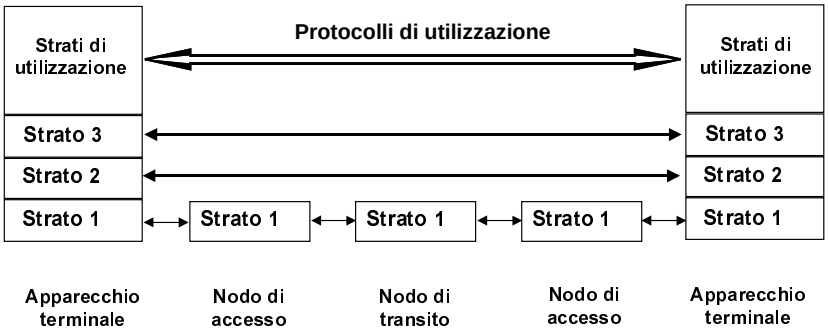


Figura IV.4.1 – Prima alternativa di architettura protocollare

Nella seconda alternativa di Fig. IV.4.2 si considera invece il caso in cui lo strato MT è quello di *rete*. Ne consegue che i protocolli dei primi tre strati del modello OSI debbono essere aperti e chiusi in ogni nodo della rete di trasporto. Questa architettura protocollare risponde quindi bene alle esigenze di un elevato grado di integrità informativa, ma con un modesto grado di trasparenza temporale.

Una terza alternativa di architettura protocollare è intermedia tra le due già considerate. È stata definita in modo da consentire un trasferimento di informazione con prestazioni indipendenti dalle caratteristiche del corrispondente servizio applicativo. Questa caratteristica è considerata di rilievo per un servizio di rete operante a sostegno di comunicazioni multimediali. L'alternativa in questione, mostrata in Fig. IV.4.3, costituisce una particolarizzazione del modello di riferimento di Fig. II.8.2. Sono spostati verso i bordi della rete i siti di svolgimento delle funzioni di protocollo, che dipendono dalle caratteristiche del servizio applicativo e sono onerose in termini di tempo richiesto per il loro svolgimento. Si allude in particolare al controllo di flusso, al recupero d'errore e alla sequenzializzazione delle UI. Ciò comporta una drastica semplificazione delle procedure svolte nei nodi di transito e, conseguentemente, una diminuzione dei ritardi di attraversamento e un aumento delle portate di nodo.

In questa terza alternativa si adotta una multiplazione dinamica e una commutazione con attraversamento ad immagazzinamento e rilancio. La scelta dello strato MT è conseguentemente il risultato del rispetto di due vincoli:

- attribuire allo strato MT un rango che sia il più basso possibile;
- differenziare lo strato MT da quello fisico per tener conto delle esigenze di bufferizzazione poste dalle funzioni di multiplazione e di commutazione.

Ciò richiede una completa ridefinizione degli strati più bassi del modello OSI e, in particolare, l'introduzione di due nuovi strati immediatamente al di sopra dello strato fisico. Questi sono in ordine gerarchico decrescente: lo strato di *adattamento* e lo strato di *modo di trasferimento*. Lo strato di adattamento contiene le funzioni dipendenti dal servizio applicativo, che hanno lo scopo di adattare le esigenze specifiche dei servizi applicativi alle potenzialità del servizio di rete. Nello strato di modo di trasferimento sono incluse le funzioni di multiplazione e di commutazione. Lo strato fisico garantisce una efficiente utilizzazione del mezzo trasmissivo e un trasferimento affidabile della sequenza di cifre binarie.

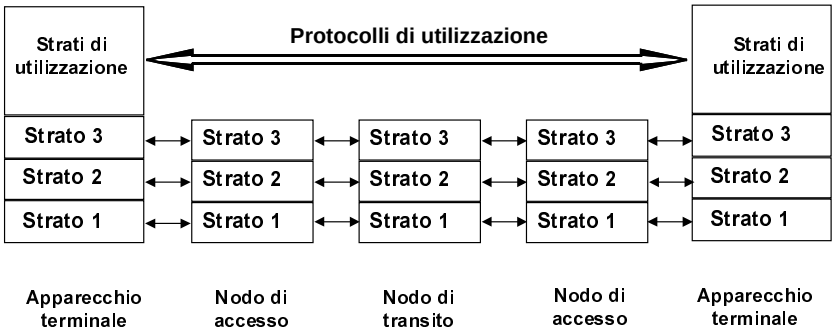


Figura IV.4.2 – Seconda alternativa di architettura protocollare

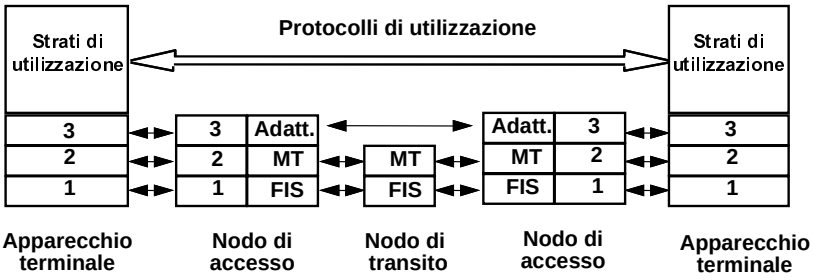


Figura IV.4.3 – Terza alternativa di architettura protocollare
*[*Adatt*.: strato di adattamento;* MT*: strato di modo di trasferimento;* FIS*: strato fisico]*

L'architettura protocollare che ne risulta è quella mostrata in Fig. IV.4.3, dove si assume che gli apparecchi terminali abbiano una struttura protocollare del tipo OSI. I nodi di transito effettuano solo funzioni di strato fisico e di strato di modo di trasferimento. Le funzioni di strato di adattamento sono localizzate nei nodi di accesso. In tal modo, il nucleo della rete di trasporto effettua solo le funzioni *comuni* a tutti i servizi applicativi. In corrispondenza dei bordi della rete di trasporto sono invece svolte le funzioni *dipendenti* dai servizi applicativi.

IV.5 Alcuni modi di trasferimento

Gli attuali orientamenti nella scelta dei modi di trasferimento più convenienti per l'attuazione di servizi di rete a supporto di applicazioni multimediali con esigenze di banda stretta o larga sono a favore dei *modi orientati al pacchetto*. Caratteristica distintiva di questi modi è l'impiego di schemi di multiplazione dinamica e di commutazione a pacchetto. Conseguentemente l'attraversamento è attuato

sempre nel modo ad immagazzinamento e rilancio, mentre le contese di utilizzazione sono normalmente risolte con meccanismi prevalentemente a ritardo.

Da un punto di vista prestazionale, i modi orientati al pacchetto garantiscono un più elevato grado di flessibilità di accesso rispetto ai *modi orientati al circuito*, che sono contraddistinti da una multiplazione statica e da una commutazione a circuito. Il minore grado di trasparenza temporale, che connotava le prime realizzazioni dei modi orientati al pacchetto (quasi esclusivamente utilizzate nelle reti dedicate per dati) è oggi sostanzialmente superata attraverso gli accorgimenti che verranno chiariti nel seguito (cfr. § IV.5.5).

Nel seguito del paragrafo riassumiamo dapprima gli aspetti specifici che caratterizzano i modi di trasferimento tradizionali e cioè quelli *a circuito* (§ IV.5.1) e *a pacchetto* (§ IV.5.2). Successivamente (§ IV.5.3) si illustrano le caratteristiche del modo di trasferimento *asincrono* (ATM - Asynchronous Transfer Mode), che fino a un recente passato appariva come il più serio candidato per la realizzazione di un ambiente di servizi integrati a larga banda. La rassegna è completata (§ IV.5.4) con il modo di trasferimento a *rilegamento di trama* (FR - Frame Relaying), che è stato standardizzato per definire una modalità più efficiente per realizzare un trasporto di dati in ambiente di integrazione dei servizi a banda stretta.

Per ognuno di questi modi, vengono precisati lo schema di multiplazione, il principio di commutazione e l'architettura protocollare, in accordo ai criteri di classificazione introdotti nei paragrafi precedenti. In particolare, per ciò che riguarda l'architettura protocollare, viene individuato lo strato di modo di trasferimento.

Il paragrafo si chiude con alcune considerazioni riguardanti i modi di trasferimento orientati al pacchetto (§ IV.5.5): in particolare viene esaminata l’influenza della lunghezza delle UI utilizzate sulle prestazioni relative al trasporto dell’informazione.

IV.5.1 Trasferimento a circuito

Un modo di trasferimento a circuito, quale è stato ed è tuttora impiegato nelle reti telefoniche, è associato ad un servizio di rete con connessione; è l’esempio decisamente più significativo, di modo orientato al circuito. Utilizza:

- una multiplazione statica, di cui l'esempio più largamente diffuso nelle reti telefoniche numeriche è costituito dalla multiplazione PCM, con un RCB costituito da 8 cifre binarie e con un sub-canale di base avente capacità di trasferimento uguale a 64 kbit/s;
- una commutazione a circuito;
- un’architettura protocollare, in cui lo strato MT è quello fisico (cfr. Fig. IV.4.3).

Nel caso di una comunicazione punto-punto su base chiamata, il trasferimento dell'informazione d'utente avviene su una connessione fisica commutata (*circuito fisico*) da utente a utente. Per instaurare tale connessione all'inizio della chiamata e per abbatterla al termine della comunicazione, avviene uno scambio di informazione di segnalazione tra i due utenti chiamante e chiamato da un lato e la rete dall'altro. Nell'ambito della rete di trasporto si ha anche un trasferimento dell'informazione di segnalazione tra i nodi coinvolti nella chiamata.

La durata richiesta per instaurare la connessione è il *ritardo di instaurazione* e può variare, a seconda delle caratteristiche della rete e dei nodi che la compongono, da alcune decine di millisecondi ad alcuni secondi. Questo ritardo è il risultato dei tempi richiesti per elaborare l'informazione di segnalazione in ognuno dei nodi coinvolti nella chiamata e per trasferire tale informazione dal chiamante al chiamato (*segnalazione in avanti*) e viceversa (*segnalazione all'indietro*); dipende anche dai tipi di protocolli di segnalazione che vengono impiegati.

In relazione poi al carico della rete, possono verificarsi condizioni di *congestione*, che impediscono la instaurazione del circuito fisico da utente a utente. Tali condizioni possono

corrispondere a temporanea insufficienza di risorse di trasferimento sia all'interno di uno tra i nodi coinvolti nella chiamata, sia all'esterno di questi. Nel primo caso si ha l'impossibilità di formare una connessione diretta ingresso-uscita e si parla allora di *congestione interna*; il secondo caso corrisponde invece a una condizione di *congestione esterna*, che si manifesta con la impossibilità di pre-assegnare individualmente alla chiamata una delle linee di giunzione uscenti dal nodo.

IV.5.2 Trasferimento a pacchetto

Un modo di trasferimento a pacchetto, quale è stato realizzato nelle reti dedicate per dati, è associato ad un servizio di rete che può essere di due tipi

- quando è senza connessione, è normalmente indicato come servizio *datagramma*;
- quando è invece con connessione, è denominato servizio *a chiamata virtuale* o a *circuito virtuale,* a seconda che la connessione sia commutata o semi-permanente, rispettivamente.

È caratterizzato da:

- una multiplazione dinamica, in cui le UI (*pacchetti*) sono, in generale, di lunghezza variabile e sono trasferite su canali con asse dei tempi indiviso; il trattamento delle contese di utilizzazione è orientato al ritardo;
- una commutazione a pacchetto con o senza connessione a seconda di quale sia il servizio di rete da realizzare;
- una architettura protocollare, in cui lo strato MT è quello di rete (cfr. fig. IV.4.4).

Con riferimento a una chiamata virtuale, la rete, a seguito di scambio di informazione di segnalazione, mette a disposizione un circuito virtuale. Tale associazione è definita nella fase di instaurazione ed è attuata, in ogni nodo coinvolto nella chiamata, tramite una corrispondenza tra un canale logico entrante ed uno uscente.

Un esempio di particolare rilievo riguardante la realizzazione del servizio datagramma è offerto da Internet, nella sua versione originaria come rete dedicata per dati.

IV.5.3 Modo di Trasferimento Asincrono

L'ATM è un modo di trasferimento orientato al pacchetto che offre un servizio orientato alla connessione e in cui:

- la multiplazione è dinamica; le UI trattate sono chiamate *celle* e hanno un formato di lunghezza fissa; l'asse dei tempi è suddiviso in IT con o senza organizzazione in trama;
- la commutazione è a pacchetto con connessione;
- l'architettura protocollare segue le linee-guida descritte nel modello di Fig. IV.4.5, dove lo strato di modo di trasferimento è chiamato *strato ATM*.

Più in particolare, una cella ha una lunghezza complessiva di 53 ottetti di cui 5 sono dedicati all'intestazione. L'etichetta identifica le celle appartenenti allo stesso canale logico per mezzo di un identificatore di connessione. Sono previsti due tipi di canale logico: i *canali virtuali* (virtual channel) e i *percorsi virtuali* (virtual path). L'informazione di indirizzamento deve essere protetta contro possibili errori; conseguentemente al contenuto dell'intestazione è applicata una procedura di controllo di errore.

Una connessione virtuale è identificata da una sequenza di identificatori di connessione. Ogni elemento di questa sequenza è assegnato, per ciascun segmento di rete, durante la fase di instaurazione e viene restituito quando non è più necessario. La connessione, cosi' identificata, rimane invariata per tutta la durata della chiamata.

Nel corso della chiamata, la capacità di trasferimento è assegnata a domanda in relazione all'attività della sorgente e alle risorse disponibili. L'integrità della sequenza di celle su un canale logico è garantita dallo strato ATM.

L'ATM offre una potenzialità di trasferimento flessibile, comune a tutti i servizi, inclusi quelli senza esigenze di connessione. A tale scopo provvedono le funzionalità dello strato di adattamento. Il limite tra lo strato ATM e gli strati superiori corrisponde a quello tra le funzioni supportate dai contenuti dell'intestazione e quelle legate all'informazione contenuta nel testo della cella.

In corrispondenza all'interfaccia utente-rete sono attivate, in forma commutata o semi-permanente, connessioni logiche multiple verso destinazioni differenziate. Nel caso di connessione commutata la corrispondente informazione di segnalazione è trasferita in modo logicamente separato rispetto all'informazione di utente. La separazione dei piani di utente e di controllo segue le linee-guida illustrate per le architetture di comunicazione rappresentate da modelli a tre piani (cfr. par. II.7).

IV.5.4 Modo di trasferimento FR

L'FR è un modo di trasferimento che, come l'ATM, è orientato al pacchetto e offre un servizio con connessione commutata o semi-permanente. Più in particolare:

- la multiplazione è dinamica; le UI trattate sono chiamate *trame* e hanno un formato di lunghezza variabile; l'asse dei tempi è indiviso; la gestione delle contese di utilizzazione è a ritardo;
- la commutazione è a pacchetto con connessione;
- l'architettura protocollare segue le linee-guida descritte nel modello OSI, in cui allo strato MT è attribuito il rango 2.

Come nell'ATM, l'instaurazione di una connessione commutata richiede un trasferimento di informazioni di segnalazione attuato in modo logicamente separato rispetto all'informazione di utente. Questa separazione tra le funzionalità relative all'informazione di utente e a quella di controllo trova anche qui corrispondenza in una architettura a tre piani (cfr. par. II.7).

L'informazione di utente è trasferita utilizzando un protocollo, che è chiamato *LAPF* (Link Access Procedure for Frame Mode). Il trasferimento attraverso la rete conserva l'ordine di emissione delle trame tra il punto di accesso di origine e quello di destinazione (*conservazione della sequenzialità*). Il servizio di rete rivela gli errori trasmissivi, di formato e operativi e scarta le trame affette da errore. Non può essere evitata la perdita di trame e non è previsto alcun meccanismo di controllo di flusso attraverso l'interfaccia utente-rete. Questo insieme di caratteristiche ha l'obiettivo di ottenere un alleggerimento delle operazioni di trasporto e quindi con minor ritardo di trasferimento.

IV.5.5 Trasferimento a pacchetto e trasparenza temporale

Nelle realizzazioni più recenti di nodi operante secondo modi di trasferimento orientati al pacchetto, il valor medio e la varianza del ritardo di trasferimento sono tali da assicurare un elevato grado di trasparenza temporale. Ciò è dovuto alla riduzione delle funzionalità di protocollo espletate dai nodi di transito e alle possibilità oggi offerte di realizzare, da un lato, nodi con una drastica riduzione dei ritardi di attraversamento e, dall'altro, sistemi di trasmissione ad alta velocità.

Questo obiettivo è conseguibile, oltre che a seguito della utilizzazione di protocolli a bassa funzionalità, anche per effetto di quanto oggi consente la tecnologia dei sistemi di elaborazione e di quelli di trasmissione. È infatti possibile realizzare elaboratori di UI con alto parallelismo e quindi con ridottissimi tempi di elaborazione per ogni UI. Inoltre l'impiego di canali trasmissivi ad alta capacità, quali sono realizzabili su fibra ottica, consente di ridurre fortemente i tempi di emissione di ogni UI.

Tuttavia l'utilizzazione di un modo di trasferimento orientato al pacchetto in un ambiente di servizi integrati implica due ulteriori componenti del ritardo di trasferimento, che assumono, per flussi informativi continui in un contesto di alta velocità di trasferimento, valori non trascurabili e, in alcuni casi, addirittura dominanti. Si allude al *ritardo di pacchettizzazione* e a quello *di de-pacchettizzazione*,

che sono introdotti, rispettivamente, all'inizio e alla fine di ogni parte di rete in cui è applicato il modo di trasferimento in esame.

Nel caso di flusso informativo continuo, il ritardo di pacchettizzazione è già stato definito in § IV.3.3 come il tempo necessario a riempire una UI; è dato dal rapporto fra la lunghezza (in cifre binarie) di una UI e il ritmo binario di emissione della sorgente. Il ritardo di de-pacchettizzazione è invece il tempo necessario a *equalizzare* i ritardi di trasferimento (cfr. § IV.1.1), se questa operazione è richiesta com'è talvolta il caso di servizi aventi esigenza di elevato grado di trasparenza temporale, e a permettere la ricostruzione del flusso informativo continuo.

Nei modi orientati al pacchetto l'informazione è trasferibile in piccole UI aventi lunghezza *fissa* o *variabile*. I vantaggi potenziali di UI con lunghezza variabile risiedono in considerazioni di flessibilità, di efficienza e di qualità di servizio. Di questi vantaggi possono essere dati due esempi, con riferimento alle applicazioni riguardanti la voce e i dati.

Le sorgenti di dati emettono segmenti informativi composti da un numero di ottetti (di cifre binarie) che può variare da 1 o 2 unità fino a varie centinaia di migliaia. In questo caso l'uso di UI con lunghezza fissa sarebbe inefficiente per il trasferimento di segmenti informativi sia più brevi della lunghezza fissa (in quanto parte della UI non verrebbe utilizzata), sia più lunghi (in quanto sarebbe necessario l'impiego di una molteplicità di UI, ognuna con la sua informazione aggiuntiva di protocollo).

Per ciò che riguarda le applicazioni vocali, queste sono basate su un vasto campo di possibili ritmi binari di codifica (da una decina di kbit/s fino ad alcune centinaia di kbit/s in corrispondenza di differenti qualità dell'informazione vocale trasferita), ma comportano in ogni caso analoghi vincoli sul valore di picco del ritardo di trasferimento. Il mancato rispetto di questi vincoli comporta l'introduzione di dispositivi per il controllo dell'eco (*cancellatori d'eco*), con conseguenti costi aggiuntivi. D'altra parte, in queste applicazioni, il ritardo di trasferimento è largamente dipendente dal ritardo di pacchettizzazione, che, come si è visto, è direttamente proporzionale alla lunghezza della UI e inversamente proporzionale al ritmo binario di emissione della sorgente. Conseguentemente, per ogni valore di questo ritmo, esiste un valore massimo della lunghezza di UI.

I vantaggi di una UI con lunghezza fissa sono principalmente legati alla semplificazione delle funzionalità di commutazione nei nodi di rete. Infatti questa soluzione facilita il dimensionamento degli organi di nodo (buffer) preposti all'immagazzinamento delle UI e assicura un flusso continuo di informazione di etichetta negli elaboratori di UI. È consentita una elaborazione più veloce del flusso asincrono di UI, dato che, nell'ambito dell'alternativa *S* dello schema di multiplazione, la delimitazione delle UI è identificabile in modo agevole, quando siano assicurate condizioni di sincronizzazione di IT.

Se viene adottata una UI con lunghezza fissa, questa deve essere la più corta possibile. Tale scelta è legata a ragioni di efficienza per ciò che riguarda la funzione di commutazione e a motivi prestazionali per quanto attiene la riduzione del ritardo di pacchettizzazione. Inoltre, l'impiego di UI di piccola lunghezza rende più tollerabile il loro eventuale scarto a seguito di rivelazione di errore sulle loro etichette.

Per concludere, è poi da considerare che la portata media di un nodo che adotta un modo di trasferimento orientato al pacchetto deve essere aumentata di oltre due ordini di grandezza rispetto a quella che è necessaria nei nodi a pacchetto operanti nelle reti per dati. Supponiamo infatti che l'informazione da trasferire sia quella emessa da una sorgente vocale in una conversazione telefonica e assumiamo che:

- la codifica della voce sia effettuata con rivelazione dei tratti di voce attiva e il grado di intermittenza della sorgente sia uguale a 2,5 (ipotesi 1);
- la lunghezza, supposta fissa, di una UI sia tale da contenere un segmento di voce attiva con durata di 10 ms (ipotesi 2);

- ogni chiamata telefonica abbia una durata media di 180 s. (ipotesi 3).

In queste condizioni, sulla base delle ipotesi 1) e 2), in ognuna delle due direzioni di trasferimento vengono emesse in media $1/(10 \bullet 10^{-3} \bullet 2{,}5) = 40$ UI/s e quindi $40 \bullet 180 = 7.200$ UI nella durata di una chiamata. Ciò si riflette sul valore di portata media richiesto per un nodo. A tale riguardo consideriamo un autocommutatore locale telefonico, che sia in grado di servire fino a 100.000 utenti. Allora, se le chiamate hanno la durata media che si è sopra ipotizzata e se l'intensità media di traffico offerto da ogni utente è di 0,1 Erl, l'autocommutatore deve essere in grado di trattare in media fino a circa 55 chiamate/s. Conseguentemente, un autocommutatore locale, che operi con un modo di trasferimento orientato al pacchetto e che abbia una capacità di trattamento di chiamata avente valore comparabile con quella dell'autocommutatore telefonico, dovrebbe avere una portata media di circa 800.000 UI/s, in accordo con quanto si è anticipato circa il confronto con le portate medie dei nodi a pacchetto per applicazioni di dati.

Valori così elevati di portata media di nodo sono oggi realizzabili attraverso i mezzi già menzionati ai fini della riduzione del ritardo di attraversamento.

IV.6 Nodi a circuito

I commutatori a circuito mettono a disposizione i mezzi (elementi di commutazione) per instaurare, a richiesta, una connessione diretta tra una terminazione di ingresso (appartenente ad un insieme *J*) e una terminazione di uscita (appartenente ad un insieme *U*). L'instaurazione e l'abbattimento della connessione sono attuate, in alternativa, tramite interventi al servizio di una specifica comunicazione ovvero tramite provvedimenti di natura gestionale al servizio delle comunicazioni tra due o più insiemi di utenti del servizio di rete. Il primo caso riguarda nodi che sono in grado di gestire connessioni commutate; il secondo è tipico di nodi che trattano connessioni semi-permanenti. In entrambi i casi la connessione è instaurata attraverso l'azionamento di uno o più elementi di commutazione; l'abbattimento è l'operazione complementare alla instaurazione.

Nel caso di nodo in grado di gestire connessioni commutate, l'instaurazione e l'abbattimento di una connessione rientrano normalmente tra gli adempimenti connessi al trattamento di chiamata e, in questo caso, avvengono all'inizio e al termine della comunicazione rispettivamente. Possono però avvenire nel corso della comunicazione quando sia richiesto un adeguamento della connessione alle esigenze variegate di una comunicazione multimediale. La gestione di connessioni commutate è effettuata in ogni caso a seguito di messaggi di segnalazione, che provvedono a coordinare i compiti dei nodi interessati alla formazione e alla liberazione di un percorso di rete da estremo a estremo al servizio di una specifica comunicazione. L'instaurazione di una connessione diretta all'interno di un nodo a circuito è una operazione che richiede una decisione di instradamento. Quest'ultima avviene nel rispetto di opportune regole.

Gli elementi di commutazione hanno un funzionamento dipendente dalla grandezza fisica (spazio, tempo, frequenza, ecc.) che viene utilizzata per distinguere le terminazioni di ingresso e di uscita. Ad un nodo di rete accedono linee (di ingresso o di uscita) che sono il supporto di uno o più flussi di informazione, associati a una o più comunicazioni contemporanee. Un commutatore a circuito può essere realizzato, in linea di principio, a divisione di spazio o a divisione di tempo o a divisione di frequenza (o di lunghezza d'onda).

Queste sono anche le tecnologie con le quali nel tempo sono state realizzate le reti di connessione, e cioè quelle strutture che sono preposte ad attuare la funzione di attraversamento in un nodo a circuito. La tecnica a divisione di tempo è quella attualmente impiegata e caratterizza la realizzazione di reti di connessione in tecnologia numerica.

Per chiarire quale sia stata l'evoluzione delle tecniche realizzative dei nodi a circuito è utile fare riferimento al loro impiego nelle reti telefoniche.

La tecnica di progettazione delle reti telefoniche, con riferimento specifico alle apparecchiature di commutazione (*autocommutatori*) e ai sistemi di trasmissione che ne fanno parte, si è sviluppata per oltre mezzo secolo in un ambiente di comunicazioni completamente analogiche, quale era consentito dalla tecnologia disponibile, e ha condotto a realizzazioni via via più perfezionate.

Dalla seconda metà degli anni '60, la rete telefonica ha iniziato l'evoluzione tecnologica e sistemistica verso un impiego generalizzato delle tecniche numeriche in tutti gli apparati componenti. Questo sviluppo, ormai giunto a conclusione, ha riguardato dapprima i sistemi di trasmissione e, successivamente (dopo circa un decennio), anche gli apparati di commutazione. L'obiettivo finale è stato la realizzazione di una *rete numerica integrata nelle tecniche* (IDN - Integrated Digital Network).

Gli elementi distintivi di una IDN telefonica sono stati e sono tuttora:

1) l'utilizzazione di *tecnologie completamente elettroniche*;
2) l'impiego di elaboratori per controllare, da un punto di vista decisionale, il funzionamento degli autocommutatori (*elaboratori di controllo*);
3) la graduale sostituzione, nella rete di trasporto, della Segnalazione Associata al Canale con la *Segnalazione a Canale Comune* (SCC) (cfr. § V.7.1. e V.7.2).

Circa il punto 1), si può ricordare che, almeno per ciò che riguarda le apparecchiature di commutazione, la tecnologia utilizzata per circa 50 anni è stata quella *elettromeccanica* e successivamente (dalla metà degli anni '60) quella *semielettronica* (e cioè in parte elettromeccanica e in parte elettronica). La transizione verso un impiego generalizzato di tecnologie elettroniche, dapprima con l'impiego di componentistica discreta a semiconduttori e, in un momento successivo, con l'utilizzazione di tecniche microcircuitali a vari livelli di integrazione, ha rivoluzionato le tecniche progettuali, le metodologie costruttive e l'organizzazione di esercizio degli impianti.

Relativamente poi al punto 2), un elaboratore controlla un autocommutatore sulla base di un opportuno programma. Un autocommutatore *con controllo a programma memorizzato* (SPC - Stored Program Control) ha presentato (e presenta tuttora) vantaggi decisivi rispetto all'impiego di logiche di tipo cablato. Tra questi si possono citare:

- la *modularità* e la *modificabilità*, e cioè la suddivisione del programma di controllo in moduli che possono essere modificati separatamente e indipendentemente, in modo relativamente agevole;
- la *adattabilità al sito* e la *estensibilità*, e cioè la possibilità di fronteggiare sia specifiche condizioni di sito all'atto della prima installazione, che mutate condizioni di esercizio dell'autocommutatore;
- la *autodiagnostica* e la *capacità di adattamento* a condizioni di funzionamento anormali, e cioè, da un lato, la possibilità offerta agli addetti di individuare e di eliminare rapidamente eventuali guasti e, dall'altro, la possibilità di modificare la configurazione di sistema per consentirne la sopravvivenza.

Con riguardo infine al punto 3), nelle IDN telefoniche le informazioni di segnalazione vengono scambiate fra i nodi di commutazione per il tramite della rete SCC.

IV.6.1 Divisione di spazio

Se *ogni linea di accesso* è il supporto di un *singolo flusso informativo* (associato quindi a una singola comunicazione), la grandezza fisica che si utilizza per distinguere una terminazione da un’altra è lo *spazio*; l’insieme dei flussi informativi che entrano nel sistema o che ne escono è allora strutturato con una particolare modalità di multiplazione statica, chiamata *multiplazione a divisione di spazio*.

Il sistema di commutazione, che opera su flussi informativi distinti su base spaziale, è detto anch’esso *a divisione di spazio* (DS). La tecnica DS è stata applicata prevalentemente in passato con l’uso di *tecnologie elettromeccaniche.*

Elemento di commutazione, che è di base per la tecnica DS, è il *punto di incrocio*, e cioè un dispositivo a due stati: quello di *apertura* e quello di *chiusura*. Un punto di incrocio è collegato, da un lato, con una terminazione *di ingresso* e, dall'altro lato, con una *di uscita*: la chiusura del punto di incrocio stabilisce una connessione diretta tra queste terminazioni; la sua apertura abbatte la connessione in atto.

Un insieme di punti di incrocio:

- a cui fanno capo $|J|$ terminazioni di ingresso e $|U|$ terminazioni di uscita;
- organizzato in una struttura matriciale a $|J|$ righe e a $|U|$ colonne, tale che all'incrocio di ogni riga con ogni colonna può essere collocato un punto di incrocio;
- in cui ad ogni riga corrisponde una terminazione di ingresso e ad ogni colonna una terminazione di uscita,

costituisce una *matrice spaziale a divisione di spazio* (S-matrice DS). In una *S-matrice DS*, la connessione della terminazione di ingresso facente capo alla riga *i*-esima con quella di uscita facente capo alla colonna *j*-esima si attua attraverso la chiusura del punto di incrocio che è collocato all'intersezione della riga *i*-esima con la colonna *j*-esima.

IV.6.2 Divisione di tempo

Se ogni linea di accesso è il supporto di una molteplicità di flussi informativi, che corrispondono ad altrettante comunicazioni in corso di svolgimento e se la grandezza fisica che si utilizza per distinguere una terminazione da un'altra è il *tempo*, l'insieme dei flussi informativi che entrano nel sistema o che ne escono è allora strutturato con una *multiplazione statica a divisione di tempo,* e cioè con asse dei tempi suddiviso in intervalli temporali (IT) e organizzato in trame.

Il sistema di commutazione, che opera su flussi informativi distinti su base temporale, è detto anch'esso *a divisione di tempo* (DT); la tecnica DT è normalmente in *forma numerica*, dato che tali sono i flussi informativi su cui opera, e si applica attualmente con l'uso *di tecnologie elettroniche*.

Con riferimento a un linea di accesso (di ingresso o di uscita) a commutatore DT in forma numerica (*bus di accesso),* l'informazione (gruppo di cifre binarie) relativa a una comunicazione è contenuta in un IT che si ripresenta periodicamente con un intervallo uguale alla durata della trama. Un IT individua quindi una terminazione del sistema di commutazione: questa terminazione è *di ingresso* se il bus è di ingresso; è invece di uscita se il bus è di uscita.

Una commutazione DT in *forma numerica* consiste quindi nello *spostare temporalmente* (nel verso dei ritardi):

- il contenuto di un IT appartenente alla sequenza di trame su uno dei bus di ingresso *(trame di ingresso)*,
- in un IT (in generale con diverso numero d'ordine rispetto all'IT in ingresso) appartenente alla sequenza di trame su uno dei bus di uscita *(trame di uscita)*.

Lo scambio di IT può riguardare una stessa trama o due trame successive (Fig. IV.6.1). Il trasferimento può avvenire con uno o più scambi di IT.

Condizione necessaria affinché questa operazione sia possibile è che i flussi numerici multiplati all'ingresso e all'uscita del sistema siano tra loro *sincroni*, e cioè supportati da sincrosegnali aventi *uguale frequenza* e *uguale fasatura di trama*: ciò equivale a dire che la differenza di fase tra gli istanti caratteristici dei corrispondenti segnali numerici deve essere *invariante nel tempo*.

Con la tecnica a divisione di tempo, il *ritardo di attraversamento*, che, come in tutte le connessioni dirette, è di valore costante, è costituito da due componenti. Una, di valore normalmente trascurabile, corrisponde al ritardo di propagazione tra l'ingresso e l'uscita del nodo. L'altra è un multiplo della durata di un IT. Se, nell'ambito di una trama con N IT, i è il numero d'ordine dell'IT corrispondente

all'ingresso (i = 1, 2, ..., N) e j quello corrispondente all'uscita (j = 1, 2, ..., N), tale multiplo è almeno uguale a

$$\begin{array}{ll} j - i & \text{se} \quad j > i \\ N + j - i & \text{se} \quad j < i \end{array}$$

e cioè a seconda che lo scambio di IT avvenga nella stessa trama o in due trame successive. Può essere maggiore, se, all'interno della rete di connessione, è effettuato più di uno scambio di IT.

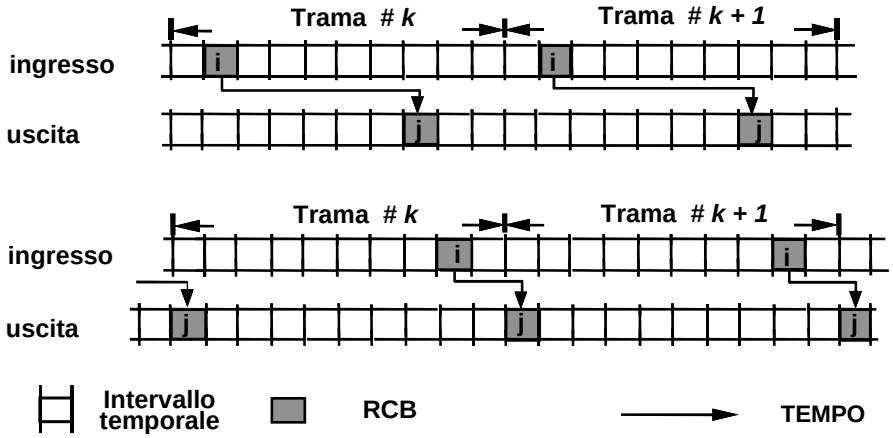


Figura IV.6.1- Tecnica di commutazione a divisione di tempo: scambio di IT (a) nella stessa trama o (b) in due trame successive.

L'elemento di commutazione , che è di base nella commutazione DT in tecnica numerica, è lo *scambiatore di IT*, e cioè un dispositivo di memoria, capace di introdurre un ritardo *costante* tra ingresso e uscita. Questo ritardo è tale da spostare, periodicamente a cadenza di trama, il gruppo di cifre binarie contenuto in un IT appartenente alle trame di ingresso in un diverso IT appartenente alle trame di uscita. Ciò equivale a connettere direttamente una terminazione di ingresso con una di uscita. Lo scambiatore di IT nella commutazione DT è quindi funzionalmente equivalente al punto di contatto nella commutazione DS.

Un elemento di commutazione DT, che può essere confrontato funzionalmente con una S-matrice DS, è la *matrice temporale* (T-matrice), e cioè un dispositivo che, nella versione più semplice, presenta un bus di ingresso e uno di uscita. La T-matrice è modellabile con due memorie (Fig. IV.6.2): la *memoria di commutazione*, che è preposta a trattare il flusso informativo attraversante il sistema, e la *memoria di comando*, che è adibita a funzioni di controllo della memoria di commutazione.

La funzione della T-matrice è trasferire il contenuto di uno qualunque degli IT appartenenti alle trame di ingresso in uno qualunque degli IT appartenenti alle trame di uscita. In una T-matrice, le terminazioni di ingresso e quelle di uscita sono in corrispondenza con gli IT appartenenti alle trame di ingresso e a quelle di uscita, rispettivamente. La connessione diretta tra la terminazione di ingresso i-esima e quella di uscita j-esima può essere attuata ritardando, periodicamente a cadenza di trama, il contenuto dell'IT i-esimo. L'entità del ritardo, che è specifica di ogni singola comunicazione e che è costante per tutta la durata di questa, deve essere commisurata alla posizione temporale dell'IT j-esimo immediatamente successivo nello stesso intervallo di trama o in quello seguente. L'operazione di ritardo è attuata dalla memoria di commutazione, mentre il controllo di questa operazione (entità del ritardo per ogni comunicazione e sua attuazione con periodicità di trama) è compito della memoria di comando.

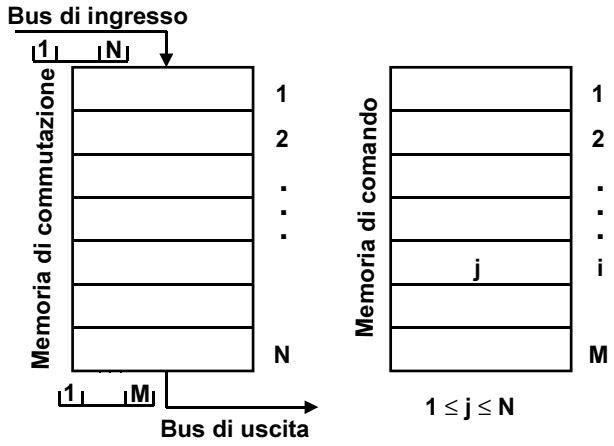


Figura IV.6.2 – Modello di T-matrice

Facendo riferimento alla connessione tra l'*i*-esimo IT appartenente alle trame di ingresso e lo *j*-esimo IT appartenente alle trame di uscita, un possibile funzionamento della T-matrice è il seguente:

- la memoria di commutazione è una RAM organizzata in celle, ognuna delle quali è in grado di memorizzare un RCB quale è contenuto in un IT appartenente alle trame di ingresso o di uscita;
- gli RCB sono depositati nella memoria di commutazione (*fase di scrittura*) rispettando l'ordine sequenziale degli IT appartenenti alle trame di ingresso; in particolare l'RCB contenuto nell'*i*-esimo IT è scritto nella cella *i*-esima;
- il contenuto della cella *i*-esima della memoria di commutazione viene successivamente prelevato (*fase di lettura*) in corrispondenza dello *j*-esimo IT appartenente alle trame di uscita;
- l'ordine di lettura della memoria di commutazione è stabilito esplorando ciclicamente i contenuti della memoria di comando; in tale memoria è disponibile, per ognuno degli IT appartenenti alle trame di uscita, il numero d'ordine dell'IT appartenente alle trame di ingresso da mettere in relazione con esso; cioè, al passo *j*-esimo dell'esplorazione della memoria di comando, si trova l'indirizzo *i* della cella appartenente alla memoria di commutazione;
- il riempimento degli indirizzi nella memoria di comando è effettuato nella fase di instaurazione della connessione corrispondente ad ogni comunicazione.

Una T-matrice, il cui bus di ingresso supporta trame con $|J|$ IT e il cui bus di uscita supporta trame con $|U|$ IT, è *funzionalmente equivalente* a una S-matrice DS con $|J|$ righe e $|U|$ colonne.

In conclusione un commutatore DT numerico, operante nell'ambito di un servizio di rete con connessione commutata su base chiamata, rende disponibile una connessione diretta tra un sub-canale di base del flusso multiplato in ingresso e un corrispondente sub-canale di base del flusso multiplato in uscita. Ad esempio, se la multiplazione statica è di tipo PCM e se quindi i suoi sub-canali di base hanno capacità di 64 kbit/s, tale è anche la capacità della connessione che viene messa a disposizione della chiamata per tutta la sua durata.

Almeno in linea di principio, possono anche essere attuate connessioni con capacità che siano multiple o sotto-multiple di quella del sub-canale di base della multiplazione statica. Nel primo caso si parla di *commutazione a multi-IT*, nel secondo di *commutazione a sotto-ritmo*. Le relative modalità di attuazione sono immediatamente deducibili, ove si tenga conto di quanto detto, a proposito della sovra e della sotto-multiplazione, rispettivamente.

Inoltre, sempre con riferimento all'evoluzione di una chiamata, durante la fase di trasferimento dell'informazione in una comunicazione punto-punto, le due parti hanno a disposizione una connessione fisica, che è caratterizzata da una portata e da un ritardo di trasferimento, che sono costanti per tutta la durata della chiamata. Questa condizione di trasparenza temporale, già più volte sottolineata come caratterizzante la tecnica a circuito ha due implicazioni. Da un lato essa offre un canale di trasferimento individuale, che viene reso disponibile per tutta la durata della chiamata e in maniera esclusiva per le parti in comunicazione. Dall'altro essa obbliga i due utenti alle estremità di questo canale a verificare, prima dell'instaurazione della connessione, la loro compatibilità in termini di ritmo binario emesso e ricevuto, di procedure di controllo, ecc.

IV.6.3 Divisione di frequenza

Se ogni linea di accesso è il supporto di una molteplicità di flussi informativi, che corrispondono ad altrettante comunicazioni in corso di svolgimento e se la grandezza fisica che si utilizza per distinguere una terminazione da un'altra è la *frequenza* (o la lunghezza d'onda), l'insieme dei flussi informativi che entrano nel sistema o che ne escono è strutturato con *una multiplazione statica a divisione di frequenza* (o di lunghezza d'onda).

Il sistema di commutazione, che opera su flussi informativi distinti sull'asse delle frequenze (o delle lunghezze d'onda), è detto anch'esso *a divisione di frequenza* (DF) (o a divisione di lunghezza d'onda). La tecnica di commutazione DF si applica attualmente con l'uso di *tecnologie fotoniche* ed è denominata *commutazione WDM* (Wavelength Division Multiplexing).

Per i sistemi di commutazione DF , possono ripetersi le stesse considerazioni svolte per i sistemi DT, con la diversità rappresentata dall'asse delle frequenze in luogo di quello dei tempi.

IV.6.4 Gestione dei sincrosegnali

Si è già osservato che una commutazione DT numerica è possibile se i flussi numerici multiplati all'ingresso e all'uscita del sistema sono tra loro sincroni. Ciò deriva dal fatto che i flussi di cifre binarie entranti e uscenti in ogni nodo a circuito numerico, così come quelli emessi e ricevuti da ogni apparecchio terminale, sono posti in una corrispondenza temporalmente trasparente.

La condizione di sincronismo può essere soddisfatta assicurando una *sincronizzazione di rete*, e cioè mettendo in atto quanto è necessario affinché le frequenze emesse dagli orologi di nodo siano in alternativa: almeno mediamente uguali (*condizione di mesocronismo*) o con scarti relativi che debbono essere contenuti entro ristrettissimi margini di tolleranza (*condizione di plesiocronismo*). Le modalità per conseguire questo obiettivo saranno illustrate in § IV.6.6.

Inoltre, occorre impiegare *organi di interfaccia* tra nodo e rete, che garantiscano, per tutti i flussi numerici entranti, condizioni *di sincronismo di cifra* e di *allineamento di trama,* quali risultano dal rispetto del vincolo di sincronismo tra i corrispondenti sincro-segnali. Infine, la temporizzazione degli apparecchi terminali e di altri apparati (multiplatori, concentratori, ecc.) nella rete di accesso deve essere *asservita* alla temporizzazione di rete.

IV.6.5 Interfacce

Per un nodo di commutazione a circuito, si distinguono (Fig. IV.6.3) l'*interfaccia di utente* (tra nodo e linea di utente) e l'*interfaccia di giunzione* (tra nodo e linea di giunzione).

Nelle attuali reti a circuito (ad es. in quelle telefoniche) la linea di utente utilizza usualmente, come mezzo trasmissivo, una coppia bifilare in rame (*doppino*) e connette un singolo impianto terminale di utente (costituito da uno o più apparecchi telefonici o telefax, apparati terminali per dati con i relativi

modem, ecc.) con l'autocommutatore di competenza. L'interfaccia tra la linea di utente e l'autocommutatore é comunemente chiamata *attacco di utente*. Si tratta di una *risorsa individuale* in quanto dedicata in modo esclusivo a ciascuna linea di utente. Presenta quindi una diffusione capillare in una rete telefonica.

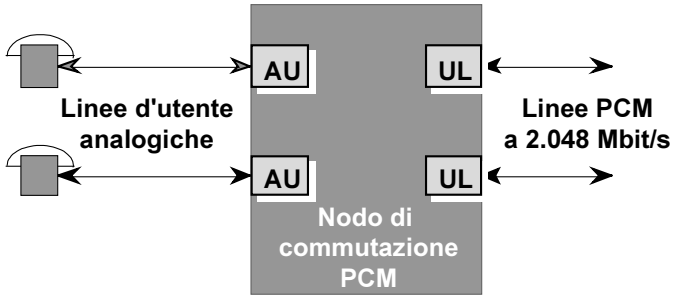


Figura IV.6.3 – Interfacce di utente (AU) e di giunzione (UL) per l'accesso a un nodo PCM

Le funzioni di un attacco di utente sono le seguenti:

- *l'alimentazione* di linea per il microfono dell'apparecchio telefonico (battery); uno schema usualmente impiegato prevede che i due fili del doppino vengano alimentati simmetricamente con una tensione, rispetto a terra, di -48 V attraverso due resistenze di 400 Ω;
- la *protezione da sovratensioni* (overvoltage protection), dovute a fenomeni atmosferici o a accoppiamenti con eventi impulsivi in impianti elettrici di potenza; ad esempio un fulmine può dare luogo a sovratensioni impulsive con ampiezze che possono raggiungere valori dell'ordine del migliaio di volt, con tempi di salita dell'ordine di alcuni microsecondi e con tempi di discesa dell'ordine del millisecondo;
- la generazione del *segnale di chiamata* (ringing); questo segnale ha la frequenza di 25 Hz e tensione di valore efficace uguale a 75 V tra i due fili del doppino;
- la *supervisione* (supervisory) dello stato della linea d'utente; ha lo scopo di sorvegliare lo stato di apertura o di chiusura del doppino e quindi di rivelare lo sgancio del microtelefono come manifestazione dell'intenzione dell'utente a dare inizio ad una chiamata;
- le *conversioni A/N e N/A* del segnale vocale, se questo é trasferito sul doppino in forma analogica; come é noto il formato del segnale numerico é usualmente quello PCM a 64 kbit/s e la funzione é quella di codifica/decodifica (codec) oggi attuabile con dispositivi individuali;
- il passaggio da due a quattro fili, tramite un dispositivo che é chiamato *forchetta* (hybrid); questa funzione è necessaria dato che un autocommutatore numerico é a quattro fili;
- l'accesso per *rilevamento di caratteristiche elettriche* (testing), da effettuare sia dal lato della linea di utente, sia da quello dell'autocommutatore; ad esempio, poiché i cavi della rete di distribuzione (sezione di accesso della rete telefonica) possono subire degradazioni delle loro caratteristiche elettriche (conduzione, isolamento, ecc.) in seguito a cause varie (infiltrazioni di umidità, ecc.), é necessario avere la possibilità di effettuare rilevamenti sulla linea di utente per riconoscere la presentazione di queste condizioni di degradazione.

Le funzioni dell'attacco di utente ora descritte sono talvolta indicate, per ragioni di sintesi, con l'acronimo BORSCHT, composto dalle iniziali delle parole *b*attery, *o*vervoltage, *r*inging, *s*upervisory, *c*odec, *h*ybrid, *t*esting.

Il secondo tipo di interfaccia che é importante considerare é quello di giunzione. Al riguardo occorre tenere conto che l'accesso a un autocommutatore PCM é concesso solo a autostrade anch'esse PCM, e cioè a linee che supportino uno schema di multiplazione statica PCM con organizzazione in trame di durata uguale a 125 μs. Si è poi già sottolineato come i segnali numerici all'ingresso e all'uscita di una rete di connessione PCM *debbano essere tra loro sincroni*. La condizione di sincronismo può essere realizzata soltanto attraverso un'opportuna interfaccia tra linee di giunzioni entranti o uscenti e l'autocommutatore. Questa interfaccia é denominata in vari modi; qui si userà il termine *unità di linea*.

Per chiarirne il funzionamento, consideriamo una linea di giunzione che si attesta al lato ricevente dell'unità di linea. Nel punto di attestazione il flusso numerico non é in generale sincrono con altri flussi numerici entranti nell'autocommutatore. Analogo asincronismo si verifica tra ognuno dei flussi entranti e ognuno dei flussi uscenti dall'autocommutatore. Relativamente a quest'ultimo, le funzioni da esso svolte sono controllate da un cronosegnale emesso da un *orologio locale* , che fornisce la temporizzazione anche ai segnali numerici emessi sulle linee uscenti. Invece, per ognuna delle linee entranti, il cronosegnale relativo ha origine da un orologio localizzato nell'apparecchiatura all'altra estremità della linea entrante (*orologio remoto*).

Per chiarezza di riferimento e con riguardo ad una specifica linea di giunzione che accede a un nodo di rete attraverso una unità di linea, chiamiamo: (i) *temporizzazione di linea* quella definita dal cronosegnale estraibile all'ingresso ricevente dell'unità di linea e avente origine dall'orologio remoto localizzato all'estremità lontana della linea di giunzione considerata; (ii) *temporizzazione di nodo* quella del cronosegnale emesso dall'orologio locale.

Per spiegare quale possa essere la relazione in frequenza e in fase tra queste due temporizzazioni, occorre tenere conto che nella frequenza istantanea $f(t)$ del cronosegnale emesso da un orologio sono distinguibili due componenti: una tempo-invariante e l'altra invece variabile nel tempo. La frequenza istantanea $f(t)$ è quindi esprimibile da :

$$f(t) = f_0 \pm \Delta f + \frac{1}{2\pi} \frac{d\psi(t)}{dt}$$

ove

- f_0 é la frequenza nominale, stabilita in sede di obiettivo progettuale ;
- Δf é lo scarto rispetto alla frequenza nominale, con il quale, ad esempio per errori di taratura, si realizza la componente di $f(t)$ che é tempo-invariante;
- $\psi(t)$ é la componente della fase istantanea del cronosegnale che costituisce scostamento rispetto a $2\pi (f_0 \pm \Delta f)t$, e cioè rispetto a un incremento linearmente crescente o decrescente nel tempo.

La derivata temporale di $\psi(t)$ rispetto al tempo esprime la rapidità di variazione di $\psi(t)$ e rende conto delle caratteristiche tempo-varianti della frequenza $f(t)$. Nelle variazioni di $\psi(t)$ si distinguono gli scostamenti *deterministici* (derive) e quelli *aleatori*. I primi sono dovuti principalmente a fenomeni di invecchiamento dei componenti dell'orologio. I secondi sono invece dipendenti da instabilità dei parametri circuitali e da variazioni termiche.

In un trasferimento trasmissivo, nel cronosegnale associato a un segnale numerico si producono ulteriori cause di fluttuazione della componente aleatoria di $\psi(t)$. Queste cause sono legate alla variazione delle caratteristiche di propagazione sul mezzo trasmissivo e alla imperfetta ricostruzione del cronosegnale nei punti di rigenerazione della catena trasmissiva (*jitter di temporizzazione*).

Supponiamo dapprima che tutti i segnali numerici all'estremità lontana delle linee di giunzione siano tra loro mesocroni. Si osserva che questa ipotesi equivale a supporre mesocroni tutti i flussi numerici che entrano nei lati riceventi delle unità di linea. Infatti, dato che ogni trasferimento trasmissivo accumula al proprio interno un numero limitato di cifre binarie e dato che non si ipotizzano

perdite di cifre binarie nel trasferimento, il ritmo binario medio in ingresso ad una linea di giunzione deve essere uguale al ritmo binario medio in uscita. Per lo stesso motivo rimangono plesiocroni flussi numerici che siano plesiocroni all'estremità lontana .

Supponiamo allora in primo luogo che all'ingresso di una unità di linea sia presente un flusso numerico che é mesocrono rispetto a tutti gli altri flussi numerici entranti nel nodo e rispetto anche ai flussi numerici uscenti. Con questa ipotesi esaminiamo le funzioni principali di una unità di linea. Svolgiamo questo esame riferendoci alla Fig. IV.6.4 e distinguendo la sezione trasmittente da quella ricevente. Per la *sezione trasmittente* sono presenti:

- l'inserimento della parola di allineamento ;
- la conversione di codice di linea, da quello interno (normalmente di tipo NRZ) a quello esterno (ad esempio di tipo HDB3); questa seconda funzione non é esplicitamente indicata nello schema a blocchi di Fig. IV.6.4.

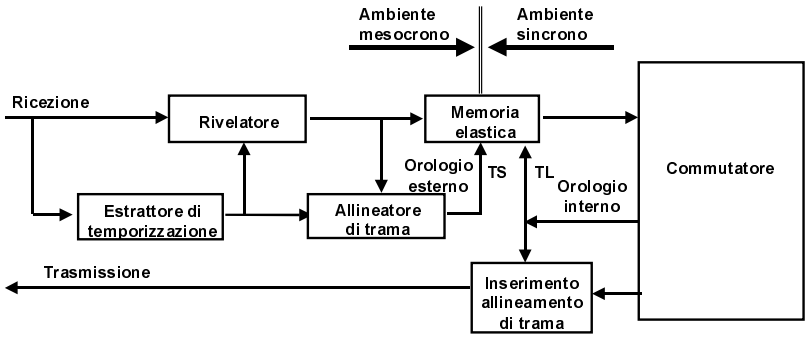


Figura IV.6.4 – Schema a blocchi di una unità di linea
(TS: temporizzazione di linea*; TL:* temporizzazione di nodo*)*

Per la sezione *ricevente* si riconoscono :

- la ricostruzione del cronosegnale esterno (temporizzazione di linea);
- la rigenerazione del segnale numerico entrante;
- la conversione di codice di linea , da quello esterno a quello interno; anche questa funzione non é esplicitamente indicata in Fig.IV.6.4;
- l'allineamento di trama (svolto dal blocco allineatore);
- il passaggio dal cronosegnale esterno a quello interno (temporizzazione di nodo); tale funzione é svolta dalla *memoria elastica;* questa é una RAM scritta con la *temporizzazione di linea TS* e letta con la *temporizzazione di nodo TL*.

L'*estrattore di temporizzazione* provvede a estrarre la temporizzazione di linea. L'*allineatore* provvede alla verifica dell'allineamento di trama e al recupero dell'allineamento corretto in caso di perdita di allineamento. La *memoria elastica* ha lo scopo di compensare le fluttuazioni di fase tra le temporizzazioni di linea e di nodo. Il cronosegnale, associato alla temporizzazione di linea e presente all'uscita dell'estrattore di temporizzazione, ha una fase di trama che è indeterminata. A eliminare questa indeterminazione provvede l'allineatore, che fornisce alla sua uscita il cronosegnale esterno, con *fasatura in trama*. Dall'orologio locale si ottiene invece il cronosegnale interno.

Un modello di memoria elastica, con significatività puramente funzionale e senza alcun riferimento a modalità realizzative, è mostrato in Fig. IV.6.5. Nel modello sono mostrate, lungo una circonferenza, le celle di memoria, ognuna capace di accumulare una cifra binaria. Si distinguono poi due puntatori, uno di *scrittura* e l'altro di *lettura.* Ogni puntatore ruota lungo la circonferenza: quello si scrittura ruota con la frequenza istantanea del cronosegnale esterno fasato in trama, mentre quello di

lettura si muove secondo la temporizzazione del cronosegnale interno. Nella figura in esame si ipotizza che i puntatori ruotino in senso orario.

La sequenza delle cifre binarie uscenti dal *rivelatore* dell'unità di linea sono scritte ordinatamente dal puntatore di scrittura nelle celle della memoria elastica e successivamente lette dal puntatore di lettura. Le operazioni di scrittura sono controllate dal cronosegnale esterno con fasatura in trama. Quindi la sequenza di cifre binarie scritte nella memoria elastica è sezionabile in trame. Le operazioni di lettura sono invece controllate dal cronosegnale interno. A valle della memoria elastica si ha quindi una sequenza di cifre binarie, che ha la temporizzazione di nodo e che è sezionabile in trame secondo la fase dell'orologio interno.

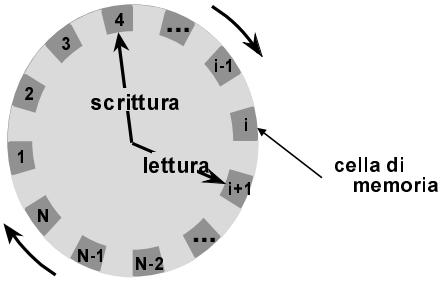


Figura IV.6.5 Modello di memoria elastica con puntatori di scrittura e di lettura

Come è ovvio, il puntatore di scrittura deve ruotare in *anticipo* rispetto a quello di lettura. Se $\Delta\varphi(t)$è questo angolo di anticipo (espresso in radianti) in un generico istante *t*, il grado di riempimento della memoria *GRM(t)* nello stesso istante è dato da $\Delta\varphi(t)/2\pi$. Le fluttuazioni di fase del cronosegnale esterno rispetto a quello interno si traducono in variazioni di *GRM(t)* (cfr. Fig. IV.6.6).

Secondo la sua definizione, *GRM(t)* è variabile nell'intervallo chiuso (0,1). Il valore *GRM(t)* = 1 corrisponde a memoria piena, mentre il valore *GRM(t)* = 0 si presenta con memoria vuota. Se *N* è la capacità della memoria, il prodotto $N \bullet GRM(t)$fornisce il ritardo (espresso in cifre binarie) che la memoria elastica introduce all'istante *t* sul flusso numerico che la attraversa. È da osservare che tale ritardo è un numero intero solo se le operazioni di scrittura e di lettura avvengono simultaneamente. Dato che questa condizione non è necessaria, il numero $N \bullet GRM(t)$ può non essere intero e quindi rappresenta solo approssimativamente il numero di bit che sono accumulati nella memoria tampone all'istante *t*.

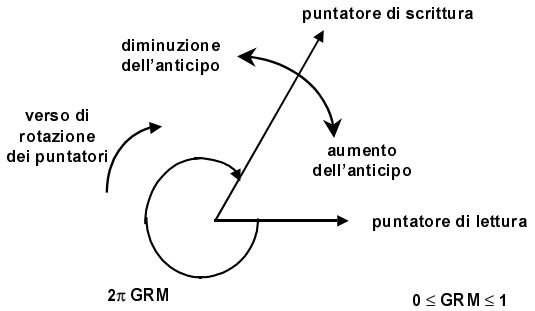


Figura IV.6.6 - Grado di riempimento GRM della memoria elastica
in relazione con la posizione relativa dei puntatori di scrittura e di lettura

Se l'anticipo della scrittura rispetto alla lettura *diminuisce* (e cioè se la scrittura rallenta rispetto alla lettura), anche *GRM(t)* diminuisce. Quando, a seguito di una diminuzione dell'anticipo, il puntatore

di scrittura si avvicina a quello di lettura nel verso contrario a quello di rotazione dei due puntatori (verso antiorario in figura), *GRM(t)* tende a 0 e la memoria *tende a svuotarsi*. Se la lettura non è distruttiva, si può verificare allora l'aggiunta di uno o più bit al flusso entrante (*slittamento per aggiunta*).

Se invece l'anticipo della scrittura rispetto alla lettura aumenta (e cioè se la scrittura accelera rispetto alla lettura), anche *GRM(t)* aumenta. Quando, a seguito dell'aumento dell'anticipo, il puntatore di scrittura si avvicina a quello di lettura nel verso di rotazione dei due puntatori, *GRM(t)* tende a 1 e cioè la memoria *tende a riempirsi*. Si può allora verificare un trabocco della memoria con la perdita di uno o più bit (*slittamento per perdita*).

Le condizioni *GRM(t)=1* e *GRM(t)=0*, che sono origine di trabocchi o di svuotamenti con conseguenti slittamenti per perdita o per aggiunta, possono ripetersi a brevi intervalli di tempo a seguito di scavalcamenti mutui multipli dei due cronosegnali. Invece la condizione *GRM(t)=1/2* (memoria riempita a metà) presenta margine massimo rispetto a riempimenti o a svuotamenti.

Per evitare le condizioni di instabilità legate a riempimenti o a svuotamenti, si può operare come segue :

- in condizioni nominali , le due temporizzazioni sono in opposizione di fase e il grado di riempimento é la metà di quello massimo (*GMR=1/2*);
- quando si verifica un riempimento o uno svuotamento della memoria, si eliminano o si aggiungono *N/2* cifre binarie in modo tale da riportare le due temporizzazioni nelle condizioni di fase nominali.

D'altra parte sia l'eliminazione che l'aggiunta di *N/2* cifre binarie comportano uno slittamento della trama di *N/2* cifre e quindi, in generale, una perdita di allineamento . Per evitare tale inconveniente é sufficiente attuare slittamenti per un numero di cifre binarie che sia uguale a un multiplo intero della lunghezza (in bit) della trama. Osserviamo però che é necessario contenere i ritardi attraverso la memoria tampone e quindi la capacità *N* di questa. Di conseguenza é conveniente effettuare slittamenti che comportino l'inserimento o la cancellazione di una singola trama.

Si conclude affermando che la memoria tampone deve avere una capacità *N* proprio uguale alla lunghezza di due trame del segnale multiplato. In queste condizioni, se la massima deviazione di fase (in anticipo o in ritardo) tra i due cronosegnali non supera

$$\Delta\phi_{max} = 2\pi\frac{N}{2}$$

e se é soddisfatta l'ipotesi di mesocronismo, non si possono verificare slittamenti incontrollati.

Riassumendo, nell'ipotesi di temporizzazioni di linea e di nodo tra loro mesocrone, la memoria elastica introduce, sul flusso numerico entrante nell'unità di linea, un *ritardo controllato* in modo da : (i) rimuovere le fluttuazioni di fase che si manifestano nel cronosegnale esterno rispetto a quello interno; (ii) allineare la fase di trama nella temporizzazione di linea con quella nella temporizzazione di nodo. Si attua quindi una conversione dall'ambiente mesocrono, che si ha esternamente all'unità di linea, a una condizione sincrona, che è essenziale per il funzionamento della rete di connessione PCM. Conseguentemente i segnali numerici sulle autostrade entranti nella rete di connessione sono tutti sincroni tra loro, con allineamento delle relative fasi di trama. Il sincronismo con relativa fasatura di trama riguarda anche la relazione dei segnali entranti con quelli uscenti dall'autocommutatore. Ciò assicura la possibilità di porre in corrispondenza trame entranti e trame uscenti in ogni stadio della rete di connessione PCM.

Sostituiamo ora l'ipotesi di mesocronismo con quella di plesiocronismo e supponiamo che la frequenza del cronosegnale esterno sia uguale a $f_0(1 \pm \varepsilon)$ e che quella del cronosegnale interno sia uguale a f_0. In questo caso la memoria tampone non può evitare il verificarsi periodico di slittamenti per

aggiunta o per perdita. È facile calcolare l'intervallo T_S di slittamento. Infatti, come illustrato in Fig. IV.6.7, si verifica riempimento o svuotamento quando $f_0 \, \varepsilon \, T_s = N/2$ e cioè per

$$T_s = \frac{N}{2\varepsilon f_0}.$$

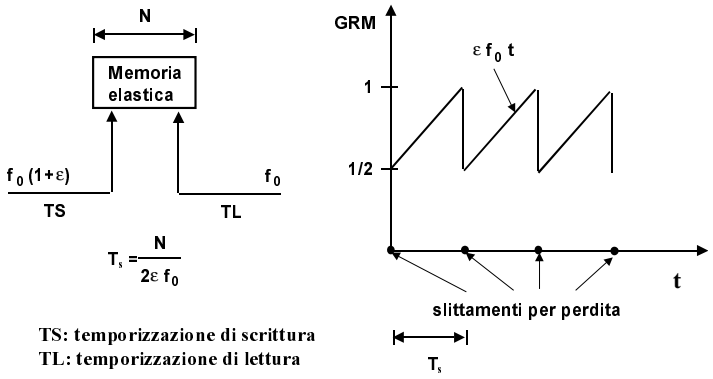


Figura IV.6.7 - Sul fenomeno degli slittamenti con cronosegnali esterno e interno tra loro plesiocroni

Conseguentemente

$$min\, T_s = \frac{N}{2 f_0 \, max\, \varepsilon} \quad ,$$

da cui si ottiene

$$max\, \varepsilon = \frac{N}{2 f_0 \, min\, T_s} \quad .$$

Ad esempio , per *N/2*= 256 (lunghezza della trama del multiplex PCM primario), per min T_S= 1 giorno = 86.400 s e per f_0= 2,048 MHz, si ottiene maxε = 256 • 10^{-6} / (86.400 • 2,048) = 1,45 • 10^{-9} . Quindi, nel caso in cui l'accesso avvenga con flussi a 2,048 Mbit/s, se si impone che gli slittamenti avvengano con un intervallo non inferiore a un giorno, la tolleranza dei cronosegnali di nodo supposti tra loro plesiocroni deve essere inferiore a circa 1,4 x 10^{-9} per un periodo sufficientemente più lungo di un giorno (ad esempio 100 giorni).

IV.6.6 Sincronizzazione di rete

In una rete fisica numerica la *sincronizzazione di rete* é la funzione che rende mesocroni o plesiocroni , alle loro origini, tutti i segnali che transitano attraverso interfacce comuni di rete. Una sincronizzazione di rete può essere attuata secondo diverse strategie, di cui le principali sono:

anarchica (o plesiocrona) in cui gli orologi di nodo sono indipendenti e i relativi cronosegnali sono plesiocroni;

dispotica (sincrona) in cui un singolo orologio (*riferimento di rete*) é distribuito a tutti nodi della rete che contengono un anello ad aggancio di fase (PLL) per sincronizzare i loro orologi interni al riferimento;

democratica (mutuamente sincrona) in cui ogni orologio di nodo é agganciato in fase agli altri orologi della rete.

Nella strategia *anarchica* ogni nodo impiega un orologio di tipo atomico con elevata precisione a lungo termine (almeno un anno per motivi di manutenzione). La stabilità annua richiesta per gli orologi

di nodo é di circa 10^{-11} e ciò conduce all'impiego di orologi atomici al cesio. Nonostante questo provvedimento, sono inevitabili eventi di perdita di informazione legati a fenomeni di slittamento incontrollato. Per ragioni di costo la strategia anarchica é oggi impiegata nella sola rete telefonica internazionale.

Per conseguire le condizioni di mesocronismo ipotizzate nel precedente paragrafo si possono impiegare gli altri due tipi di strategie di sincronizzazione di rete .

Secondo la strategia *dispotica*, detta anche "master -slave", tutti gli orologi dei centri dei nodi della rete sono direttamente asserviti, istante per istante, ad un unico orologio di elevata precisione e affidabilità, detto *orologio principale* della rete. Questa soluzione non comporta perdita di informazione e consente l'uso di orologi a bassa precisione nei vari centri di commutazione. È tuttavia necessario disporre di un'insieme di collegamenti di sincronizzazione stellari che consentano all'orologio principale di raggiungere ogni altro orologio di nodo. La conseguenza é il sorgere di problemi di affidabilità a causa dei possibili guasti di tali collegamenti e dell'orologio principale .

La strategia *democratica* prevede che ad ogni orologio di nodo sia applicato un controllo dinamico da parte di tutti gli altri orologi della rete: ciò al fine di assicurare che in regime permanente la frequenza media di tutti gli orologi della rete sia la stessa. Tale strategia non comporta quindi alcuna perdita di informazione in regime permanente e consente inoltre un elevato grado di affidabilità. Questa soluzione richiede però la costituzione di una rete di sincronizzazione prevalentemente a maglia tra i nodi e un accurato progetto dei sistemi di controllo degli orologi per garantire la stabilità asintotica delle frequenze e per limitare i fenomeni transitori conseguenti ad ogni possibile perturbazione .

IV.6.7 Caratteristiche generali di una rete a circuito

Per illustrare il trasferimento dell'informazione in una rete a circuito è comodo fare ulteriore riferimento ad una rete telefonica e alla connessione tra due generici utenti facenti capo ad autocommutatori diversi. La connessione si attua attraverso una via fisica scelta in un insieme di possibili percorsi di rete, che comprendono sia i percorsi interni agli autocommutatori, sia le linee di giunzione tra questi.

Per l'economia di sistema questo numero di percorsi di rete deve essere molto minore del numero degli utenti potenzialmente chiamanti, tenuto conto della bassa intensità media di traffico offerto da ciascun utente. Per soddisfare questa condizione, la instaurazione di una connessione telefonica richiede lo svolgimento di tre operazioni:

- quella, svolta nell'autocommutatore di origine, per trasferire la chiamata dalla linea dell'utente chiamante all'insieme dei percorsi di rete ammissibili;
- quella di impegno, per tutta la durata della comunicazione, di uno di questi percorsi di rete verso la destinazione desiderata;
- quella, svolta nell'autocommutatore di destinazione, per trasferire la chiamata dai percorsi di rete alla linea dell'utente chiamato.

Queste tre operazioni sono dette di *concentrazione*, di *distribuzione* e di *espansione* rispettivamente. La concentrazione è l'operazione di trasferimento del traffico a basso valore specifico offerto dalle linee degli utenti chiamanti verso un minor numero di linee di giunzione, caratterizzate da un elevato rendimento di utilizzazione; tale operazione si effettua in prevalenza nell'autocommutatore di origine.

La espansione è l'operazione contraria alla concentrazione, con il risultato di trasferire il traffico ad elevato valore specifico offerto dalle linee di giunzione verso il più elevato numero delle linee a cui sono connessi gli utenti chiamati; tale operazione si effettua prevalentemente nell'autocommutatore di destinazione.

La distribuzione è l'operazione di smistamento del traffico ad elevato valore specifico tra linee di giunzione entranti ed uscenti appartenenti a fasci di uguale potenzialità; tale operazione, a differenza delle precedenti, si effettua, oltre che in ambedue gli autocommutatori di origine e di destinazione, anche negli eventuali autocommutatori di transito.

Con il termine *stadio di utente* vengono denominati gli organi di connessione che effettuano entrambe le operazioni di concentrazione e di espansione necessarie affinchè ogni utente possa essere sia chiamante che chiamato. Con il termine *stadio di gruppo* vengono invece denominati gli organi di connessione che effettuano l'operazione di distribuzione e che, come tali, permettono collegamenti tra giunzioni facenti capo ad altri stadi di gruppo ovvero a stadi di utente.

Gli autocommutatori locali sono composti da stadi di utente e da stadi di gruppo. Gli stadi di utente possono avere la stessa collocazione fisica degli stadi di gruppo ovvero possono essere posti in posizione a questi remota. Gli autocommutatori di transito sono invece composti da soli stadi di gruppo.

Un modello astratto di un autocommutatore a circuito con controllo a programma memorizzato è mostrato in Fig. IV.6.8. Tale modello comprende:

- le *terminazioni* di ingresso e di uscita, che costituiscono l'interfaccia tra autocommutatore e rete; da queste possono essere estratte o inserite le informazioni di segnalazione, nel caso in cui queste ultime siano trasferite con modo associato al canale;
- una *rete di connessione*, attraverso la quale vengono effettuate le connessioni dirette ingresso-uscita;
- un *sistema di comando*, che è preposto al trattamento delle chiamate e alla gestione della rete;
- un insieme di *collegamenti di segnalazione* che, da un lato, fanno capo al sistema di comando e, dall'altro, possono fare capo, in alternativa, alle terminazioni di ingresso e di uscita ovvero ai sistemi di comando di altri autocommutatori; mentre la prima possibilità è quella adottata per la segnalazione associata al canale, la seconda fa riferimento all'impiego della segnalazione a canale comune (cfr. § V.7.1 e V.7.2).

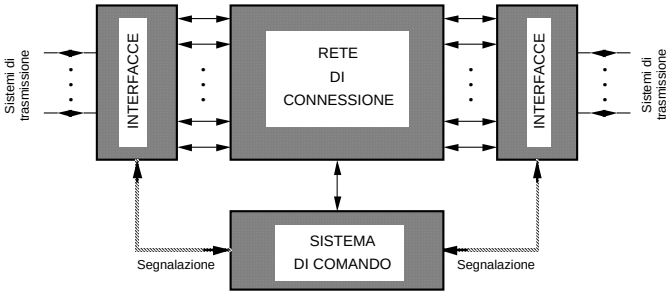


Figura IV.6.8 - Modello di un autocommutatore a circuito.

In particolare, un autocommutatore a circuito in tecnica numerica PCM (*autocommutatore PCM*) è caratterizzato dal fatto che le informazioni entranti e uscenti sono strutturate in flussi multiplati in modo statico di tipo PCM e che la rete di connessione è realizzata con tecnica a divisione di tempo numerica.

La potenzialità di un autocommutatore a circuito è descritta da due parametri prestazionali principali. Il primo di questi è la massima intensità media di traffico che può globalmente essere offerta all'autocommutatore con una prefissata *probabilità di rifiuto* per le connessioni dirette ingresso-uscita e per ognuno dei fasci di giunzione uscenti. Questo parametro rende conto dei fenomeni di contesa di

pre-assegnazione individuale relativi alle risorse di trasferimento interne ed esterne all'autocommutatore.

Il secondo parametro riguarda invece la capacità del sistema di comando (e quindi delle risorse di elaborazione del nodo) a trattare, con *vincoli di tempo reale* (cfr. § IV.6.8), i tentativi di chiamata che gli pervengono: è normalmente espresso mediante il numero massimo di *tentativi di chiamata* che possono essere trattati in un fissato intervallo temporale (ad esempio, in un'ora) con il rispetto di detti vincoli.

IV.6.8 Trattamento di chiamata

In una rete a circuito il trattamento di chiamata ha lo scopo di

- mettere a disposizione degli utenti, quando ne fanno richiesta all'inizio della comunicazione, quanto loro occorre per comunicare con altri utenti;
- supervisionare lo svolgimento della comunicazione;
- prendere atto della conclusione della chiamata.

Le funzioni di trattamento di chiamata sono svolte nei nodi di rete e possono individuarsi nei seguenti quattro tipi:

- una funzione *sensoriale* che è svolta da opportuni organi, inseriti sulle terminazioni di ingresso e di uscita; questi provvedono a rivelare la presentazione di una richiesta di servizio, a sorvegliare le modalità di soddisfacimento di questa richiesta, nonché ad accertare il completamento della fornitura del servizio richiesto;
- una funzione *di memoria*, che consente di immagazzinare sia i *dati* informativi necessari per l'attuazione di un servizio specifico, sia i *programmi* preposti a regolare le modalità di azione dell'autocommutatore a fronte delle richieste di servizio che gli vengono presentate;
- una funzione *decisionale*, che si attua attraverso organi di elaborazione che operano le scelte del caso in base a informazioni provenienti sia dall'esterno che dall'interno dell'autocommutatore; le prime sono ricevute direttamente o tramite la funzione sensoriale; le seconde sono quelle relative ai dati e ai programmi contenuti in memoria;
- una funzione *di attuazione*, che consente l'esecuzione delle azioni definite dagli organi decisionali.

Il trattamento di chiamata è svolto in ogni nodo di rete dal sistema di comando in cooperazione con le terminazioni di ingresso/uscita e con la rete di connessione; il segnale controllante è rappresentato dall'informazione di segnalazione.

Per esemplificare, seppure parzialmente, i compiti connessi al trattamento di chiamata, possiamo fare riferimento alla cosiddetta *fase di pre-selezione* in un autocommutatore telefonico, e cioè a quella fase del trattamento di una chiamata telefonica, che è compresa tra lo sgancio del microtelefono e l'invio all'utente chiamante dell'*invito a selezionare*. In questa fase devono essere svolte nell'autocommutatore varie operazioni, e cioè

- la rivelazione dello sgancio;
- l'impegno di un registro;
- l'analisi della categoria dell'abbonato;
- la ricerca di un ricevitore di selezione libero;
- l'impegno di un ricevitore di selezione;
- la ricerca di un percorso libero, all'interno dell'autocommutatore, tra la terminazione e il ricevitore di selezione;
- la connessione dell'utente al ricevitore di selezione;
- l'invio dell'invito a selezionare.

Nello svolgere la funzione di trattamento di chiamata, il sistema di comando di un autocommutatore telefonico deve rispondere alle richieste dell'utenza nel tempo più rapido possibile: deve cioè soddisfare *requisiti di tempo reale*.

Si possono distinguere più livelli di tempo reale. Sono più severi quelli legati al trattamento dell'informazione di segnalazione. Per questo trattamento è infatti richiesto un ritardo massimo di risposta dell'ordine di una decina di millisecondi. Un secondo livello di requisiti di tempo reale riguarda il resto dei programmi di trattamento di chiamata. I ritardi di risposta alle azioni degli utenti debbono essere globalmente inferiori a valori dell'ordine del secondo, affinché il servizio offerto possa essere giudicato soddisfacente. Con riferimento ad esempio, alla fase di pre-selezione e alle sue operazioni tipiche sopra elencate, ciascuna di queste deve effettuarsi in media, in meno di un centinaio di millisecondi. Questo ordine di grandezza (da qualche decina a un centinaio di millisecondi) si applica conseguentemente a tutte le azioni riguardanti un trattamento di una chiamata telefonica.

IV.7 Nodi a pacchetto

Nel termine "*commutatore a pacchetto"* vengono qui fatti rientrare tutti i dispositivi di rete che, operando in un modo di trasferimento orientato al pacchetto per comunicazioni mono- o multi-mediali, svolgono una funzione di rilegamento.

Le componenti essenziali di un nodo a pacchetto, nell'ipotesi che le contese di utilizzazione siano risolte a ritardo, includono un *buffer di ingresso*, un processore preposto al trattamento dei pacchetti (*processore di nodo*) e un buffer per ognuna delle linee uscenti (*buffer di uscita*). Il buffer di ingresso memorizza (completamente o parzialmente) i pacchetti entranti per consentirne il trattamento. Il processore di nodo tratta l'intestazione dei pacchetti in cui è collocata la pertinente informazione di controllo protocollare. Ogni buffer di uscita memorizza i pacchetti per consentire loro l'accesso alla relativa linea di uscita, secondo quanto richiesto da una multiplazione dinamica con contese risolte a ritardo.

I nodi a pacchetto trattano in generale tre tipi di traffico (Fig. IV.7.1):

- un *traffico originario*, proveniente da apparecchi terminali (ad es. da DTE) siti localmente in prossimità del nodo considerato;
- un *traffico terminale*, proveniente da altri nodi nella rete e indirizzato ad apparecchi terminali (ad es. a DTE) siti localmente in prossimità del nodo considerato;
- un *traffico di transito*, proveniente da altri nodi della rete e instradato verso ulteriori altri nodi nella rete.

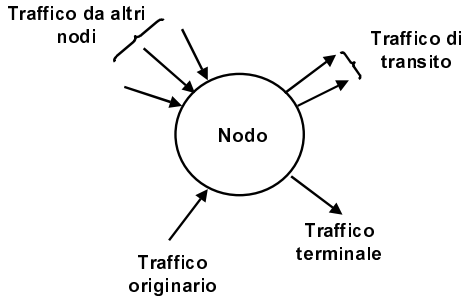


Figura IV.7.1 – Tipi di traffico trattati da un nodo a pacchetto

Nei confronti di queste componenti di traffico, un nodo a pacchetto che tratta connessioni commutate può essere caratterizzato mediante due insiemi di parametri prestazionali. Dal punto di vista del trattamento delle chiamate che gli vengono presentate e quindi della sua capacità nel risolvere le

contese di pre-assegnazione, i parametri prestazionali da utilizzare sono analoghi a quelli descritti nel caso di un nodo a circuito (cfr. § IV.6.7). Dal punto di vista, invece, della sua capacità di trattare le UI che lo attraversano, un parametro comunemente impiegato è la *portata media* del nodo, espressa ad esempio, in UI che transitano mediamente nell'unità di tempo. La caratterizzazione di nodi con connessione semi-permanente o senza connessione riguarda solo il secondo punto di vista.

IV.7.1 Processore di nodo

Nel caso di un nodo operante in un servizio di rete *con connessione* (commutata o semi-permanente), in corrispondenza della ricezione di un pacchetto durante la *fase di trasferimento*, il processore di nodo:

- deve cooperare, per quanto di sua competenza, allo svolgimento delle funzioni di controllo protocollare definite nell'intestazione del pacchetto;
- deve leggere il numero di canale logico entrante che è contenuto nell'intestazione;
- in base alla consultazione della tabella di attraversamento (cfr. § IV.3.2), deve trovare il numero di canale logico uscente, che è in corrispondenza con quello entrante;
- deve modificare, nell'intestazione del pacchetto, il numero di canale logico entrante in quello uscente (traduzione di etichetta);
- deve inoltrare il pacchetto così modificato verso il ramo di uscita, che è in corrispondenza con il numero di canale logico uscente.

Nel caso specifico di un nodo operante in un servizio di rete con *connessione commutata*, in corrispondenza dell'inizializzazione di una nuova comunicazione, il processore di nodo:

- deve trattare l'informazione di segnalazione, che riceve sotto forma di particolari pacchetti e che gli è trasferita dagli apparati di rete (apparecchi terminali o altri nodi) a monte lungo il costituendo percorso da estremo a estremo;
- sulla base dell'indirizzo di destinazione della comunicazione contenuto nei pacchetti di segnalazione e svolgendo un algoritmo di instradamento, deve decidere *quale sia il ramo* che deve concorrere al percorso di rete da estremo a estremo;
- applicando poi un opportuno *criterio di assegnazione* (cfr. § IV.7.3), deve decidere se assegnare o meno un canale logico sostenuto da quel ramo;
- deve *riempire la tabella di attraversamento* con il numero di canale logico entrante (contenuto nel pacchetto di segnalazione) e con quello di canale logico uscente (oggetto della decisione di instradamento da parte del processore).

Nel caso di un nodo operante in un modo di trasferimento *senza connessione*, in corrispondenza della ricezione di un pacchetto, il processore di nodo:

- deve cooperare, per quanto di sua competenza, allo svolgimento delle funzioni di controllo protocollare definite nell'intestazione del pacchetto;
- deve leggere il campo di indirizzo contenuto nell'intestazione del pacchetto;
- sulla base degli indirizzi di origine e di destinazione e svolgendo un algoritmo di instradamento, deve decidere quale sia il ramo verso cui instradare il pacchetto.

IV.7.2 Instaurazione di un circuito virtuale

Tra le funzioni svolte al servizio di una comunicazione inizializzata a chiamata, ne consideriamo due di particolare rilievo: l'*instaurazione di un circuito virtuale* e l'*accettazione di chiamata*. La prima di queste funzioni mira a rendere disponibile una connessione commutata al servizio della chiamata. La seconda è l'applicazione del criterio che consente di accettare o meno una nuova prenotazione di

accesso: la richiesta di prenotazione è avanzata da un utente del servizio di rete e fa riferimento alle risorse virtuali di un nodo.

Cominciamo dalla modalità di instaurazione di un circuito virtuale, rinviando l'accettazione di chiamata a § IV.7.3. Indichiamo con

Y_i il nodo a cui ci riferiamo per descrivere la funzione di *instaurazione*;

Y_{i-1} l'apparecchiatura (terminale di origine o altro nodo), che è collocata a monte di Y_i sul circuito virtuale in corso di instaurazione.

Per la chiamata considerata, il canale logico entrante in Y_i è preassegnato da Y_{i-1} e fa parte dell'insieme di canali logici che sono pre-assegnabili per la multiplazione dinamica, sul ramo di rete che connette Y_{i-1} con Y_i (*ramo entrante in* Y_i) . Il canale logico uscente da Y_i è pre-assegnato da quest'ultimo a seguito della ricezione di un *messaggio di segnalazione*, che richiede l'instaurazione del circuito virtuale. Tale pre-assegnazione è attuata attraverso i seguenti passi:

- conoscendo il nodo di destinazione (precisato dall'informazione di segnalazione), viene applicato un *algoritmo di instradamento* e, sulla base dei risultati da questo forniti, viene individuato un ramo uscente da Y_i; tale ramo concorrerà a far parte del percorso di rete da utente a utente;
- nell'insieme dei canali logici pre-assegnabili per la multiplazione dinamica su questo ramo, ne viene prescelto uno tra quelli liberi da precedenti impegni.

I due canali logici, quello entrante e quello uscente, sono identificati da due numeri d'ordine, scelti in due insiemi di numerazione univoca. Il *numero di canale logico entrante* è univoco nell'insieme utilizzato da Y_{i-1} per il ramo entrante in Y_i ed è comunicato a quest'ultimo tramite il messaggio di segnalazione, che richiede l'instaurazione del circuito virtuale. Il *numero di canale logico uscente* è univoco nell'insieme di numerazione utilizzato da Y_i per individuare i canali logici attivabili su tutti i rami da esso uscenti. In tal modo, a conclusione della fase di instaurazione, ogni nodo, che appartiene al corrispondente percorso di rete, può memorizzare, nella *tabella di attraversamento* e in corrispondenza al ramo entrante che compone quel percorso, una *coppia di numeri*; di questi:

- il *primo* identifica il canale logico pre-assegnato per il trasferimento dei pacchetti scambiati nell'ambito di quella chiamata e entranti nel nodo;
- il *secondo* identifica il canale logico pre-assegnato per il rilancio dei pacchetti scambiati nell'ambito di quella stessa chiamata e uscenti dal nodo.

Ciò equivale a dire che il circuito virtuale instaurato per una particolare chiamata è descritto da una *sequenza di numeri di canale logico*. Gli elementi, che compongono questa sequenza, sono *memorizzati* nelle apparecchiature di rete per tutta la durata della chiamata fino alla conclusione della fase di abbattimento.

IV.7.3 Accettazione di chiamata

L'accettazione è il criterio che consente di accettare o meno una nuova prenotazione di accesso da parte di una chiamata che ne fa richiesta. Si tratta della funzione svolta dal processore di un nodo a pacchetto e rientra tra gli adempimenti relativi al trattamento di chiamata. Il criterio di accettazione si basa su dati forniti da ogni chiamata richiedente e su una loro elaborazione effettuata dagli organi preposti alla decisione.

Come già detto in precedenza (cfr. § III.2.2), si fa riferimento a:

- i *dati di traffico preesistente* e i relativi *requisiti prestazionali* per ognuna delle comunicazioni, le cui richieste di accesso sono già state accolte in precedenza;
- dati analoghi relativi alla comunicazione la cui richiesta è oggetto della decisione.

Allora l'elaborazione consente di valutare: lo *stato di carico di riferimento* sulla base dei dati preesistenti e lo *stato di carico di riferimento modificato* sulla base dei dati richiesti. Alla valutazione

dei due stati segue *una valutazione delle prestazioni*. Se i gradi di accessibilità per le chiamate già accolte e per quella presentante la nuova richiesta rispettano i requisiti prestazionali sia del pregresso che dell'attuale, l'autorizzazione di accesso viene *accordata*. In caso contrario viene *negata*.

Tra i criteri di accettazione di chiamata, se ne descrivono qui di due tipi: l'*assegnazione a domanda media* e l'*assegnazione a domanda di picco*, che introduciamo con riferimento a un ambiente in cui:

- la sorgente *k*-esima (k = 1, 2…..) tra quelle che concorrono per una autorizzazione di accesso sia caratterizzata da un ritmo binario di picco R_{pk} e da un grado di intermittenza B_k ;
- il canale multiplato abbia capacità di trasferimento C_m .

Nell'*assegnazione a domanda media,* se le contese di utilizzazione sono risolte con *modalità a ritardo*, l'obiettivo prestazionale riguarda il valor medio $E[\Xi]$ del *ritardo di attraversamento* Ξ che una UI deve subire per accedere al ramo uscente. Questo valor medio $E[\Xi]$ non deve essere superiore ad una soglia prefissata Ξ_{max} ; cioè

$$E[\Xi] \leq \Xi_{max} .$$

Per l'applicazione di questo criterio si suppone che lo stato di utilizzazione della capacità di trasferimento del ramo uscente sia completamente definito in *termini medi* tramite il valore assunto dal rendimento di utilizzazione U e che esista un legame tra $E[\Xi]$ e U , in cui $E[\Xi]$ è una funzione monotona crescente di U . Siano

U_0 il dato rappresentativo (in termini medi) dello stato di *carico di riferimento*;

U_1 il dato rappresentativo dello stato di *carico di riferimento modificato*.

Nell'ipotesi allora che la nuova richiesta sia presentata dalla sorgente *k*-esima, risulta

$$U_1 = U_0 + \Delta U \quad ,$$

in cui

$$\Delta U = \frac{R_{pk}}{B_k C_m} \quad .$$

La regola di decisione può essere così applicata

- si valuta il valor medio $E[\Xi_1]$ del ritardo di attraversamento che si avrebbe in corrispondenza di $U = U_1$ e si confronta $E[\Xi_1]$ con Ξ_{max} ;
- se $E[\Xi_1] \leq \Xi max$, la nuova richiesta può essere accolta; in caso contrario deve essere rifiutata.

Nell'*assegnazione a domanda di picco* e sempre nell'ipotesi che le contese di utilizzazione siano risolte a ritardo, in ogni stato di utilizzazione della capacità di trasferimento del ramo uscente, deve essere emulato il comportamento prestazionale di commutazione a circuito. Deve cioè essere assicurato il massimo grado di trasparenza temporale. Per perseguire questo obiettivo, si impone che il ritmo binario di picco totale autorizzato ad accedere al ramo uscente non superi la capacità di trasferimento di quest'ultimo. Siano

R_{pT0} il ritmo binario totale di picco che viene assunto come stato di *carico di riferimento* e che è dato dalla somma dei ritmi binari di picco che caratterizzano le chiamate in corso di evoluzione;

R_{pT1} il ritmo binario totale di picco che viene assunto come stato di *carico di riferimento modificato* (quale risulterebbe se la nuova richiesta di autorizzazione fosse accolta).

Nell'ipotesi che la nuova richiesta sia presentata dalla sorgente *k*-esima (k = 1, 2…), risulta

$$R_{pT1} = R_{pT0} + R_{pk} .$$

La regola di decisione è allora la seguente:

- si confronta R_{pT1} con C_m ;
- se $R_{pT1} \leq C_m$, la nuova richiesta può essere accolta; in caso contrario deve essere rifiutata.

IV.7.4 Orologi di nodo

Nel caso di nodi a pacchetto, possono essere impiegate due relazioni di temporizzazione:

- quella sincrona, in cui, come nel caso dei nodi a circuito numerici, è richiesta una completa sincronizzazione di rete;
- quella asincrona, in cui non c'è necessità di sincronizzazione di rete, in quanto la temporizzazione dei flussi di cifre binarie entranti può essere adattata all'orologio del nodo senza perdita di informazione.

Il funzionamento asincrono può essere utilizzato se le UI sono completamente memorizzate entro il nodo, elaborate e successivamente inoltrate verso la loro destinazione. In questo caso, le memorie di ingresso e/o di uscita, se hanno capacità adeguata, costituiscono una separazione tra le temporizzazioni dei flussi entranti e di quelli uscenti. Questa separazione fornisce la possibilità di un adattamento fra la temporizzazione del flusso entrante e quella dell'orologio di nodo; che a sua volta è disposta ad emettere il sincro-segnale a supporto del flusso uscente.

L'impiego del funzionamento sincrono può essere richiesto quando si opera con memorizzazioni parziali delle UI, avendo lo scopo di ridurre il ritardo di attraversamento. Un esempio significativo è offerto dall'applicazione del metodo del "*cut-through*" (cfr. par. IV.3).

V FUNZIONI DI RETE

Fra le svariate funzioni svolte per scopi di trasferimento nell'ambito della fornitura di servizi di rete, questo capitolo tratta quelle di maggior rilievo, e cioè *l'accesso multiplo* (par. V.1), il *controllo d'errore* (par. V.2), *l'indirizzamento* (par. V.3) e *l'instradamento* (par. V.4), queste due ultime funzioni in una *inter-rete* (par. V.5.) e infine il *controllo del traffico e della congestione* (par.V.6).

V.1 Accesso multiplo

Abbiamo già più volte incontrato i mezzi di comunicazione, come sostegno di trasferimenti di informazione a distanza. Qui il concetto viene ripreso per introdurre un argomento a cui si è già accennato, e cioè quello riguardante l'*accesso multiplo*.

Un mezzo di comunicazione consente il trasferimento dell'informazione tra due o più punti topologicamente a distanza; supporta per questo scopo la via fisica (*canale trasmissivo*) su cui si propagano i segnali informativi; può essere *punto-punto* o *multi-accesso*:

- è *punto-punto* quando consente trasferimenti da una sola estremità emittente ad una sola ricevente (*mezzo mono-accesso*); in questo caso il segnale ricevuto dipende *unicamente* dal segnale trasmesso e dal disturbo su quel canale trasmissivo;
- è *multi-accesso* quando comprende, come in Fig. V.1.1, due o più sistemi terminali distinti che sono sorgenti e/o collettori di informazione e che sono indicati tradizionalmente come *stazioni*; in questo caso il segnale ricevuto in una stazione dipende dal segnale trasmesso da due o più tra le altre stazioni ed è la somma delle versioni attenuate di questi segnali, corrotti da disturbi e da ritardi.

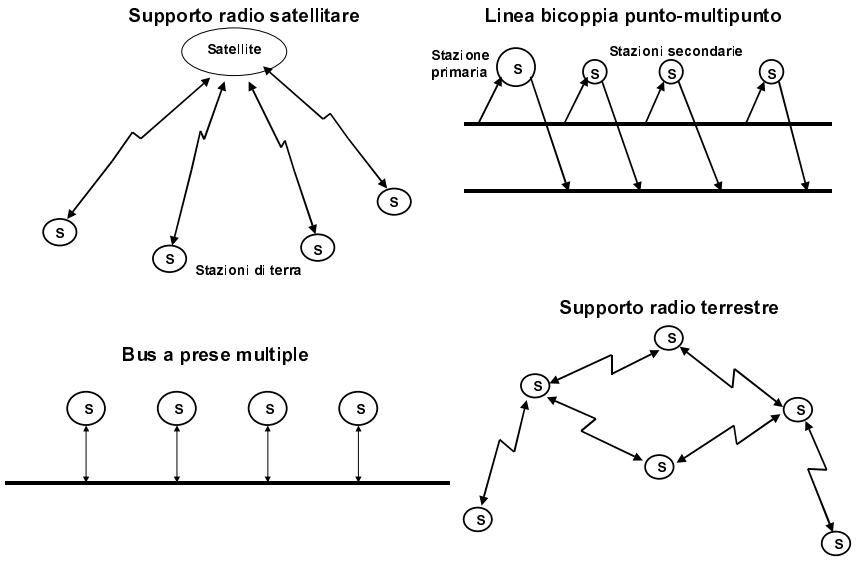


Figura V.1.1 – Esempi di mezzi di comunicazione multi-accesso (S: stazione)

La *funzione di accesso* multiplo riguarda i mezzi di comunicazione multi-accesso. Come indicato in Fig. V.1.2, si distinguono accessi multipli *con allocazione statica* e *con allocazione dinamica*. Nel caso di allocazione statica, il mezzo sostiene una molteplicità di comunicazioni con una suddivisione in sub-canali, ognuno dei quali è *pre-assegnato individualmente* alle necessità di trasferimento di una particolare comunicazione. Nel caso invece di una allocazione dinamica, il mezzo viene trattato come

una risorsa singola a cui si può accedere in accordo a una opportuna procedura di controllo: la banda disponibile è allora *assegnata a domanda* o *pre-assegnata collettivamente*. Come è evidente, esiste una sostanziale analogia tra queste due modalità e quelle omonime riguardanti la *funzione di multiplazione*.

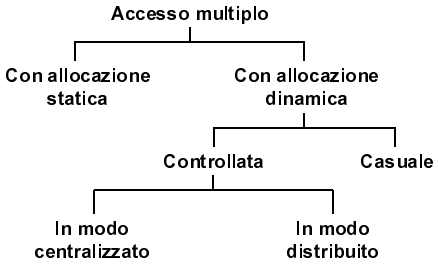


Figura V.1.2 – Alternative di accesso multiplo

Il paragrafo fornisce dapprima qualche ulteriore precisazione sull'accesso multiplo con allocazione statica (§ V.1.1) e con allocazione dinamica (§ V.1.2). Le parti successive del paragrafo sono tutte dedicate all'accesso multiplo con allocazione dinamica. In particolare:

- si fornisce una classificazione delle procedure di controllo (*protocolli*) per l'accesso al mezzo (§ V.1.3);
- si affrontano gli aspetti prestazionali che vengono considerati per caratterizzare un sistema ad accesso multiplo (§ V.1.4);
- si parla poi dei protocolli ad accesso casuale, con riferimento dapprima a quelli operativamente più semplici (§ V.1.5) e successivamente ad altri con obiettivi di migliori prestazioni in termini di smaltimento di traffico (§ V.1.6);
- si conclude con l'esame dei protocolli ad accesso controllato (§ V.1.7).

V.1.1 Accesso multiplo con allocazione statica

Il caso di accesso multiplo con allocazione statica è stato risolto nel tempo con modalità varie, che si distinguono in base al *dominio,* che consente di suddividere in *sub-canali* la capacità di trasferimento del mezzo di comunicazione.

Una prima modalità di allocazione statica è quella nel *Dominio della Frequenza* (*FDMA*: Frequency Division Multiple Access). Ad ogni stazione è *pre-assegnata individualmente* una specifica sotto-banda di frequenze (*banda di stazione*) del canale trasmissivo. Le bande di stazione sono affiancate nell'intervallo di frequenze che è passante nel canale trasmissivo e intervallate con *bande di guardia.* La multiplazione che così si attua è quindi a divisione di frequenza (Fig. V.1.3).

Una seconda modalità di accesso multiplo con allocazione statica è nel *Dominio del Tempo* (*TDMA*: Time Division Multiple Access). Ad ogni stazione è pre-assegnato individualmente uno specifico *intervallo temporale* (IT) del canale trasmissivo. Gli IT sono affiancati sull'asse temporale, con eventuali *intervalli di guardia*, e *organizzati in trama*. La stazione utilizza l'IT assegnatole con periodicità di trama. La multiplazione che così si attua è quindi a divisione di tempo (Fig. V.1.4). L'asse dei tempi è gestito con *modalità SF* ed è trattato con *pre-assegnazione individuale*. Si presentano quindi solo contese di pre-assegnazione.

Una terza tecnica è quella nel *Dominio del Codice* (*CDMA*: Code Division Multiple Access). Ad ogni stazione è *pre-assegnato individualmente* uno specifico codice di *m* bit (*sequenza di chip*). Per emettere un "1" la stazione emette la sua sequenza di chip, mentre per emettere uno "0" emette il complemento a uno di tale sequenza.

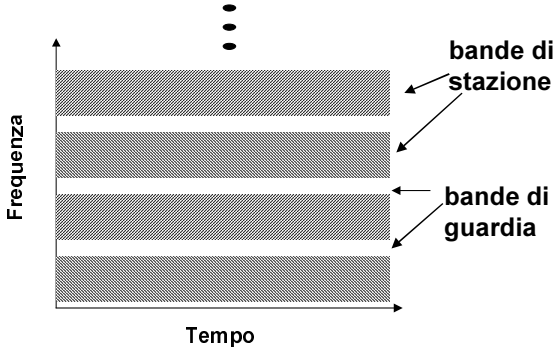


Figura V.1.3 – Principio della tecnica FDMA

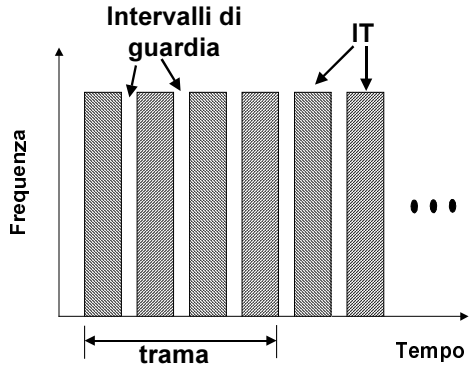


Figura V.1.4 – Principio della tecnica TDMA

V.1.2 Accesso multiplo con allocazione dinamica

Per l'accesso multiplo con allocazione dinamica, ci limitiamo a considerare le sue attuazioni nel dominio del tempo. Come nell'omonima tecnica di multiplazione, l'asse dei tempi è gestito con le *modalità U, SF* o *SU*. La risorsa è normalmente *assegnata a domanda* e le contese di utilizzazione possono essere risolte con due modalità alternative (cfr. Fig. V.1.2): l'accesso *casuale* e quello *controllato*.

Nel caso di accesso casuale, ogni stazione può iniziare a emettere una unità informativa (UI) senza alcun coordinamento con le altre stazioni. Ciò implica la presentazione di contese di utilizzazione del mezzo condiviso. Questi eventi, chiamati *collisioni*, si manifestano quando due o più stazioni emettono UI che impegnano *contemporaneamente* il canale supportato dal mezzo. Allora, per effetto della mutua interferenza tra i relativi segnali, il contenuto informativo di queste UI viene corrotto e può risultare non utilizzabile dal punto di vista delle applicazioni. Per questo motivo occorre prevedere meccanismi, che garantiscano il corretto trasferimento delle informazioni anche se questi conflitti accadono.

Nel caso di accesso controllato, invece, ogni stazione può emettere solo quando riceve una specifica autorizzazione. Viene così evitato l'accesso contemporaneo e quindi non si verificano collisioni. La gestione dell'autorizzazione a emettere può (cfr. Fig. V.1.2) essere demandata ad una stazione speciale (*controllo centralizzato*), oppure può essere condivisa fra tutte le stazioni del sistema (*controllo distribuito*). In questo secondo caso il controllo passa ordinatamente da stazione a stazione. Le stazioni inattive (senza cioè esigenze di emissione) non impegnano risorse. Si consegue così una

efficiente ripartizione della capacità di trasferimento del mezzo di comunicazione fra le sole stazioni attive.

In tutti i tipi di accesso multiplo con allocazione dinamica la regolazione dell'accesso al mezzo, con la risoluzione di eventuali contese di utilizzazione tra le stazioni, è affidata a protocolli di accesso al mezzo, e cioè ai *protocolli MAC* (Medium Access Control).

L'accesso multiplo con allocazione dinamica è utilizzato, ad esempio, nelle reti in area locale (LAN) con struttura sia cablata (wired-LAN), che su portante radio (wireless-LAN): in entrambi questi casi si utilizzano infatti mezzi di comunicazione multi-accesso. In Fig. V.1.5 sono mostrate alcune topologie di LAN con struttura cablata. Da un punto di vista logico si distinguono sostanzialmente il *bus* (bidirezionale o unidirezionale) e l'*anello*.

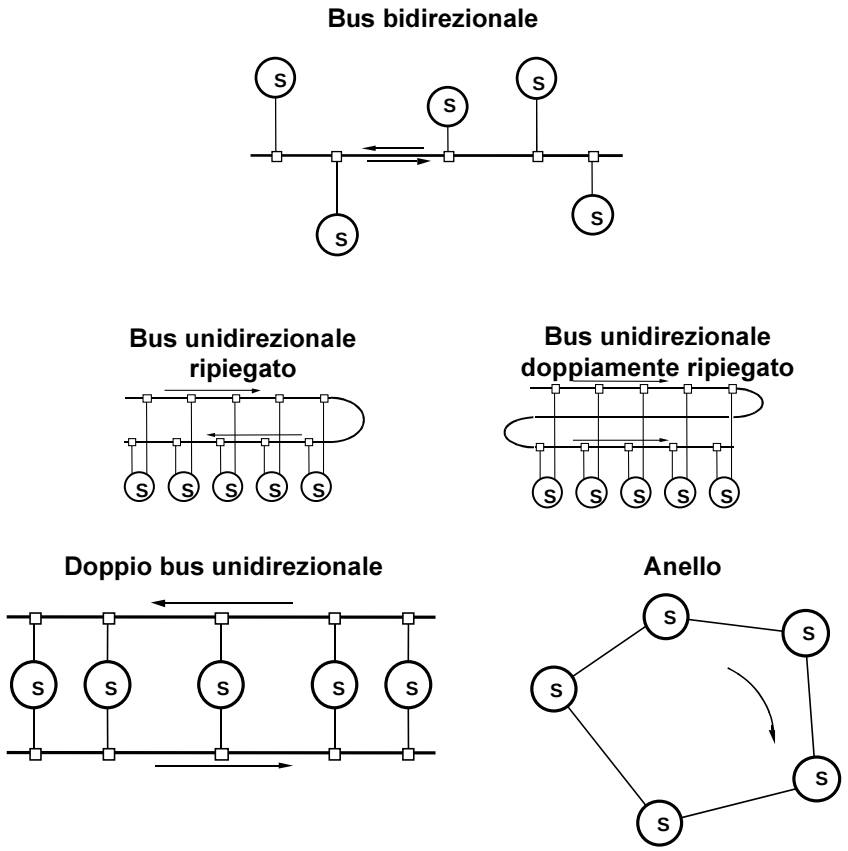


Fig. V.1.5 – Esempi di topologie di LAN a struttura cablata (S: Stazione)

V.1.3 Protocolli MAC

Come già detto in § V.1.2, i protocolli MAC possono essere raggruppati in due categorie:

- *protocolli ad accesso casuale*;
- *protocolli ad accesso controllato*.

Nell'accesso casuale, ogni stazione emette quando ha una UI pronta e dopo aver eventualmente accertato, nei limiti delle sue possibilità di verifica, lo stato di occupazione del mezzo di comunicazione. Se il mezzo è effettivamente "libero", l'emissione della UI ha successo. Se invece il mezzo è già occupato e questo stato non è stato accertato dalla stazione emittente, si verifica collisione,

con possibile perdita dell'informazione contenuta nella UI e con conseguente possibile necessità di ri-emissione. Se il traffico offerto alla rete aumenta, cresce anche la probabilità di collisione. Ciò può limitare il traffico utile che è mediamente smaltibile dal sistema e può essere causa di instabilità nel funzionamento della rete. Il verificarsi di eventi di collisione caratterizza quindi i protocolli ad accesso casuale, che conseguentemente sono anche detti *protocolli MAC con collisione*.

Nell'accesso controllato, invece, non sono possibili collisioni. Per tale ragione questa classe di protocolli MAC è anche detta *senza collisioni*. Come si è già visto, in questo ambito si hanno protocolli MAC con controllo centralizzato e altri con controllo distribuito. Un esempio di controllo centralizzato è fornito dal protocollo *ad interrogazione* (polling), mentre esempi di controllo distribuito sono offerti dal protocollo *a testimone* e da quello *a coda distribuita*. Nel seguito ci occuperemo esclusivamente dei protocolli a controllo distribuito.

V.1.4 Aspetti prestazionali

Per qualificare prestazionalmente una funzione di accesso multiplo, soprattutto nel caso di accessi con allocazione dinamica, si utilizzano normalmente due parametri:

- la *portata* della struttura di accesso, e cioè il numero di cifre binarie che la struttura è in grado di trasferire con successo nell'unità di tempo; ogni valore di portata è in corrispondenza con un valore di *carico*, e cioè di numero complessivo di cifre binarie che tentano l'accesso nell'unità di tempo e che includono in generale, accanto alle cifre trasferite con successo, anche quelle soggette a insuccesso;
- il *ritardo di trasferimento*, e cioè l'intervallo di tempo tra l'istante di generazione di una unità informativa alla stazione di origine e l'istante in cui questa UI è correttamente ricevuta dalla stazione di destinazione.

Nella determinazione di questi parametri, si considerano usualmente solo i contributi legati alla gestione dell'accesso multiplo. Inoltre portata, carico e ritardo di trasferimento sono variabili in generale caratterizzabili solo in termini probabilistici e sono normalmente valutate attraverso i loro momenti (in particolare attraverso i loro valori medi ed eventualmente le loro varianze), nell'ipotesi che sussistano condizioni di equilibrio statistico tali da assicurare l'indipendenza di questi momenti dall'istante di osservazione del sistema di accesso e dalle condizioni iniziali della sua evoluzione. Infine portate e carichi medi sono usualmente espressi in *forma normalizzata* assumendo come riferimento la capacità di trasferimento del mezzo di comunicazione (cfr. § III.1.1). Anche il valore medio del ritardo di trasferimento è usualmente espresso in termini normalizzati: in questo caso il riferimento è la durata media di una unità informativa nel suo trasferimento attraverso il sistema multi-accesso.

Si osserva che la *portata media normalizzata*, nel seguito denotata con la lettera S, coincide in questo contesto con il *rendimento di utilizzazione* del mezzo di comunicazione. Trattasi quindi di una quantità che non è mai superiore all'unità e che esprime l'*intensità media del traffico smaltito* dalla struttura di accesso. Corrispondentemente il *carico medio normalizzato,* nel seguito indicato con il simbolo G, può essere interpretato come *intensità media di traffico offerto alla rete*: i suoi valori possono essere considerati come variabili indipendenti, con la possibilità di essere superiori all'unità e con l'unica limitazione imposta dall'eventuale esigenza di non sollecitare il sistema di accesso verso condizioni di instabilità.

I parametri S e G e la variazione del primo parametro in funzione del secondo sono spesso utilizzati per indicare in quale misura un sistema multi-accesso smaltisce mediamente il traffico che gli viene offerto, sempre in termini medi, dalle stazioni accedenti.

Circa i ritardi di trasferimento si possono sottolineare due punti importanti per la caratterizzazione di uno specifico sistema multi-accesso:

- un primo punto riguarda come il valor medio del ritardo di trasferimento si comporta in funzione della portata media; come è intuitivo, massimizzare la portata media e minimizzare il valor medio del ritardo di trasferimento potrebbero apparire obiettivi attraenti, ma in generale sono anche obiettivi contrastanti;
- un secondo punto si riferisce al comportamento della varianza del ritardo di trasferimento sempre in funzione della portata media; per un fissato valore di *S*, il valore assunto da questa varianza è una possibile misura del grado di trasparenza temporale che caratterizza il sistema multi-accesso considerato nelle fissate condizioni di smaltimento di traffico.

Una misura dell'efficienza di un protocollo di accesso è fornita dall'*utilizzazione U del canale:* questa rappresenta la massima portata media normalizzata che la rete è in grado di smaltire e, coerentemente con quanto detto in precedenza, è definita come il rapporto fra la massima intensità di traffico smaltito dalla rete e la capacità di trasferimento del mezzo di comunicazione.

Per mostrare come la condivisione del mezzo influenza i valori di utilizzazione U del canale, ricaviamo una espressione di questo parametro nel caso di un *accesso perfetto,* in cui l'impegno del canale nel trasferimento di una UI è determinato solo dal tempo di trasmissione della UI e dal ritardo di propagazione tra le stazioni di origine e di destinazione. Ciò consentirà di mettere in evidenza quali siano i vincoli che agiscono su U indipendentemente da specifici meccanismi di accesso. A tale scopo, con riferimento ad un accesso multiplo con allocazione dinamica, indichiamo con:

R la *capacità di trasferimento* (in bit/s) del mezzo di comunicazione;
D il *massimo ritardo di propagazione* (in s) fra due stazioni che accedono al mezzo;
L la lunghezza (in bit) di una UI tipica che transita sul mezzo di comunicazione;
α la *lunghezza del canale trasmissivo* normalizzata rispetto alla lunghezza di una UI tipica.

Osserviamo che il prodotto RD rappresenta la *lunghezza del canale trasmissivo* espressa in bit, e cioè il numero massimo di bit che possono essere in transito tra le estremità del mezzo di comunicazione ad un qualsiasi istante di osservazione. In base alle definizioni risulta allora:

$$\alpha = \frac{RD}{L} = \frac{D}{L/R}.$$

Pertanto il parametro α è anche uguale al rapporto tra

- il massimo ritardo di propagazione D tra due stazioni che accedono al mezzo;
- il tempo L/R richiesto per la trasmissione di una UI tipica.

Valori del parametro α in una LAN sono nell'intervallo 0,01 - 0,1. Esistono tuttavia casi (ad es. nell'accesso multiplo in collegamenti via satellite) in cui il ritardo di propagazione prevale sul tempo di trasmissione, con la conseguenza che il parametro α assume valori maggiori dell'unità.

Supponiamo, con riferimento ad un *accesso perfetto*, che:

- il meccanismo di accesso consenta un solo trasferimento alla volta;
- il traffico offerto al sistema comporti trasferimenti senza soluzione di continuità;
- i trasferimenti non richiedano extra-informazione;
- il ritardo di propagazione non varii modificando le stazioni tra cui avviene il trasferimento.

In base a tali ipotesi, il valore massimo $maxS$ della portata media è esprimibile come rapporto tra i valori netti e lordi dei tempi necessari per trasferire una UI: mentre il valore netto è il tempo di trasmissione L/R, il valore lordo differisce da quello netto per il ritardo di propagazione D. Si ha quindi

$$max\ S = \frac{L/R}{D + L/R} = \frac{1}{1 + RD/L} = \frac{1}{1+\alpha}.$$

Conseguentemente l'utilizzazione U del canale è espressa da

$$U = \frac{1}{1+\alpha}.$$

Si vede allora che il parametro U è unitario solo nel caso ideale in cui α è uguale a zero (cioè quando il ritardo D è nullo) e assume un valore uguale a $1/(1+\alpha)$, indipendentemente dal meccanismo di accesso. Per ogni fissato valore di α, che dipende unicamente dalle caratteristiche fisiche del trasferimento (ritardo di propagazione, capacità di trasferimento del mezzo, lunghezza delle UI), risulta fissato il valore di U. Più in particolare:

- valori di U maggiori di 0,5 possono essere ottenuti scegliendo le quantità R e L in modo che il rapporto L/R sia prossimo a D ($\alpha \cong 1$); ne segue che, in presenza di elevati valori di ritardo di propagazione, il raggiungimento di questo obiettivo può ottenersi con lunghezze L decisamente maggiori della capacità R;
- valori di U maggiori di 0,9 richiedono rapporti L/R superiori a $10D$ ($\alpha \leq 0{,}1$); conseguentemente se il ritardo di propagazione è di modesta entità (come si verifica, ad es., nelle LAN) e se si desiderano elevate capacità di trasferimento (ad es. maggiori di 100 Mbit/s) è necessario impiegare UI di lunghezza particolarmente elevata (ad es., con i valori di D e di R sopra ipotizzati, L dovrebbe avere valori maggiori di 3.200 byte.

In Fig. V.1.6 è mostrato l'andamento della portata media in funzione del carico medio in un accesso multiplo perfetto: da questo andamento risulta che S cresce linearmente con G fino a raggiungere il valore *maxS ;* i valori in ascissa e in ordinata sono normalizzati rispetto alla capacità R e i massimi di portata media sono parametrati al variare di α.

Si nota infine che, con i valori tipici del parametro α in una LAN (normalmente minori di 0,1), in questo ambiente *maxS* non dovrebbe essere inferiore a circa 0,91. In effetti l'influenza del metodo di accesso è tale che, per vari protocolli MAC utilizzati nelle LAN, *maxS* assume valori decisamente inferiori.

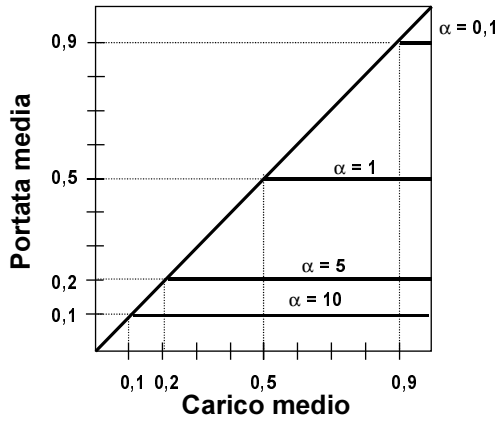


Figura V.1.6 –Andamento della portata media in funzione del carico medio in un accesso multiplo perfetto e per vari valori del parametro α

V.1.5 Protocolli ad accesso casuale

A causa della semplicità delle loro procedure, i protocolli ad accesso casuale sono facilmente realizzabili e relativamente poco costosi ed hanno avuto una larga applicazione sia in LAN di tipo commerciale, che in altri sistemi multi-accesso. In particolare, essi si adattano ad ambienti caratterizzati da un elevato numero di stazioni, ognuna delle quali offre un traffico di tipo interattivo, altamente intermittente.

Esistono numerosi schemi di protocolli ad accesso casuale, che si differenziano principalmente per le strategie seguite nel ridurre il numero di collisioni, o, comunque, per limitarne l'effetto sulle prestazioni della rete. A caratterizzare ciascuna di queste strategie e a condizionarne le prestazioni è l'*intervallo di vulnerabilità*, e cioè l'intervallo massimo di tempo entro cui una stazione può iniziare una emissione e collidere con un'altra emissione.

In questi schemi di protocolli si incontrano due alternative per la gestione dell'asse dei tempi: questo è infatti *indiviso* (unslotted) o *suddiviso in intervalli temporali* (slotted); nel secondo caso le stazioni debbono essere sincronizzate in modo che la emissione di una UI cominci con l'inizio di un intervallo temporale (IT).

Nel seguito di questa sezione saranno presentati i protocolli MAC della famiglia ALOHA e ne verranno valutate le prestazioni utilizzando un modello elementare. Altri protocolli ad accesso casuale saranno trattati nella sezione seguente.

V.1.5.1 Il protocollo "ALOHA puro"

Il protocollo ALOHA è stato sviluppato (agli inizi degli anni '70) per una rete radio multiaccesso presso l'Università delle Hawaii, ove più stazioni periferiche erano logicamente connesse da un unico canale ad una stazione centrale. Il suo principio-chiave si adatta però ad una qualsiasi rete con esigenze di accesso multiplo con allocazione dinamica. E' il più semplice protocollo ad accesso casuale. Verrà qui chiamato "*ALOHA puro*", per distinguerlo da una successiva versione, che verrà trattata nel seguito. In un sistema di accesso basato sullo schema ALOHA puro l'asse dei tempi è *indiviso*.

Il principio base del protocollo ALOHA puro è il seguente:

- una stazione può emettere una UI non appena questa è disponibile; viene quindi applicata la direttiva "*trasmetti subito* (senza preoccuparti delle iniziative di altre stazioni)";
- se la stazione emittente non riceve un riscontro positivo dalla stazione di destinazione entro un determinato intervallo di tempo (time-out), si assume che si sia verificata una *collisione*;
- occorre allora riemettere la UI corrispondente; per evitare nuove collisioni si rende *casuale_la riemissione*; cioè questa viene ritardata di un intervallo di tempo calcolato in base ad un *algoritmo di subentro* (back off).

Per valutare le prestazioni di questo e di altri protocolli ad accesso casuale, assumiamo un modello elementare, secondo il quale:

- tutte le UI hanno uguale lunghezza;
- alla emissione di una UI possono seguire solo due eventi:
  - *collisione* con distruzione completa dell'informazione portata da quella UI;
  - *ricezione esente da errori* (recapito con successo);
- la reazione sulle stazioni emittenti al verificarsi di questi due eventi è *istantanea*; cioè chi emette riceve informazione immediata sull'esito della sua emissione;
- la emissione di ogni UI che ha subito una collisione viene ripetuta finché la UI non è stata recapitata con successo; una stazione con una UI da riemettere si dice "*prenotata*" (backlogged).
- le stazioni sono in *numero infinito* e ogni nuovo arrivo di UI perviene ad una nuova stazione.

Indichiamo poi con R, D e L le quantità definite in § V.1.4. A complemento del modello per accesso multiplo casuale e con riferimento specifico a protocolli della famiglia ALOHA si osserva che il processo di arrivo complessivo delle UI alla rete è la sovrapposizione di due componenti: una è costituita dai *nuovi arrivi* e cioè da UI che vengono emesse per la prima volta; l'altra è formata dalle *UI che vengono ri-emesse*. Per entrambi questi processi si assume un *modello di arrivo poissoniano*;

conseguentemente poissoniano è anche il processo di arrivo complessivo. Più in particolare, nell'ambito di queste ipotesi, si indica con:

λ la frequenza media del processo dei nuovi arrivi;

G/T la frequenza media del processo degli arrivi complessivi,

ove G è il carico medio normalizzato e $T = L/R$ *è* il tempo di trasmissione di una UI.

Si nota che

- assumere poissoniano il processo dei nuovi arrivi rientra in una scelta modellistica che è normalmente lontana dalla realtà, ma che può essere accettabile ove la si consideri un riferimento per soli scopi di confronto;
- assumere poissoniano anche il processo degli arrivi legati alla ri-emissione è giustificabile solo se le ri-emissioni da parte delle stazioni "prenotate" sono sufficientemente casualizzate.

Come è evidente deve essere $G/T>\lambda$. Inoltre, sempre in base alle ipotesi fatte, l'*intervallo di vulnerabilità V* è in questo caso uguale a due volte la durata T di una UI

$$V = 2T;$$

infatti, come appare in Fig. V.1.7, se una UI viene emessa a partire dall'istante t_0 (l'emissione dura quindi fino a t_0+T), si ha collisione se almeno un'altra UI è già in corso di emissione nell'intervallo (t_0 - T, t_0) ovvero se almeno una ulteriore UI comincia ad essere emessa nell'intervallo (t_0, t_0+T).

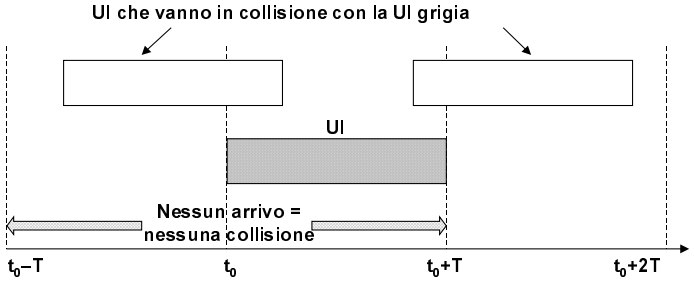


Figura V.1.7 – Per la determinazione dell'intervallo di vulnerabilità nel protocollo ALOHA puro

Con questo modello si può determinare la *portata media normalizzata S* del sistema di accesso, e cioè il numero di tentativi di emissione andati a buon fine in un intervallo di tempo uguale alla durata T di una UI. Questa portata è uguale a

$$S = p_0 G,$$

ove p_0 è la probabilità che una UI non subisca collisione nell'intervallo di vulnerabilità $V = 2T$. Tenendo conto che, in base al carattere poissoniano degli arrivi complessivi, la probabilità $P_k(X)$ di avere k arrivi complessivi in un intervallo di tempo di durata X è uguale a

$$P_k(X) = \frac{\left(G\frac{X}{T}\right)^k}{k!} e^{-G\frac{X}{T}}$$

e, poiché per definizione $p_0 = P_0(V)$, ne segue che

$$p_0 = e^{-2G},$$

da cui si ottiene

$$S = G\, e^{-2G}.$$

Con riferimento al protocollo ALOHA puro, la Fig. V.1.8 mostra l'andamento, secondo l'espressione ora ricavata, della portata media *S* in funzione del carico medio *G*. Entrambi i valori *S* e *G* sono normalizzati.

Come è facile verificare, il massimo valore della portata media si ottiene per *G=0,5* ed è uguale ad $S = 1/2e \cong 0{,}184$. Ciò stabilisce che il canale non può essere mediamente utilizzato per più del 18% della sua capacità di trasferimento.

Per $G > 0{,}5$, la portata media diminuisce, come attestazione di uno stato di instabilità. Ciò può essere spiegato con la circostanza che, all'aumentare del carico medio, la probabilità di collisione via via aumenta, determinando una diminuzione progressiva della quota parte di impegno utile del canale.

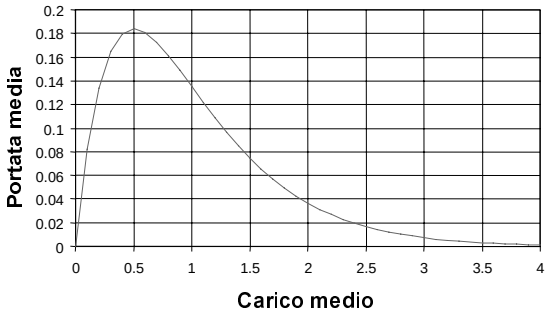


Figura V.1.8 – Portata media del protocollo ALOHA puro

V.1.5.2 Il protocollo "Slotted ALOHA"

Alla famiglia dei protocolli ALOHA appartiene anche lo "*slotted ALOHA*". Questo opera come il protocollo ALOHA puro, ma con un asse dei tempi *suddiviso in intervalli temporali* (IT) di durata fissa e quindi con una sincronizzazione tra le stazioni. La durata di un IT è uguale al tempo di trasmissione *T* di una UI. Ogni stazione è vincolata ad iniziare l'emissione delle proprie UI in corrispondenza dell'*inizio di un IT*. Con questo accorgimento l'intervallo di vulnerabilità *V* nella emissione si riduce alla durata *T* di un IT

$$V = T\ .$$

La portata media del protocollo slotted ALOHA si ottiene con procedimento del tutto analogo a quello seguito per il protocollo ALOHA puro, tenendo conto che in questo caso $V = T$; quindi

$$p_0 = e^{-G}.$$

La portata media normalizzata *S* è data allora da

$$S = G\, e^{-G}.$$

In Fig. V.1.9 è mostrato l'andamento corrispondente a questa espressione. L'andamento di *S* in funzione di *G* è anche confrontato con quello corrispondente nel caso del protocollo ALOHA puro. Come risulta, la diminuzione dell'intervallo di vulnerabilità comporta un sensibile miglioramento della portata media di rete, che in questo caso raggiunge il suo massimo per un carico medio $G = 1$ e con un valore corrispondente che è uguale al doppio di quello relativo al caso ALOHA puro: cioè $maxS = 1/e \cong 0{,}368$.

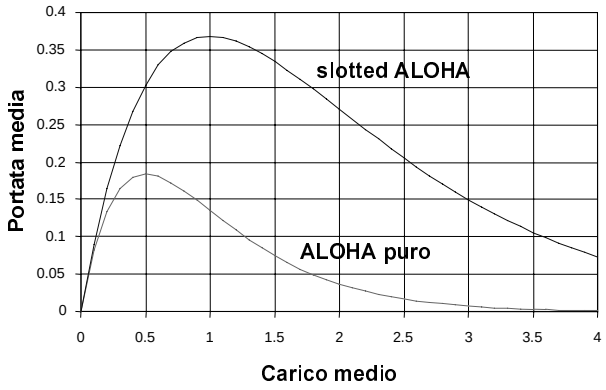


Figura V.1.9 – Portata media del protocollo slotted ALOHA

V.1.6 I protocolli CSMA

Questi protocolli appartengono ancora alla famiglia delle procedure ad accesso casuale. Il loro capostipite è il protocollo *CSMA* (Carrier Sense Multiple Access), che adotta la direttiva "*ascolta prima di parlare*".

Secondo questa procedura, una stazione che desidera emettere ascolta se il canale è occupato da una emissione precedente: se il canale è libero, la stazione emette; se il canale è occupato, la stazione ritarda l'emissione ad un istante successivo. Nonostante questa cautela nell'emissione di una UI, il protocollo CSMA non evita le collisioni: ciò a causa dei ritardi di propagazione che non consentono ad una generica stazione in ascolto di avere una informazione completa sullo stato di occupazione del canale. Ciò è mostrato in Fig. V.1.10, ove viene mostrato che tra due stazioni avviene una collisione se esse accedono al canale in istanti che distano tra loro meno del ritardo di propagazione tra le due stazioni.

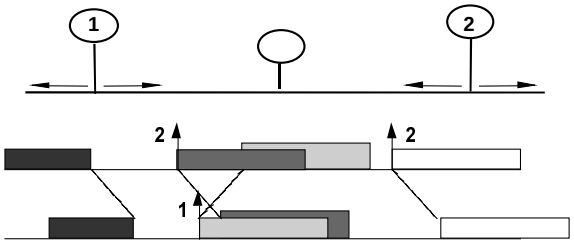


Figura V.1.10 – Eventi di collisione nel protocollo CSMA

V.1.6.1 Procedure di persistenza

Il protocollo CSMA deve gestire due problemi connessi alla reiterazione dei tentativi di accesso

- in presenza di canale occupato;
- a seguito di collisioni.

Nel caso di *canale occupato*, l'istante successivo di emissione è determinato in base ad una *procedura di persistenza*; questa può essere di tre tipi:

1-persistente;
0-persistente;
p-persistente ($0 < p < 1$);

il loro comportamento è mostrato in Fig. V.1.11.

La procedura *1-persistente* (Fig. V.1.11A) prevede che, se al momento dell'ascolto, il canale è occupato, la stazione continui l'ascolto del canale ed esegua la sua emissione non appena il canale diventi libero.

La procedura *0-persistente* (Fig. V.1.11B) differisce completamente dalla precedente. Infatti, se il canale è occupato, la stazione ritarda la emissione di un intervallo determinato attraverso l'esecuzione di un *algoritmo di subentro* (backoff). Questo ha lo scopo di casualizzare il nuovo accesso al canale e quindi di ridurre la probabilità di collisione.

Un comportamento intermedio è fornito dalla procedura *p-persistente* (Fig. V.1.11C). In questo caso, se il canale è occupato, la stazione attende che ritorni nello stato di libero. A questo punto la emissione avviene con probabilità *p*, mentre con probabilità *1-p* la stazione attende ulteriormente un intervallo la cui durata si calcola con l'algoritmo di subentro.

L'importanza dell'algoritmo di subentro è dimostrata dal seguente esempio. Durante la emissione di una UI, più di una stazione può effettuare l'ascolto del canale ottenendo lo stesso risultato di canale occupato. Se il nuovo tentativo di emissione avvenisse dopo un intervallo di tempo fisso, tutti i terminali eseguirebbero le operazioni di ascolto del canale alle stesse distanze temporali reciproche in cui hanno effettuato la medesima operazione la prima volta. Ciò causerebbe un ulteriore ritardo nella emissione per effetto dell'occupazione del canale.

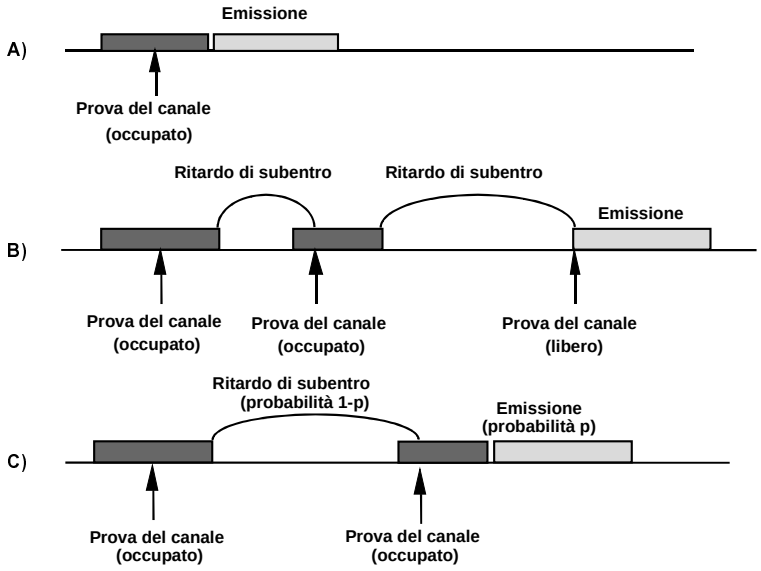


Figura V.1.11 – Effetto delle procedure di persistenza:
A) 1- persistente; B) 0-persistente, C) p-persistente

Confrontando le varie procedure di persistenza, si può osservare che:

- la procedura *1*-persistente potrebbe portare a buoni valori di portata media in quanto il canale non può andare a riposo finché esiste almeno una stazione che deve emettere, ma ha lo svantaggio di causare una collisione, con probabilità 1, se più stazioni ascoltano il canale durante una emissione;
- la procedura *0*-persistente è in grado di ridurre questo svantaggio rendendo casuale l'istante di inizio di un tentativo di emissione;
- la procedura *p*-persistente consente, valutando opportunamente il valore del parametro *p*, di ottimizzare le prestazioni del protocollo in funzione dei particolari parametri di rete e di traffico incontrati.

V.1.6.2 Intervallo di vulnerabilità

Come appare chiaro dalla descrizione del protocollo CSMA, può accadere che, anche se il canale è rivelato libero, due o più stazioni emettano contemporaneamente: ciò provoca una collisione e causa quindi la perdita delle informazioni. Infatti, detto D il ritardo di propagazione da estremo a estremo sul canale, una stazione A, che abbia inserito una unità informativa in rete può subire interferenze da parte delle stazioni che hanno ascoltato il canale trovandolo libero e che hanno quindi cominciato a emettere nell'intervallo compreso fra $-D$ e $+D$ rispetto all'inizio della emissione dell'unità informativa da parte della stazione A. Pertanto, nel protocollo CSMA, l'*intervallo di vulnerabilità* è dato da

$$V = 2D\ .$$

È evidente che V è un parametro che influenza pesantemente le prestazioni del protocollo, giacché il numero di collisioni, e quindi l'intervallo di tempo necessario a emettere con successo una unità informativa, risulta crescente con l'aumentare di V e del carico medio. Tanto più la rete è estesa, tanto più elevati sono i valori di V e della probabilità di collisione. Per questa ragione il protocollo CSMA non può essere utilizzato in reti molto estese, pena una sensibile riduzione del valor massimo della loro portata media.

V.1.6.3 Il protocollo CSMA/CD

Per superare queste limitazioni è stato definito il protocollo *CSMA/CD* (*Collision Detection*), che si distingue dal CSMA per il fatto che le stazioni, oltre alla strategia "*ascolta prima di parlare*", utilizzano anche quella "*ascolta mentre parli*" per tutta la durata della emissione. In questo modo le stazioni sono in grado di rivelare le collisioni confrontando la sequenza di bit emessa con quella ricevuta. Tale accorgimento ha il vantaggio di consentire l'interruzione della emissione di una unità informativa nell'istante in cui la collisione viene rivelata e quindi di ridurre il tempo di occupazione non utile del canale. Quando viene rivelata una collisione, la emissione viene, come detto, interrotta. L'istante in cui sarà effettuato il tentativo successivo è determinato utilizzando l'algoritmo di subentro.

Circa la modalità di rivelazione di una collisione, questa funzione è svolta sulla base di un confronto tra ciò che si sta trasmettendo e ciò che è ascoltabile sul canale. Se ciò che viene letto è differente da ciò che viene emesso, la stazione deduce che è avvenuta una collisione. E' quindi importante che la codifica del segnale sia tale da rendere possibile questo confronto e che il risultato della avvenuta collisione pervenga alla stazione emittente mentre l'emissione è ancora in corso. Per rispondere poi al secondo requisito, occorre impiegare UI di lunghezza maggiore di un limite definito. Per determinare questo limite, indichiamo con:

- $T1$ il tempo necessario alla rivelazione di una collisione;
- $T2$ è il tempo di permanenza del canale nello stato di avvenuta collisione;
- Tc il tempo totale necessario affinché tutti i terminali interrompano la emissione.

Si riconosce facilmente che

$$T_c = 2D + T_1 + T_2\ ,$$

come viene chiarito in Fig. V.1.12: infatti se le stazioni A e B sono poste agli estremi del canale trasmissivo ed iniziano a emettere rispettivamente negli istanti t_{0A} e $t_{0A}+D$, la stazione A è in grado di rivelare la collisione nell'istante $t_{0A}+2D+T_1$ per poi mantenere lo stato di collisione del canale fino all'istante $t_{0A}+2D+T_1+T_2$, istante entro il quale tutte le emissioni sono interrotte.

Conseguentemente, per mettere in condizioni tutte le stazioni, in particolare quelle a distanza massima, di rivelare una collisione, la durata T di una *PDU di strato MAC CSMA/CD,* deve essere non inferiore al minimo di T_c e cioè a $2D+T_1$. Ne segue che deve risultare

$$T = \frac{L}{R} \geq 2D + T_1 \ ;$$

quindi una PDU di strato MAC non può avere lunghezza inferiore a

$$min\, L = (2D + T_1)R \ .$$

Nel caso di una LAN in cui D è uguale a 25,6 μs, la lunghezza minima di una PDU di strato MAC è uguale a 64 bytes quando R = 10 Mbit/s. Se invece fosse R = 1 Gbit/s, risulterebbe min L = 6.400 bytes.

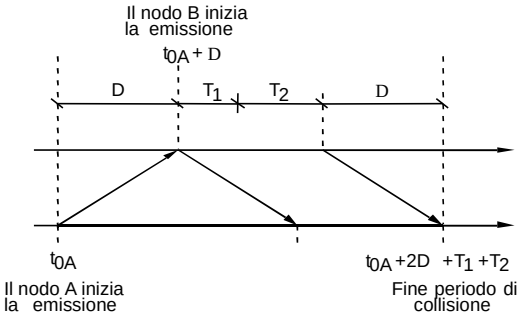


Figura V.1.12 – Tempo di rivelazione di una collisione nel protocollo CSMA/CD

V.1.6.4 Efficienza del canale nell'accesso CSMA/CD

Il problema più importante che occorre considerare nell'uso del protocollo CSMA/CD, ed in generale in tutti i protocolli ad accesso casuale, riguarda la degradazione delle prestazioni ad alto traffico. In particolare, questa si verifica poiché, quando il carico medio sulla rete cresce, il numero delle collisioni tende ad aumentare e quindi aumenta il tempo in cui il canale non è disponibile per le emissioni. Se il traffico supera un determinato valore di soglia, la durata di occupazione del canale per le collisioni e quindi per emissioni senza successo diviene via via preponderante rispetto a quella spesa in emissioni utili, fino ad impedire le emissioni stesse. Per questa ragione l'utilizzazione del canale non può andare oltre a valori nell'intervallo 0,6÷0,7.

E' interessante, se non altro a scopo di confronto con lo schema di accesso perfetto e con i protocolli della famiglia ALOHA, valutare la portata media di un sistema che adotta il protocollo CSMA/CD. Ci riferiamo, per semplicità, ad una *procedura 1-persistente* e assumiamo il modello elementare di accesso multiplo casuale già utilizzato per i protocolli ALOHA.

Supponiamo che all'istante t_0 una delle stazioni accedenti abbia completato l'emissione di una UI; allora una qualunque delle stazioni aventi una UI da emettere può tentare di farlo. Se due o più stazioni decidono di emettere simultaneamente ci sarà una collisione; ognuna di queste stazioni rivelerà la collisione e di conseguenza cesserà la sua emissione, aspettando che il canale trasmissivo si liberi prima di procedere ad un successivo tentativo. Pertanto nel funzionamento del sistema di accesso si distinguono tre intervalli temporali:

- uno *di emissione*, nel corso del quale il canale trasmissivo è impegnato nel trasferimento di una UI;
- uno *di contesa*, comprendente i tentativi di emissione da parte di stazioni che hanno UI da emettere;
- uno *di riposo*, che si presenta quando tutte le stazioni sono in uno stato inattivo.

La prestazione del protocollo dipende dalla durata dell'intervallo di contesa (oltre che da quella dell'intervallo di emissione). Per analizzare la durata del primo di questi intervalli entriamo nel dettaglio dell'algoritmo di accesso.

Supponiamo che due stazioni comincino entrambe ad emettere esattamente al tempo t_0; si renderanno conto che si è verificata una collisione dopo un intervallo di tempo che è al più uguale a $2D$. In altre parole, nel caso peggiore, una stazione non può essere sicura di avere occupato il canale con successo finché essa non ha trasmesso per una durata uguale a $2D$ senza aver percepito una collisione. Per questa ragione, l'intervallo di contesa può essere rappresentato come suddiviso in intervalli temporali (*IT di contesa*) di durata $2D$. Per brevità, nel seguito, gli IT di contesa saranno indicati con ITC.

Per valutare la durata dell'intervallo di contesa, ipotizziamo che la struttura di accesso operi in condizioni di carico elevato e costante e che di conseguenza esistano sempre N stazioni ($N>0$) pronte ad emettere. Indichiamo poi con

π la probabilità che una stazione (una tra le N) emetta durante un ITC;

P la *probabilità di successo*, e cioè la probabilità che una stazione (tra le N) acquisisca l'accesso durante quell'ITC.

Con queste notazioni la probabilità di successo è esprimibile con

$$P = N\pi\,(1-\pi)^{N-1}.$$

La probabilità P, come funzione di π, presenta un massimo, max$P$, per $\pi=1/N;$ questo massimo è uguale a

$$max\ P = \left(1 - \frac{1}{N}\right)^{N-1}$$

ed è una funzione di N monotona decrescente, che tende ad $1/e$ quando N tende all'infinito. E' tuttavia da osservare che, già per $N \geq 5$, max$P \leq 0{,}41$ (valore molto prossimo a $1/e \cong 0.37$).

D'altra parte, poiché la probabilità che l'intervallo di contesa contenga esattamente n ITC è uguale a $P(1-P)^{n-1}$, il numero medio di ITC in un intervallo di contesa è dato da

$$\sum_{n=0}^{\infty} nP(1-P)^{n-1} = \frac{1}{P}\ .$$

Dato che, per ipotesi, ogni ITC ha una durata $2D$, l'intervallo di contesa ha una durata media C che è uguale a $2D/P$. Assumendo il valore di π che massimizza P, il numero medio di ITC per intervallo di contesa non è mai maggiore di e C è quindi al più uguale a $2De$.

In conclusione, per trasferire una UI, il meccanismo del protocollo CSMA/CD con procedura 1-persistente richiede che al tempo di trasmissione L/R si aggiunga la durata media C di un intervallo di contesa; conseguentemente la portata media S è esprimibile con

$$S = \frac{L/R}{L/R + 2D/P} = \frac{1}{1 + 2\alpha/P}\ ,$$

ove $\alpha = DR/L$ è il parametro definito in § V.1.6. Inoltre l'efficienza U del canale (tenendo conto che $C \leq 2De$) è data da

$$U = \frac{1}{1 + 2e\alpha}\ .$$

V.1.7 Protocolli ad accesso controllato

Vengono qui considerati i protocolli a testimone su bus e su anello.

V.1.7.1 Protocollo “a testimone su bus”

Il protocollo *a testimone su bus* (token bus) è ad accesso controllato in quanto consente ai terminali connessi ad una rete a bus di accedere al mezzo solo quando sono in possesso di una particolare PDU di controllo chiamata *testimone*. È descritto nello standard IEEE 802.4.

Il testimone è trasferito da una stazione all'altro lungo un *anello logico* sul bus. La definizione della sequenza dei terminali sull'anello logico è completamente indipendente dalla loro posizione fisica sul bus. Una stazione viene in possesso del testimone quando rivela, sul bus, il transito del testimone recante il proprio indirizzo.

Durante il funzionamento normale, la rete passa ciclicamente attraverso due fasi: 1) la *fase di emissione*, in cui la stazione che possiede il testimone emette la propria unità informativa; 2) la *fase di trasferimento del testimone* in cui la stazione che ha concluso la sua emissione trasferisce il testimone alla stazione che lo segue sull'anello logico, dopo aver posto l'indirizzo di questo all'interno del testimone stesso.

Per consentire il trasferimento corretto del testimone tra le varie stazioni, ognuna di queste deve conoscere il proprio indirizzo, quello della stazione che lo precede e quello della stazione che la segue sull'anello logico.

In Fig. V.1.13 è rappresentato un possibile anello logico ricavato tra i terminali di un bus bidirezionale.

È evidente che la sequenza delle stazioni sull'anello logico può essere qualsiasi e non dipende in alcun modo dalla disposizione fisica di queste sul bus. La circolazione del testimone avviene nel seguente modo. Una stazione, ad esempio la 5, rivela sul bus la presenza del testimone a lei indirizzato; se ha informazione da emettere, esegue la emissione. Al termine di questa rilascia il testimone con l'indirizzo della stazione 3, che è quella successiva nell'anello logico.

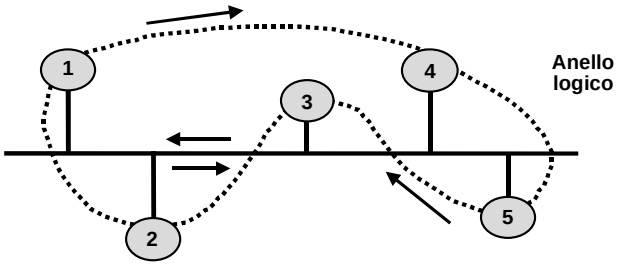


Fig. V.1.13 - Creazione di un anello logico su un bus bidirezionale.

Una caratteristica fondamentale del protocollo a testimone su bus è che il *tempo di ciclo*, e cioè il tempo che intercorre tra due successive visite del testimone ad una stessa stazione e quindi tra due successive opportunità di emissione, ha un valore massimo finito ed uguale alla somma dei tempi di trasferimento del testimone e di emissione delle MAC PDU in tutti i terminali. Tale caratteristica consente quindi di garantire ad una stazione, in qualsiasi condizione, un valore minimo di traffico smaltibile dalla rete.

Da tale caratteristica emerge che il principale vantaggio offerto dal protocollo a testimone su bus, rispetto a quello CSMA/CD, consiste nell'assenza di fenomeni degenerativi delle prestazioni. I protocolli a testimone riescono infatti a garantire un valore stabile di traffico smaltito anche se il traffico offerto supera la portata massima della rete.

I vantaggi precedentemente elencati sono ottenuti al prezzo di una considerevole complicazione della struttura della rete e delle funzionalità di ogni singola stazione. Infatti, ognuna di queste, oltre a quelle di accesso, deve eseguire le seguenti funzioni di gestione:

- inizializzazione dell'anello logico e creazione del testimone;
- gestione del testimone, cioè sua ri-emissione in caso di distruzione, riconoscimento e cancellazione di testimoni multipli, ecc.;
- inserimento delle stazioni nell'anello logico e loro rimozione;
- recupero delle funzionalità di rete in caso di guasti;
- riconoscimento ed aggiornamento degli indirizzi.

La gestione del testimone è l'elemento più delicato del protocollo. Infatti per cause accidentali può accadere che il testimone sia perso, causando l'interruzione del servizio di rete. Le ragioni che possono portare alla perdita del testimone sono da individuarsi nel guasto di una stazione, che non rilancia il testimone, o negli errori di linea, che rendono inintelligibile il contenuto del testimone stesso. Per fronteggiare situazioni di questo tipo occorre prevedere opportune procedure di controllo della rete.

V.1.7.2 Il protocollo “a testimone su anello”

Il protocollo *a testimone su anello* (token ring) è basato sullo stesso principio di quello a testimone su bus, ma è definito per essere utilizzato su reti con topologia ad anello. È descritto nello standard IEEE 802.5.

Il testimone circola sull'anello dando ciclicamente ad ogni stazione l'opportunità di accedere alla rete (Fig. V.1.14A). Una stazione che è in attesa di emissione può impadronirsi del testimone e iniziare la emissione quando il testimone transita attraverso la propria interfaccia (Fig. V.1.14B). La stazione che inserisce una unità informativa in rete deve successivamente rimuoverla ed emettere un nuovo testimone (Fig. V.1.14C).

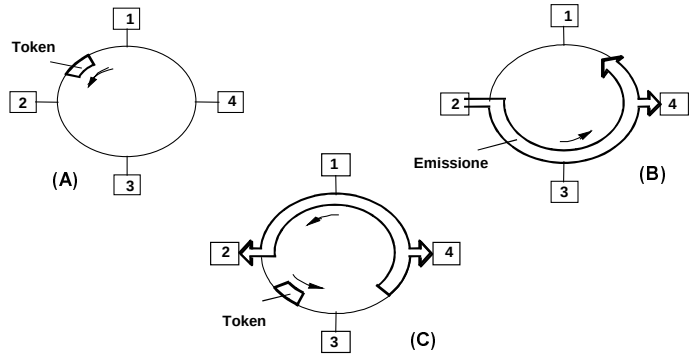


Fig. V.1.14- Fasi del protocollo a testimone su anello: (A) circolazione del token; (B) cattura del token da parte di una stazione; (C) riemessione del token.

V.2 Controllo di errore

Per rivelare errori in ricezione su una stringa di bit (*stringa da proteggere*), è necessario in emissione aggiungere a questa uno o più bit addizionali (*extra-bit*), che sono la *ridondanza* del *codice a rivelazione di errore*. Se

K è la lunghezza della stringa da proteggere;

Z è la lunghezza della stringa di extra-bit,

le *parole di codice* sono di lunghezza uguale a *K+Z.*

V.2.1 Codici con controllo di parità

I metodi di codifica per rivelare errori nel trasporto dei dati rientrano usualmente nella categoria dei *codici con controllo di parità* (parity check codes). A questa categoria appartengono

- il controllo di *parità singola*;
- il controllo di *parità a blocchi*;
- il controllo a *ridondanza ciclica* (CRC, Cyclic Redundancy Check).

L'efficacia di un codice a rivelazione di errore è misurata da tre parametri:

- la *distanza minima, min D*, del codice;
- la potenzialità *B* a *rivelare raffiche* (burst) *di errori*;
- la *probabilità* P_{enr} che una stringa di bit completamente casuale sia *accettata in ricezione come priva di errori*.

La *distanza minima* del codice è definita come il *più piccolo* numero di errori che possono trasformare una parola di codice in un'altra. La *lunghezza* di una raffica di errori in una trama è il numero di bit dal primo errore all'ultimo, inclusi. La potenzialità a rivelare raffiche di errore in un codice è definita come *il più grande intero B* tale che il codice possa rivelare tutte le raffiche di *lunghezza non superiore a B*.

Circa la probabilità P_{enr}, con riferimento a una stringa da proteggere di lunghezza K e a una stringa di Z extra-bit, osserviamo che:

- una stringa di lunghezza $K+ Z$ è *completamente casuale* (cifre binarie indipendenti) quando viene ricevuta con probabilità $2^{-(K+Z)}$;
- esistono 2^K parole di codice.

Inoltre la probabilità di *errori non rivelati* in una stringa completamente casuale è uguale alla probabilità che questa stringa sia *una delle parole di codice*. Tale seconda probabilità (ignorando la possibilità che la stringa casuale ricevuta sia la stessa della trama emessa) e' uguale a $2^{-(K+Z)}\ 2^K = 2^{-Z}$. Questa è usualmente una buona stima della *probabilità* P_{enr} .

Nel *controllo di parità singola* si aggiunge un singolo bit (*bit di parità*) alla stringa da proteggere (ad es. ad un carattere). Il bit di parità ha un *valore "1"* se il numero di "1" nella stringa è *dispari*; altrimenti è posto al valore "0". Operativamente, il bit di parità è la *somma modulo 2* dei valori di cifra binaria contenuti nella stringa da proteggere.

Il metodo con controllo di parità singola consente la rivelazione di un *numero dispari di errori*, mentre fallisce quando il numero di errori è pari. In questa tecnica di codifica:

- la *distanza minima min D del codice* è uguale a 2;
- la *potenzialità B a rivelare raffiche di errore* è uguale a 1.

Nel *controllo di parità a blocchi* si organizza la stringa da proteggere in una forma matriciale a due dimensioni e si applica il controllo di parità singola ad ogni riga e ad ogni colonna (Fig V.2.1). Il controllo di parità relativo all'angolo destro in basso può essere visto come un controllo di parità su riga, o come un controllo di parità su colonna ovvero come un controllo di parità sull'intera stringa da proteggere.

Il controllo di parità a blocchi è efficace in presenza:

- di un *numero dispari di errori* che colpiscono una singola riga ovvero una singola colonna: la rivelazione è effettuabile con la parità di riga e di colonna, rispettivamente;
- di errori che colpiscono *una singola riga* o *una singola colonna*: la rivelazione di ogni singolo errore è effettuabile con la parità di colonna e di riga, rispettivamente.

Fallisce però quando si verificano, ad esempio, quattro errori confinati a due righe e a due colonne secondo una configurazione a rettangolo (cfr. Fig. V.2.1). La *distanza minima min D* del codice è uguale a 4; la *potenzialità B a rivelare raffiche di errore* è uguale a 1 più la lunghezza di una riga (assumendo che le righe siano emesse una dopo l'altra).

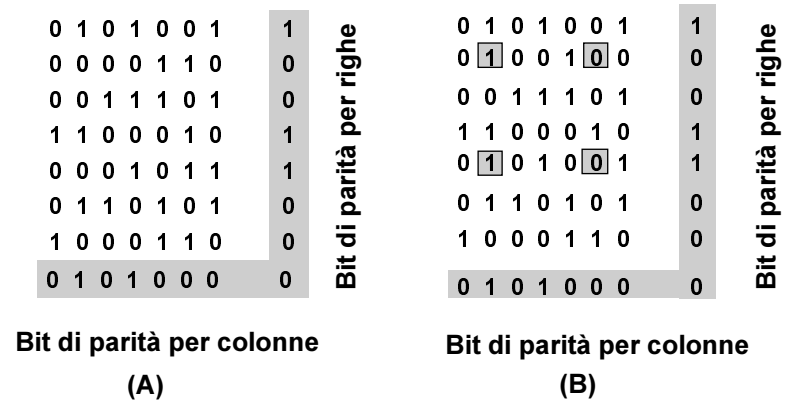


Figura V.2.1 – Controllo di parità a blocchi: (A) Blocco di carattere originario e bit di parità; (B) errori non rivelati.

V.2.2 Codici CRC

Le singole cifre binarie di una stringa da proteggere sono trattate come coefficienti (di valore "0" o "1") di un polinomio *P(x)*. Le cifre binarie della stringa di lunghezza uguale a *K* sono quindi considerate come i coefficienti di un polinomio completo di grado *K-1*. In particolare, l'*i*-esimo bit della stringa è il coefficiente del termine x^{i-1} di *P(x)* .

Le entità emittente e ricevente utilizzano un polinomio comune *G(x)*, detto *polinomio generatore*, che qualifica il codice a rivelazione di errore. Il polinomio *G(x)*

- gode di *opportune proprietà* nell'ambito della *teoria dei campi algebrici;*
- i suoi coefficienti sono *binari*, come quelli di *P(x)*;
- il suo grado è uguale a *Z*;
- tra i suoi coefficienti, quelli dei termini di grado massimo e di grado nullo debbono entrambi essere uguali a 1.

La *entità emittente* utilizza *G(x)* come divisore del polinomio $x^Z P(x)$

$$\frac{x^Z P(x)}{G(x)} = N(x) + \frac{R(x)}{G(x)}$$

dove si indica con:

N(x) il polinomio *quoziente*;
R(x) il polinomio *resto*.

La divisione è l'ordinaria divisione di un polinomio per un altro; la particolarità risiede in:

- i coefficienti di dividendo e di divisore sono binari;
- l'aritmetica viene svolta modulo 2.

Dato il grado del polinomio generatore, il grado del polinomio resto *R(x)* è al più uguale a *Z-1;* conseguentemente *R(x)* può essere sempre rappresentato con Z coefficienti (binari), ponendo uguali a "0" i coefficienti dei termini mancanti. Ottenuto il resto *R(x),* l'entità emittente inserisce i coefficienti di questo polinomio in un apposito campo della stringa da emettere (*campo CRC*), che deve quindi avere lunghezza *Z*.

Nella stringa emessa trovano quindi posto le cifre binarie da proteggere (in numero uguale a *K*) e le cifre CRC (in numero uguale a *Z*): in totale *K+Z* cifre binarie, che sono rappresentative di un polinomio *T(x)* di grado *K+Z-1*

$$T(x) = x^Z P(x) + R(x)$$

e che costituiscono una *parola di codice.* Tenendo conto che, per definizione,

$$x^Z P(x) = N(x)G(x) + R(x)$$

e poiché addizione e sottrazione modulo 2 si equivalgono, si ottiene

$$T(x) = N(x)G(x) ;$$

cioè la stringa emessa (rappresentativa del polinomio *T(x)*) è divisibile per il polinomio generatore *G(x).* Si conclude che tutte le parole di codice sono divisibili per il polinomio generatore e tutti i polinomi divisibili per *G(x)* sono parole di codice.

I codici CRC rientrano nella famiglia dei *codici a blocchi lineari* , cioè gli extra bit dipendono solo dai bit da proteggere e la dipendenza è lineare; godono delle seguenti specificità:

- sono codici *ciclici*, e cioè una permutazione ciclica di una parola di codice è ancora una parola di codice;
- sono codici *polinomiali*, e cioè sono codici ciclici in cui ciascuna delle parole di codice, considerata nella sua forma polinomiale, è un multiplo del polinomio generatore;
- in particolare, se la parola di codice ha lunghezza *K+Z*, il polinomio generatore (di grado *Z*) è un divisore di $x^{K+Z} + 1$.

Quest'ultima proprietà è un vincolo importante dato che x^{K+Z} non ammette molti divisori per pressoché tutti gli interi *K+Z.*

La *entità ricevente* esegue, con il polinomio generatore, l'operazione di divisione effettuata in emissione. In questo caso opera però sul polinomio *V(x),* rappresentato dalle *K+Z* cifre binarie ricevute. Supponiamo che nel trasferimento si siano verificati errori, con una sequenza rappresentata dal *polinomio E(x)*: ogni errore nella stringa corrisponde ad un coefficiente non nullo in *E(x)*. Allora

$$V(x) = T(x) + E(x) ,$$

ove l'addizione è svolta modulo 2. Ogni bit "1" in *E (x)* corrisponde ad un bit che è stato invertito e quindi a un errore isolato. Se ci sono *k* bit "1" in *E (x)*, sono avvenuti *k* errori di un singolo bit. Un singolo *errore a raffica* di lunghezza *k* è caratterizzato in *E (x)* da un "1" iniziale, una mescolanza di "0" e "1", e un "1" finale, mentre tutti gli altri bit sono "0".

$$E(x) = x^i (x^{k-1} + ... + 1),$$

ove *i* determina quanto la raffica è lontana dall'estremità destra della stringa emessa.

Il ricevitore calcola il resto della divisione di *V(x)* per *G(x);* le modalità sono le stesse utilizzate nell'emettitore. Poiché *T(x)* è divisibile per *G(x),* ne segue che

$$Resto\left[\frac{V(x)}{G(x)}\right] = Resto\left[\frac{T(x)+E(x)}{G(x)}\right] = Resto\left[\frac{E(x)}{G(x)}\right].$$

Conseguentemente la regola applicata dal ricevitore è la seguente:

- se il resto della divisione *V(x)/G(x)* è nullo, la stringa ricevuta è assunta "senza errori";
- in caso contrario, si sono verificati uno o più errori nel corso del trasferimento.

Si nota che sono *non rivelabili* le configurazioni di errore per le quali il relativo polinomio *E (x)* contiene *G(x)* come fattore.

Un codice polinomiale, in cui il polinomio generatore contiene *x+1* come fattore primo, è in grado di rivelare

- tutti gli errori *singoli* o *doppi*;
- tutti gli errori isolati con una *molteplicità dispari*;
- tutti gli errori a raffica di lunghezza $\leq Z$.

Se la lunghezza della raffica è *Z+1* e se tutte le combinazioni della raffica sono considerate equiprobabili, la probabilità che l'errore a raffica non sia rivelato è uguale a $2^{-(Z-1)}$. Infine, se la raffica singola ha lunghezza maggiore di *Z+1* o se si verificassero varie raffiche più corte, nell'ipotesi di equiprobabilità delle configurazioni di errore, la probabilità di errore non rivelato è uguale a 2^{-Z}.

La distanza minima *min D* di un codice CRC, con polinomio generatore divisibile per *x+1*, è uguale a 4. La potenzialità *B* a rivelare raffiche di errore è non inferiore a *Z*. La probabilità P_{enr} di errori non rivelati in una stringa di cifre completamente aleatoria è uguale a 2^{-Z}.

Circa i polinomi generatori utilizzati, sono standard:

$$G(x) = x^{16} + x^{12} + x^5 + 1;$$

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

Entrambi sono divisibili per *x+1* e quindi danno luogo a codici CRC con le proprietà suddette.

V.2.3 Mezzi per il recupero di errore

Le procedure di recupero d'errore hanno lo scopo di assicurare che il flusso di PDU (in particolare di quelle contenenti dati di utente), trasferite tra le entità di strato in cui viene applicata la procedura, pervenga a destinazione:

- *senza errori,* almeno nei limiti consentiti dalla capacità di protezione del codice utilizzato;
- *senza perdite*, *senza duplicazioni* e *senza fuori-sequenza*, almeno per ciò che riguarda la consegna delle relative SDU allo strato superiore.

Dette procedure, che sono usualmente indicate con la sigla ARQ (*Automatic Repeat Request*), hanno l'ulteriore scopo di *controllare il flusso di informazioni*, in modo che sia consegnata a destinazione solo la *portata* (volume di informazione nell'unità di tempo) *smaltibile da chi riceve*. Tra le procedure di recupero, se ne distinguono tre tipi principali:

- *riemissione con arresto e attesa;*
- *riemissione continua;*
- *riemissione selettiva,*

facenti riferimento alla modalità di recupero mediante riemissione delle PDU che non rispettano gli scopi della procedura. Per ognuno di questi tre tipi sono state proposte più varianti; nel seguito se ne considera solo una per ogni tipo di procedura.

Nelle procedure di recupero si utilizzano, in alternativa o in unione, uno o più dei mezzi seguenti:

- *riscontri positivi* (ACK) o *negativi* (NAK) (Fig. V.2.2);
- *temporizzatori* (timer) (Figg. V.2.3 e V.2.4);
- *numeri di sequenza* nelle trame e relative *finestre scorrevoli* (Fig. V.2.5).

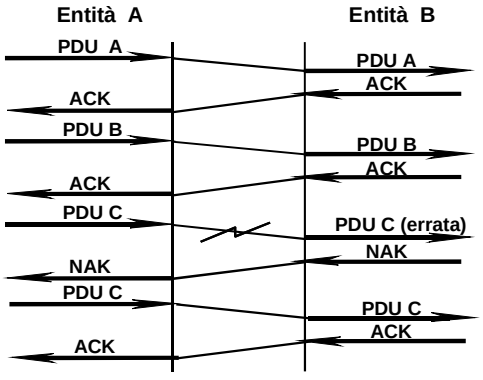


Figura V.2.2 – Procedura di recupero con uso dei riscontri positivi e negativi

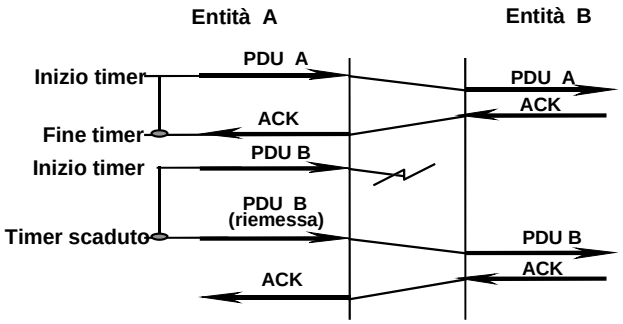


Figura V.2.3 – Procedura di recupero con uso dei temporizzatori: recupero di errore su una PDU

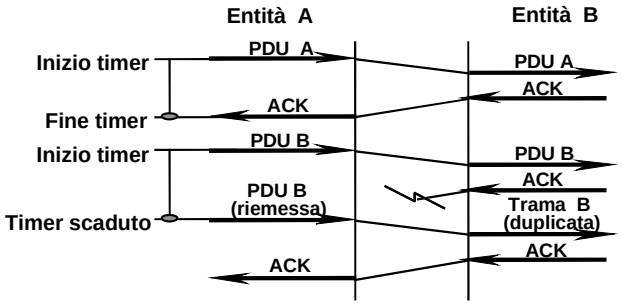


Figura V.2.4 – Procedura di recupero con uso dei temporizzatori: duplicazione di una PDU in seguito a perdita di un riscontro

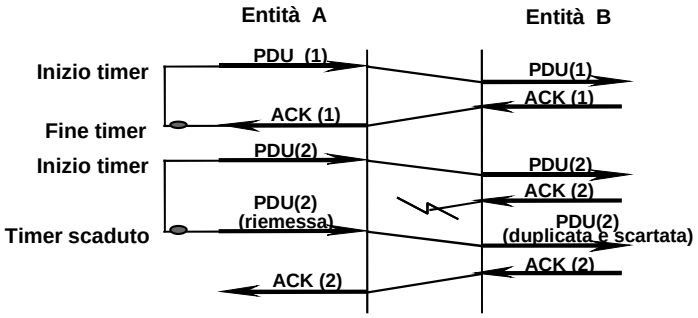


Figura V.2.5 - Procedura di recupero con uso dei temporizzatori e della numerazione delle PDU

I *riscontri positivi* emessi da chi riceve per informare chi emette sulla ricezione *accettata* di una o più PDU sono trasferibili, in alternativa o in unione:

- con *PDU* apposite (*PDU-ACK*), contenenti solo informazione di controllo, ma anch'esse protette con un CRC;
- con *PDU* contenenti dati di utente (*PDU-dati*) e trasferite nel verso contrario a quello dell'informazione da riscontrare: il riscontro è contenuto nell'intestazione della PDU come parte dell'informazione di controllo;

Questa seconda modalità di riscontro è nota come tecnica dell'"*addossamento*" (piggybacking) e consiste nell'inserire le informazioni di riscontro nelle PDU-dati che sono trasferite da chi ha ricevuto verso chi ha emesso.

Ogni PDU-dati può contenere due numeri di sequenza:

- il *numero di sequenza in emissione NS*, che è il numero d'ordine sequenziale caratterizzante la PDU uscente dall'entità agente come emittente;
- il *numero di sequenza in ricezione NR*, che è l'*NS* della PDU che l'entità agente come ricevente (e quindi come riscontrante) si aspetta di ricevere.

Il numero *NR* è *riscontro positivo implicito* per la PDU con *NS=NR-1* (*riscontro individuale*) ovvero per tutte le PDU non ancora riscontrate e con *NS* minore o uguale a *NR-1* (*riscontro cumulativo*).

Poiché entrambi i numeri di sequenza sono rappresentabili con stringhe binarie di lunghezza limitata *M*, ne deriva che per la loro rappresentazione occorre ricorrere a una *numerazione modulo M* (numerazione *ciclica*) come in Fig. V.2.6.

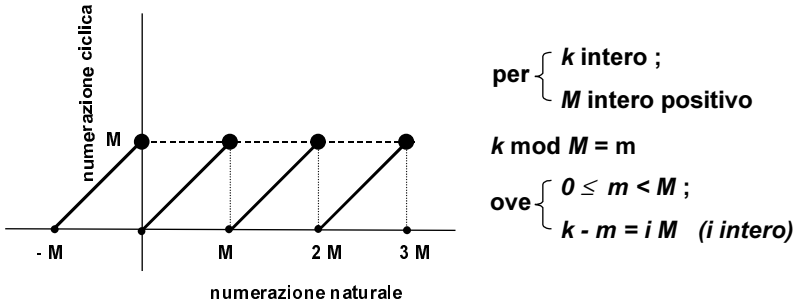


Figura V.2.6 – Rappresentazione di una numerazione modulo M

Questa numerazione è quindi rappresentabile su una circonferenza, suddivisa in *M* segmenti di uguale lunghezza (Fig. V.2.7).

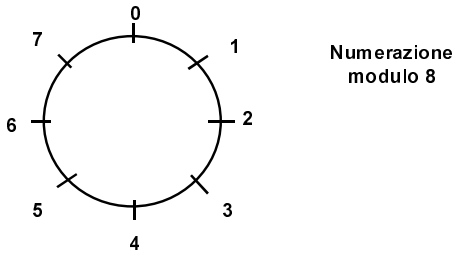


Figura V.2.7 – Rappresentazione alternativa di una numerazione modulo M (M=8)

L'uso di una numerazione modulo *M* impone un vincolo sul numero massimo *SUP* di PDU che si possono emettere (o ammettere) senza avere un riscontro per alcuna di esse. Il vincolo deriva dalla esigenza di evitare ambiguità nei riscontri. Questo numero massimo *SUP* deve essere non superiore a

$$SUP \leq M - 1.$$

Per convincersene, consideriamo il caso *M*=8 e mostriamo che il numero di PDU emesse senza riscontro non può superare il valore 7 in quanto, se questo limite non fosse rispettato, si avrebbe ambiguità nei riscontri. Supponiamo (senza perdita di generalità) che la prima PDU non riscontrata abbia *NS* = 0. Allora, chi emette sta ricevendo un riscontro *NR*= 0, cioè viene riscontrata positivamente

la PDU con *NS*= 7. Se si emettessero 8 PDU (compresa quella con *NS= 0*) e cioè se la sequenza di emissione fosse composta da PDU con i seguenti *NS*

0 1 2 3 4 5 6 7 ,

il riscontro *NR=0*, che per ipotesi continua ad essere ricevuto da chi emette, potrebbe essere interpretato come riscontro positivo per la nuova PDU con *NS= 7* e non per quella di ugual *NS* appartenente al ciclo precedente.

In una numerazione modulo *M*, si chiama *finestra* (window) la sequenza di numeri consecutivi che sono compresi tra un limite inferiore L_{inf} e un limite superiore L_{sup}, entrambi inclusi. I numeri consecutivi compresi tra L_{inf} e L_{sup} sono così "*messi in finestra*". La cardinalità *W* di detta sequenza è chiamata *larghezza della finestra*

$$W = (L_{sup} - L_{inf} + 1) \bmod M \ .$$

La finestra è *scorrevole* (sliding window) se si aumenta L_{inf} mantenendo costante la sua larghezza *W*. Ad ogni aggiornamento di L_{inf} la finestra *scorre in senso orario* e modifica corrispondentemente la sequenza di numeri messi in finestra (Fig. V.2.8).

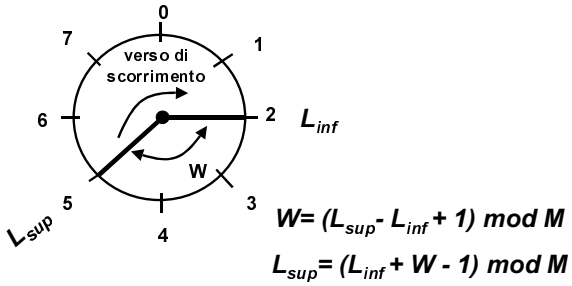


Figura V.2.8 – Finestra scorrevole

Una *finestra in emissione* è il numero massimo di PDU che una entità può *emettere sequenzialmente* senza ricevere riscontro per alcuna di esse. Questo numero massimo è la larghezza W_S della finestra in emissione. Il limite inferiore L_{inf} viene aggiornato dalla entità ricevente in relazione alla propria disponibilità ad ammettere nuove PDU. L_{inf} viene quindi posto uguale al numero di sequenza in ricezione *NR* emesso dall'entità ricevente.

La finestra in emissione può essere collocata in una qualsiasi entità *agente come emittente*; quindi, con riferimento alla relazione tra le entità *A* e *B*, supponiamo che questa finestra sia collocata in *A*. Allora i numeri di sequenza in emissione NS_A che *possono essere emessi da A* sono quelli compresi nel seguente intervallo

$$NR_B \le NS_A \le (NR_B + W_S - 1) \bmod M \quad ,$$

ove NR_B è un numero di sequenza in ricezione emesso da *B*. Ad ogni variazione di NR_B (aggiornamento dei riscontri positivi emessi da *B*) la finestra "scorre" consentendo l'"*uscita*" (cioè l'emissione) di nuove PDU da *A.* L'entità *B* può quindi controllare l'*emissione da parte di A* agendo sui propri riscontri (ad es. rallentandone o accelerandone l'emissione). Un esempio del funzionamento di una finestra in emissione collocata nell'entità A è offerto in Fig. V.2.9.

Una *finestra in ricezione* è il numero massimo di PDU che una entità può *ammettere* senza emettere riscontro per alcuna di esse. Questo numero massimo è la larghezza W_R della finestra in ricezione. Il limite inferiore L_{inf} viene aggiornato dalla entità ricevente *al suo interno* in relazione alla propria disponibilità ad ammettere nuove PDU. L_{inf} viene quindi posto uguale al numero di sequenza in ricezione *NR* emesso dalla stessa unità ricevente.

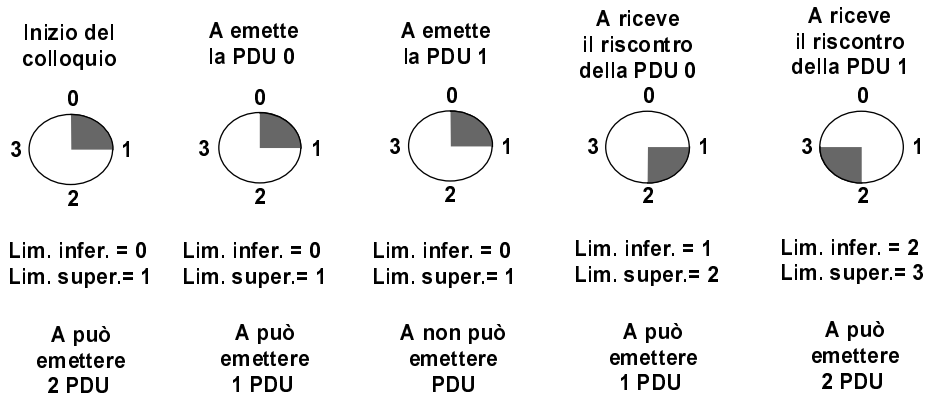


Figura V.2.9 – Finestra in emissione

La finestra in ricezione può essere collocata in una qualsiasi entità *agente come ricevente*; quindi, con riferimento alla relazione tra le entità *A* e *B*, supponiamo che questa finestra sia collocata in *A*. Allora i numeri di sequenza in emissione NS_B che *possono essere ammessi in A* sono quelli compresi nel seguente intervallo:

$$NR_B \le NS_A \le (NR_B + W_R - 1) \bmod M,$$

ove NR_A è un numero di sequenza in ricezione emesso da *A*. Ad ogni variazione di NR_A (aggiornamento dei riscontri positivi emessi da *A*), la finestra "scorre" consentendo l'ammissione di nuove PDU in *A*; l'entità *A* può quindi controllare *la propria ricettività* agendo sui propri riscontri (ad es. rallentandone o accelerandone l'emissione). Le PDU che pervengono in *A* e che sono verificate senza errore sono *ammesse* da *A* solo se i loro NS_B sono compresi *entro la finestra di ricezione*, anche se fuori sequenza. Invece le PDU che pervengono in *A al di fuori della finestra di ricezione* vengono *scartate*, anche se verificate senza errore. Una PDU *ammessa* viene *accettata* da *A* e quindi riscontrata positivamente solo se l'entità ricevente è in grado di collocarla in un corretto ordine sequenziale. Un esempio del funzionamento di una finestra in ricezione collocata nell'entità A è offerto in Fig. V.2.10.

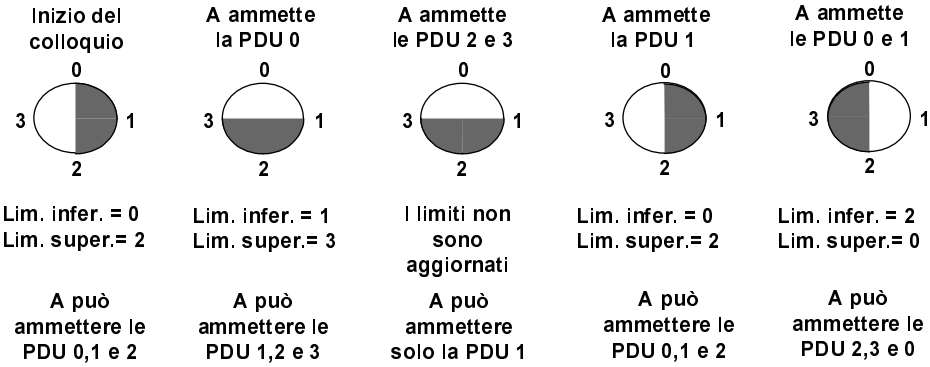


Figura V.2.10 - Finestra in ricezione

La larghezza delle finestre (in trasmissione o in ricezione) è limitata superiormente dal valore del modulo di numerazione *M*; deve infatti essere

$$W \le M - 1).$$

Quindi, anche in assenza di una *finestra esplicita* di larghezza *W<M-1*, esiste in ogni caso una *finestra implicita*, che ha larghezza uguale a *M-1*.

Se le larghezze delle due finestre sono maggiori di 1, tali larghezze sono soggette ad un ulteriore vincolo in relazione al valore del modulo di numerazione *M*. Questo ulteriore vincolo (in aggiunta a quello ora visto *W≤ M-1)* ha lo scopo di assicurare che l'entità ricevente sia in grado di indentificare senza ambiguità la PDU pervenutale e quindi di poterla riscontrare in modo univoco. Nel caso particolare in cui le due larghezze W_S e W_R sono uguali e maggiori di 1

$$W_S = W_R = W > 1,$$

la larghezza comune *W* deve rispettare il vincolo

$$W \leq M / 2,$$

cioè deve essere non superiore alla metà del modulo di numerazione. Con il rispetto di questo vincolo, si riesce infatti ad evitare che, ad ogni scorrimento della finestra in ricezione, il nuovo insieme di numeri di sequenza ammessi (*nuovi numeri*) si sovrapponga a quello valido prima dello scorrimento (*vecchi numeri*). Una eventuale sovrapposizione tra nuovi e vecchi numeri creerebbe nel ricevitore l'impossibilità di distinguere tra due casi che si possono presentare quando, dopo l'aggiornamento della finestra, perviene al ricevitore un *gruppo di PDU.* Queste possono essere il risultato di:

- *riemissioni,* se tutti i riscontri di PDU precedentemente ricevute sono andati persi;
- *nuove emissioni,* se tutti i riscontri sono stati ricevuti.

Si nota che questo vincolo non si presenta quando la larghezza della finestra in ricezione è *unitaria*; infatti in questa condizione, indipendentemente dalla larghezza W_S , si può avere aggiornamento della finestra in ricezione solo con un "*nuovo numero*", a cui può corrispondere solo una "nuova emissione".

V.2.4 Modello di recupero di errore

Supponiamo che:

- siano utilizzati tutti i mezzi di recupero precedentemente definiti;
- i riscontri positivi siano trasferiti con PDU-dati o con PDU-ACK;
- i riscontri negativi siano congetturati dall'*assenza di un riscontro positivo* entro un fissato *tempo-limite* (time-out) dall'istante di emissione della PDU da riscontrare; tale assenza di riscontro positivo è evidenziata dallo *scadere di un temporizzatore*;
- sia le PDU-dati che le PDU-ACK siano protette da un CRC;
- siano utilizzati entrambi i numeri di sequenza in emissione *NS* e in ricezione *NR*, che sono numerati modulo *M* (*M≥ 2*); questi numeri sono entrambi presenti nelle PDU-dati; nelle PDU-ACK è invece contenuto il solo *NR*;
- le PDU-dati debbano rispettare (attraverso i loro *NS*) i vincoli posti in emissione e in ricezione dalle rispettive finestre scorrevoli; le larghezze W_S e W_R di queste sono in generale oggetto di negoziazione tra le parti; per alcune procedure una o entrambe le larghezze sono unitarie;
- i numeri *NS* e *NR* inseriti nelle PDU siano definiti da due *variabili di stato VS* e *VR*, che
  - sono associate a ogni entità, agente come emittente e come ricevente;
  - assumono valori compatibili con una numerazione modulo *M*;
  - vengono incrementate via via in relazione all'evoluzione della procedura.

Si fa riferimento allo stato riguardante una entità con il duplice ruolo di emittente e di ricevente, e relativo a un istante qualunque nell'evoluzione della procedura; in questo stato:

- *VS* è la *variabile di stato in emissione*, che fornisce *il più grande* numero d'ordine *NS* della PDU che l'entità agente come emittente si propone di *formare*, accogliendo la relativa SDU dallo strato superiore, e curandone poi l'emissione;

- *VR* è la *variabile di stato in ricezione*, che fornisce *il più piccolo* numero d'ordine *NS* della PDU che l'entità agente come ricevente si propone di *accettare*, trasferendo la relativa SDU allo strato superiore.

La variabile *VS* viene incrementata di una unità per ogni nuova SDU *proveniente* in ordine sequenziale *dallo strato superiore*. La variabile *VR* viene incrementata di una unità in corrispondenza ad ogni SDU *trasferita* con corretta sequenzialità *allo strato superiore*.

Per ogni nuova PDU formata da una entità che agisce come emittente e che si trova in uno stato in cui le relative variabili abbiano le determinazioni *VS* e *VR*

- nel campo previsto per il *numero di sequenza in emissione* viene collocato $NS = VS$;
- nel campo previsto per il *numero di sequenza in ricezione* viene collocato $NR = VR$.

Con riferimento alla *finestra in emissione*, il limite inferiore L_{inf} indica il più piccolo numero di sequenza in emissione NS_{min} che non è stato ancora riscontrato. Poiché, per definizione,

$$L_{sup} = (NS_{min} + W_S - 1) \bmod M,$$

un incremento della variabile di stato in emissione *VS* è possibile solo se, in base allo stato attuale, risulta

$$(VS - NS_{min} + 1) \bmod M \le W_S .$$

In queste condizioni si può porre

$$VS \leftarrow (VS + 1) \bmod M,$$

e cioè incrementare *VS* di una unità.

Con riferimento alla *finestra in ricezione*, il limite inferiore L_{inf} indica il più piccolo numero di sequenza in emissione (proveniente da una entità emittente) che può essere accettato: coincide quindi con la variabile di stato in ricezione *VR*:

$$L_{inf} = VR.$$

Ogni qualvolta viene ricevuta una PDU:

- verificata senza errori,
- rientrante entro i limiti della finestra in ricezione,
- con numero d'ordine $NS = VR$,

la PDU viene accettata e trasferita allo strato superiore. Può allora porsi

$$VR \leftarrow (VR + 1) \bmod M,$$

e cioè incrementare *VR* di una unità.

V.2.5 Procedure di recupero di errore

Qualunque sia la procedura impiegata, è sempre richiesta la cooperazione dell'entità agente come ricevente (*entità ricevente*) con quella agente come emittente (*entità emittente*). Si assume il modello di recupero precedentemente descritto e si distinguono i compiti svolti dall'entità ricevente (*azioni del ricevitore*) e dall'entità emittente (*azioni dell'emettitore*).

L'entità *ricevente*, per ogni PDU proveniente dall'entità emittente

- verifica nella PDU la presenza o meno di errori, utilizzando il metodo di rivelazione che è adottato nella procedura; si hanno due casi
  (1) PDU *verificata senza errori*;
  (2) PDU *verificata con errori*.

Limitatamente al *caso (1)*, le entità ricevente

- *controlla* la PDU per rivelare la presenza o meno di altre situazioni anomale (ad es. perdite, duplicazioni, ecc.); il controllo è effettuato accertando in primo luogo se il relativo *NS* è all'interno o all'esterno della finestra in ricezione; si hanno due casi:

(a) PDU controllata *senza altre anomalie* (all'interno della finestra);
(b) PDU controllata *con altre anomalie* (all'esterno della finestra).

In relazione ai casi (1), (2), (a) e (b), l'entità ricevente

- *accetta* la PDU ricevuta solo quando si verifica l'unione dei due casi (1) e (a) e la PDU in questione è collocabile in un corretto ordine sequenziale;
- *scarta* la PDU ricevuta quando si presenta anche uno solo dei casi (2) o (b).

Limitatamente al caso di *PDU accettata,* l'entità ricevente

- *trasferisce* la relativa SDU allo strato superiore;
- *invia* all'entità emittente, con un possibile ritardo, un *riscontro positivo*.

Limitatamente al caso (1) in unione con quello (b) (cioè per una PDU verificata senza errori e controllata con altre anomalie) ovvero nel caso in cui non sia possibile ricostruire la corretta sequenzialità delle SDU pervenute,

- *invia* all'entità emittente un *riscontro negativo* esplicito.

L'*entità emittente*, per ogni SDU proveniente dallo strato superiore,

- *accoglie* questa SDU e contestualmente
- *inserisce* la SDU nella PDU;
- *codifica* la stringa da proteggere nei confronti degli errori;
- *attribuisce* alla PDU un opportuno numero di sequenza *NS* quando sia stato verificato che questo numero è all'interno della finestra in emissione;
- *emette* la PDU con il rispetto di due vincoli:
  - viene *rispettata*, nell'ordine di emissione, la *sequenzialità* delle SDU provenienti dallo strato superiore;
  - viene *attivato un temporizzatore*, e cioè il mezzo disponibile per evitare condizioni di stallo (ad es. dovute ad attesa indefinita di un riscontro positivo);
- *riemette* la stessa PDU il numero di volte necessario a ottenere un riscontro positivo.

Un primo protocollo di recupero di errore è chiamato *a riemissione con arresto e attesa* o "stop and wait". Il suo principio è il seguente: ogni entità nel ruolo di *emittente*

- emette una sola PDU alla volta;
- prima di emettere la successiva, attende un riscontro positivo;
- se non riceve questo riscontro, riemette la PDU e ripeterà questa operazione finché riceve un riscontro positivo.

Nel ruolo di *ricevente* una entità accetta PDU solo se verificate senza errori e controllate in corretta sequenza.

Con riferimento al modello di recupero precedentemente introdotto, si sottolineano le seguenti particolarità:

- il modulo di numerazione per *NS* e *NR* è uguale a 2;
- le finestre scorrevoli in emissione e in ricezione hanno entrambe *larghezza unitaria*.

L'evoluzione della procedura a riemissione con arresto e attesa è esemplificata in Fig. V.2.11.

Una seconda procedura di recupero di errore è chiamata *a riemissione continua o* "go back n", ove *n* è la larghezza della finestra in emissione. Il suo principio è il seguente: ogni entità quando agisce da *emittente*

- emette al massimo W_S PDU consecutive senza attendere il riscontro per alcuna di queste;
- prima di emettere un'altra PDU, attende un *riscontro positivo*, che può essere individuale o cumulativo;
- se per alcuna delle PDU emesse non riceve il riscontro positivo, provvede a riemettere questa PDU e tutte le seguenti emesse e in attesa di riscontro;

- se riceve dall'entità ricevente una PDU che segnala un fuori-sequenza non recuperabile (*riscontro negativo*), provvede a ristabilire la sequenzialità riemettendo la prima PDU non riscontrata e tutte le seguenti già emesse e in attesa di riscontro.

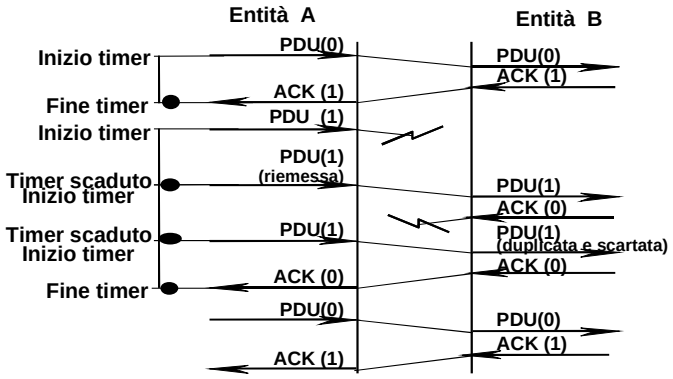


Figura V.2.11 – Evoluzione della procedura a riemissione con arresto e attesa

Ogni entità quando agisce da *ricevente* opera nello stesso modo dell'entità ricevente nel protocollo "riemissione con arresto e attesa", cioè accetta solo PDU-dati che siano verificate senza errore e che siano in corretta sequenza.

Con riferimento al modello per le procedure di recupero, si sottolineano le seguenti particolarità:

- il modulo di numerazione per *NS* e *NR* ha un valore *M*, che costituisce uno dei parametri del protocollo aventi incidenza importante sulle sue prestazioni;
- la finestra scorrevole in ricezione ha larghezza di *valore unitario*;
- la finestra scorrevole in emissione ha larghezza W_S , il cui valore può essere fissato nella inizializzazione della procedura con l'unico vincolo che sia

$$W_S \leq M - 1.$$

L'evoluzione della procedura a riemissione continua è esemplificata in Fig. V.2.12

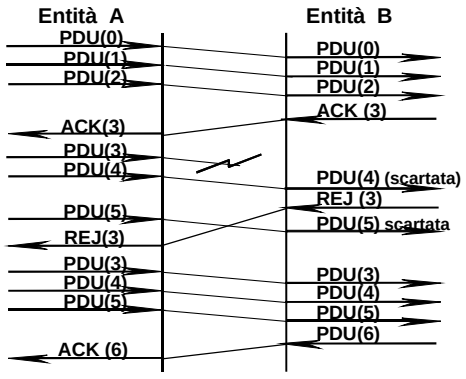


Figura V.2.12 – Evoluzione della procedura a riemissione continua

La terza procedura di recupero di errore qui considerata è chiamata *a riemissione selettiva* o "selective repeat". Il suo principio è il seguente: ogni *entità emittente* si comporta come nel caso del protocollo a riemissione continua, con la differenza che, nel caso di assenza di riscontro per una specifica PDU, provvede a riemettere solo questa PDU. Una *entità ricevente*, a differenza di quella con lo stesso ruolo nei protocolli "riemissione con arresto e attesa" e "riemissione continua",

- ammette PDU verificate senza errore e contenute nella finestra in ricezione, anche se fuori-sequenza;
- cerca poi di ricostruire la corretta sequenzialità delle PDU ammesse e, se non ci riesce, invia un *riscontro negativo*.

Con riferimento al modello per le procedure di recupero, si hanno le seguenti particolarità:

- le larghezze delle due finestre sono assunte uguali

$$W_S = W_R = W > 1;$$

- il loro valore comune deve essere scelto con il vincolo $W \leq M/2$, in modo da assicurare che l'entità ricevente sia in grado di identificare senza ambiguità le PDU pervenutegli.

L'evoluzione della procedura a riemissione selettiva è esemplificata in Fig. V.2.13.

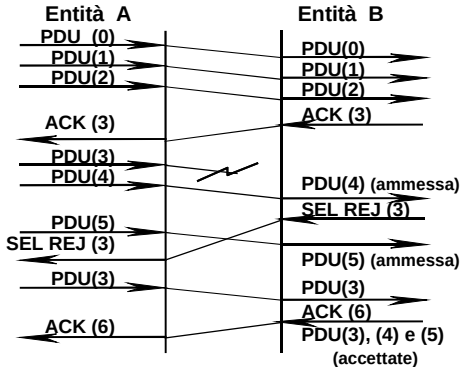


Figura V.2.13 – Evoluzione della procedura a riemissione selettiva

V.2.6 Efficienza di recupero

L'efficienza di una procedura di recupero viene qui valutata mediante il *rendimento di trasferimento* ρ, che è definito da

$$\rho = \frac{\theta_I}{\theta_R}$$

ove

θ_I è il valore atteso del tempo necessario a trasferire i dati di utente in una PDU tendendo conto del solo ritmo binario di trasferimento;

θ_R è il valore atteso del tempo necessario a trasferire una PDU, tenendo conto, oltre che del ritmo binario di trasferimento, anche

- della extra-informazione legata alla PCI della PDU;
- dei vincoli posti dalla procedura di recupero in assenza di errori;
- delle riemissioni della PDU per ottenere un trasferimento senza errori.

Per brevità, le procedure di recupero saranno indicate con

P1 "*riemissione con arresto e attesa*";
P2 "*riemissione continua*";
P3 "*riemissione selettiva*".

V.2.6.1 Modello per calcolare l'efficienza di recupero

Nel calcolo di ρ per queste tre procedure, si farà riferimento ai seguenti tre casi di complessità crescente:

caso a) : trasferimento in assenza di errori e di una qualche procedura di recupero;
caso b) : trasferimento in assenza di errori e con una procedura di recupero;
caso c) : trasferimento in presenza di errori e con una procedura di recupero.

Ovviamente le espressioni di ρ desiderate sono quelle relative al caso c).

Denotiamo con

F	la lunghezza del campo-dati di una PDU;
H	la lunghezza della PCI nella PDU;
β	il rapporto H/F (*quota di extrainformazione*);
L	la lunghezza complessiva di una PDU $[L = F+ H = F (1+ \beta)]$;
D	il ritardo di propagazione tra emettitore e ricevitore;
R	il ritmo binario di trasferimento;
ε	la probabilità che una PDU sia colpita da errore;
T	il tempo di ciclo, e cioè l'intervallo di tempo tra l'emissione del primo bit di una PDU e la ricezione del relativo riscontro;
Q	il numero medio di emissioni/riemissioni per ottenere il trasferimento di una PDU senza errori.

Supponiamo che

- le PDU abbiano lunghezza costante;
- per i riscontri, si impieghi la tecnica dell'addossamento, si trascuri la lunghezza della parte di PDU dedicata al riscontro e si adotti una emissione senza ritardi;
- i riscontri siano solo individuali;
- ogni emettitore operi in condizioni di pieno carico (cioè con esistenza continua di una esigenza di trasferimento).

Conseguentemente per tutte e tre le procedure P1, P2 e P3

$$\theta_I = \frac{F}{R}$$

$$T = \frac{L}{R} + 2D = \frac{L}{R}\left(1 + \frac{2RD}{L}\right)$$

ove il rapporto $2RD/L$ è la *lunghezza andata-ritorno del mezzo di comunicazione* espressa in PDU, (ovvero il numero di PDU che il mezzo di andata-ritorno può contenere).

Il rapporto RD/L verrà indicato con α in accordo con la notazione introdotta in § V.1.4

$$\alpha = \frac{RD}{L}$$

e quindi

$$T = \frac{L}{R}(1 + 2\alpha) = \theta_I (1 + \beta)(1 + 2\alpha).$$

V.2.6.2 Finestre critiche

Con riferimento alle sole procedure P2 e P3 e nell'ipotesi (già fatta) che si operi in condizioni di pieno carico, sono ora valutate le condizioni affinché l'*emissione avvenga senza soluzione di continuità*. Per questo scopo è sufficiente che l'emettitore non si arresti per l'evento "la finestra di

emissione si è riempita e non è ancora pervenuto un riscontro per la prima PDU non ancora riscontrata". Dato che con le altre ipotesi fatte

- il tempo di riscontro Δ_1 è uguale a $2D$;
- l'intervallo di tempo Δ_2 per riempire $W-1$ posizioni della finestra è uguale a $(W-1)\,L/R$,

la condizione di emissione senza soluzione di continuità si esprime imponendo che il tempo Δ_1 sia non superiore all'intervallo Δ_2 ; quindi

$$2D \le (W-1)\frac{L}{R};$$

da cui si ricava

$$W \ge 1+2\alpha.$$

Pertanto, per ottenere il risultato desiderato, è sufficiente assumere una larghezza di finestra che sia *il più piccolo intero non inferiore a* $1+2\alpha$. Si osserva che, se

$$W \ge 1+2\alpha \quad (caso\ \delta_1),$$

la finestra in emissione non si riempie mai (anche in condizioni di pieno carico) dato che, come si è imposto, arriva un riscontro prima del riempimento. In questo caso e in assenza di errori, il trasferimento avviene senza alcun vincolo posto dalle procedure P2 e P3.

Se invece

$$W < 1+2\alpha \quad (caso\ \delta_2),$$

l'emettitore, dopo aver emesso W PDU e quindi aver riempito la finestra, deve fermarsi e attendere che arrivi il riscontro per la prima di queste PDU. In questo caso e in assenza di errori, le procedure P2 e P3 (che da questo punto di vista producono lo stesso effetto) introducono un vincolo sul trasferimento. Pertanto, nell'analisi del caso (b) e senza necessità di distinguere tra le procedure P2 e P3, occorre considerare due sottocasi:

- *sottocaso (b-*δ_1*)*: unione dei casi (b) e (δ_1)
- *sottocaso (b-* δ_2*)*: unione dei casi (b) e (δ_2).

Inoltre, dato che

- la presenza degli errori produce lo stesso effetto nei casi (δ_1) e (δ_2), ma in modo differenziato per le procedure P2 e P3;
- la determinazione di ρ nel caso (c) si ottiene modificando l'espressione di ρ nel caso (b) con un fattore moltiplicativo,

anche nell'analisi del caso (c) e con distinzione tra le procedure P2 e P3, occorre considerare 2 sottocasi

- *sottocaso (c-*δ_1*):* unione dei casi (c) e (δ_1)
- *sottocaso (c-* δ_2*)*: unione dei casi (c) e (δ_2) .

V.2.6.3 Efficienza della procedura "riemissione con arresto e attesa"

Calcoliamo ρ *nel caso (a)* e cioè in assenza di recupero. Poiché

$$\theta_{Ra} = \frac{L}{R} = \theta_1(1+\beta)$$

si ottiene

$$\rho_a = \frac{1}{1+\beta}$$

Circa l'espressione di ρ nel *caso (b)* per la procedura P1

$$\theta_{Rb} = T = \theta_I (1+\beta)(1+2\alpha)$$

$$\rho_b = \frac{1}{(1+\beta)(1+2\alpha)} \quad .$$

Passando a determinare ρ nel *caso (c)* per la procedura P1, si ha

$$\theta_{Rc} = Q\theta_{Rb} = Q\theta_I (1+\beta)(1+2\alpha)$$

ove

$$Q = \sum_{i=1}^{\infty} i\varepsilon^{i-1}(1-\varepsilon) = \frac{1}{1-\varepsilon};$$

conseguentemente

$$\rho_c = \frac{\rho_b}{Q} = \frac{1}{(1+\beta)(1+2\alpha)}(1-\varepsilon).$$

V.2.6.4 Efficienza della procedura “riemissione continua”

Dapprima si calcola ρ nel *sottocaso (b-δ_1)* ($W \geq 1+2\alpha$) per le procedure P2 e P3. Dato che, in assenza di errori, le procedure P2 e P3 non vincolano il trasferimento, risulta

$$\rho_b = \rho_a = \frac{1}{1+\beta} \quad .$$

Si passa poi a calcolare ρ nel *sottocaso (b-δ_2)* ($W < 1+2\alpha$) per le procedure P2 e P3. Dato che, in assenza di errori e dopo aver emesso W PDU, l’emettitore deve fermarsi e attendere un riscontro, si ottiene

$$\theta_{Rb} = \frac{T}{W} = \frac{1}{W}\theta_I (1+\beta)(1+2\alpha)$$

da cui

$$\rho_b = \frac{W}{(1+\beta)(1+2\alpha)} \quad .$$

Si osserva poi che, per la procedura P2, se si indica con $K \leq W$ il numero di PDU entro la finestra, deve essere

$$K\frac{L}{R} \geq T \quad e \quad (K-1)\frac{L}{R} < T$$

ovvero

$$(1+2\alpha) \leq K < 2(1+\alpha).$$

Inoltre

$$Q = \sum_{i=0}^{\infty} (iK+1)\varepsilon^i (1-\varepsilon) = 1 + K\frac{\varepsilon}{1-\varepsilon} =$$

$$= \frac{1}{1-\varepsilon}\left[1+\varepsilon(K-1)\right] = \frac{1+2\alpha\varepsilon}{1-\varepsilon}$$

ove l’ultima uguaglianza è stata ottenuta approssimando K con $1+2\alpha$; ne seguono le espressioni di ρ_c per la procedura P2.

Quindi ρ nel *sottocaso (c-δ_1)* ($W \geq 1+2\alpha$) per la procedura P2 è dato da

$$\rho_c = \frac{1}{1+\beta}\,\frac{1-\varepsilon}{1+2\alpha\varepsilon}.$$

Invece ρ nel *sottocaso (c-δ_2)* ($W<1+2\alpha$) per la procedura P2 è espresso da

$$\rho_c = \frac{W}{(1+\beta)(1+2\alpha)}\,\frac{1-\varepsilon}{1+2\alpha\varepsilon} \quad .$$

V.2.6.5 Efficienza della procedura “riemissione selettiva”

Poiché in questo caso, con ragionamento analogo a quello per la procedura P1, si ottiene

$$Q = \frac{1}{1-\varepsilon} \quad ,$$

ne derivano le espressioni di ρ_c per la procedura P3. Con riferimento al *sottocaso (c-δ_1)* ($W\geq 1+2\alpha$) per la procedura P3, si ricava

$$\rho_c = \frac{1}{1+\beta}\,(1-\varepsilon) \quad .$$

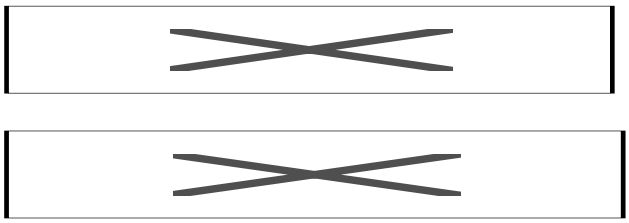
Infine, nel *sottocaso (c-δ_2)* ($W<1+2\alpha$) sempre per la procedura P3, si ottiene

$$\rho_c = \frac{W}{(1+\beta)(1+2\alpha)}\,(1-\varepsilon) \quad .$$

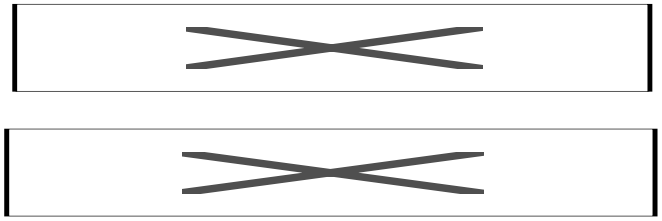
V.2.6.6 Commenti conclusivi

Riassumendo, per la procedura P1 (*ritrasmissione con arresto e attesa*) e nel caso c), il rendimento di trasferimento è dato da

Invece, sempre nel caso c), per la procedura P2 (*ritrasmissione continua*), il rendimento ρ è espresso da



Infine, ancora nel caso c), la procedura P3 (*ritrasmissione selettiva*) offre un rendimento di trasferimento che è dato da



L’efficienza delle tre procedure in assenza di errori è *limitata*, in primo luogo e in egual misura, dalla *quota* β di extra-informazione rispetto all’informazione utile.

Sempre in assenza di errori, l'efficienza della procedura P1 è ulteriormente *limitata dal parametro* α: se questo ha valore superiore a qualche decimo, il rendimento del trasferimento ne risulta sensibilmente penalizzato e questa penalizzazione è crescente al crescere di β.

L'efficienza delle procedure P2 e P3 in assenza di errori dipende anch'essa da α, in quanto questo parametro condiziona la larghezza della finestra critica ($1+2\alpha$) in emissione e quindi anche il valore del modulo di numerazione; se viene assunta una larghezza di finestra maggiore o uguale di quella critica, non si ha una ulteriore dipendenza da α. Dal punto di vista degli errori e in aggiunta alle limitazioni sopra descritte, l'efficienza della procedura P3 (come del resto quella della procedura P1) è limitata da $1-\varepsilon$, e cioè direttamente dalla probabilità che il trasferimento di una PDU avvenga senza errore. Si tratta di una limitazione, che ben può definirsi "fisiologica", dato che solo le PDU senza errore possono contribuire alla portata media del trasferimento.

Sempre per ciò che riguarda gli errori, la procedura P2 presenta una limitazione aggiuntiva rispetto a P3; la presenza del termine $1+2\alpha\varepsilon$ a denominatore di ρ ci sottolinea che:

- in P2, ogniqualvolta si manifesta un errore, vengono ritrasmesse tutte le PDU contenute nel mezzo di comunicazione avanti-indietro;
- ρ dipende dal prodotto $2\alpha\varepsilon$, nel senso che, al crescere di α o di ε o di entrambe queste quantità, si ha una sua penalizzazione;
- la procedura P2 presenta un rendimento significativamente peggiore di quello di P3 solo quando il prodotto $2\alpha\varepsilon$ è apprezzabilmente prossimo a 1.

V.3 Indirizzamento

L'indirizzo di un utente identifica in modo univoco quest'ultimo nell'ambito di una rete di telecomunicazione. La modalità con cui sono definiti ed assegnati gli indirizzi (*piano di numerazione*) ha significative conseguenze sulla *funzione di instradamento*, e cioè la funzione il cui scopo è guidare l'informazione di utente verso la destinazione voluta. Si possono distinguere tre modalità di indirizzamento:

1- nella prima l'indirizzo è stabilmente legato ad un luogo fisico e il piano di indirizzamento è definito in modo tale che l'indirizzo, oltre a identificare un utente, dia anche informazioni su dove lo stesso si trovi;
2- nella seconda l'indirizzo di per se stesso non consente di dedurre in modo immediato la localizzazione dell'utente, anche se esiste una corrispondenza stabile tra l'utente e il luogo ove esso si trova;
3- nella terza non esiste una corrispondenza stabile tra indirizzo e luogo fisico in cui l'utente si trova in un dato momento.

Nel caso di cui al precedente punto 2 l'algoritmo di instradamento deve prima stabilire *dove* l'utente si trova e quindi scegliere una strada per raggiungerlo. Nel terzo caso bisogna prima stabilire *dove* l'utente si trova *in un dato momento* e quindi scegliere una strada per raggiungerlo.

Per esemplificare prendiamo in considerazione diversi casi, in ordine crescente di complessità, essendo il caso più semplice quello in cui un indirizzo ci da' direttamente informazioni stabili sulla localizzazione della destinazione voluta:

- nella rete telefonica fissa gli indirizzi (=numeri telefonici) sono organizzati in modo gerarchico per cui un indirizzo, oltre ad identificare uno specifico utente, fornisce precise informazioni sulla localizzazione dello stesso; ad esempio il numero telefonico "39 06 44 58 XXXX" è legato ad uno specifico luogo (e apparato telefonico): 39 è l'identificativo dell'Italia, 06 il prefisso per la città di Roma, 44 può specificare una particolare centrale di

commutazione, 58 un centralino privato e XXXX un interno del centralino; quando all'algoritmo di instradamento giunge una richiesta di connessione verso tale numero è immediato sapere dove si trova la destinazione;

- nell'attuale Internet basata sul protocollo IP senza prestazioni di mobilità, l'associazione tra indirizzo e luogo fisico in cui l'utente si trova può essere considerata relativamente stabile ma l'indirizzo fornisce limitate informazioni su dove il relativo utente si trova (il piano di indirizzamento non è organizzato in modo gerarchicamente legato alla localizzazione territoriale degli utenti: non esistono prefissi internazionali, interurbani etc.); ciò significa che parte integrante dell'algoritmo di instradamento è anche stabilire la localizzazione dell'utente; si noti però che una volta stabilita l'associazione indirizzo-luogo questa può essere assunta stabile, a meno di cambiamenti che possono però avvenire solo in seguito ad operazioni di carattere gestionale e non quindi sotto il controllo di un generico utente;
- si noti che un caso simile si presenta anche nella rete telefonica fissa quando si usufruisce del servizio "numero verde"; in tal caso l'indirizzo (del tipo 167 X...X) non rientra nel piano gerarchico di indirizzamento, non si usano i consueti prefissi inter-urbani e non è possibile, esaminando il numero, stabilire la localizzazione del relativo utente di destinazione; in tal caso devono quindi essere invocate opportune procedure (nella fattispecie ottenute mediante la cosiddetta Rete Intelligente) per ottenere la corrispondenza indirizzo->localizzazione prima di poter iniziare a scegliere una strada opportuna;
- nelle reti radio-mobile cellulari un indirizzo è in corrispondenza con un dato utente ma la posizione di quest'ultimo varia nel tempo; l'algoritmo di instradamento deve quindi stabilirne la localizzazione in modo dinamico, prima di poter scegliere una strada attraverso la quale fargli pervenire una data informazione; infatti, un numero telefonico di un terminale cellulare non fornisce alcuna indicazione su dove il relativo utente si trovi in un certo momento.

L'evoluzione delle reti di telecomunicazione ha reso inoltre la questione appena discussa ancora più complessa. Nei classici paradigmi di comunicazione vi è infatti la tendenza ad identificare un terminale con l'utente (o gli utenti) che ne fa (fanno) uso e l'indirizzo si riferisce al terminale ma anche, per estensione, all'utente. Il concetto di mobilità è stato però esteso dalla mobilità di terminale e cioè dalla possibilità di accedere ad un servizio di telecomunicazione con un terminale in movimento (vedi punto precedente) alla mobilità personale. Nel primo caso l'indirizzo è legato ad un terminale fisico; quest'ultimo può spostarsi da un luogo ad un altro ma permane un'associazione tra terminale ed utente, ovvero l'indirizzo è legato ad uno specifico terminale. Nel secondo caso (mobilità personale) l'indirizzo è legato all'utente che può quindi registrarsi su diversi terminali, ovvero usare diversi terminali mantenendo la propria identità.

Il sistema GSM è un esempio di tale nuovo concetto di mobilità: un utente è identificato da un indirizzo contenuto in una carta: inserendo quest'ultima in un terminale si personalizza il terminale. Le chiamate indirizzate verso quell'utente saranno instradate sul terminale su cui l'utente stesso si è registrato e la tariffazione delle chiamate effettuate da un terminale sarà addebitata all'utente registrato su quel terminale. Tale possibilità è in progetto di essere estesa anche alle reti non mobili: un utente potrà registrarsi su qualunque terminale, sia fisso che mobile, e la rete, per instradare l'informazione verso un dato utente, dovrà determinare su quale terminale un dato utente è registrato, poi dove quel terminale si trova in quel dato momento (se si tratta di un terminale mobile) e quindi scegliere una strada che porti a quel terminale.

V.4 Instradamento

L'instradamento è una funzione di natura logica avente lo scopo di guidare l'informazione di utente verso la destinazione voluta. Decidere un criterio di instradamento significa stabilire come scegliere un percorso attraverso la rete logica in modo tale che l'informazione emessa da un utente giunga ad un altro utente *rispettando* opportuni requisiti prestazionali. Più in dettaglio si può dire che l'instradamento è una funzione decisionale che si attua con lo svolgimento di un opportuno algoritmo: per ogni coppia origine-destinazione occorre individuare un percorso, soddisfacendo determinati criteri prestazionali sia dal punto di vista degli utenti, (qualità del servizio da questi percepita: ritardo di trasferimento, perdita di informazione, eventuale ritardo di instaurazione, etc.) sia da quello degli operatori di rete (efficienza di utilizzazione delle risorse: traffico totale che la rete può trasportare, resistenza ai guasti, economicità e semplicità di implementazione e di gestione, etc.).

La questione della localizzazione di un indirizzo, trattata nel par. V.3, può essere isolata come una prima parte della funzione di instradamento (alcuni la considerano una funzione preventiva all'instradamento vero e proprio). Quale che sia il piano di indirizzamento usato esisterà dunque una prima fase (più o meno complessa) in cui dall'indirizzo si risalirà alla localizzazione dell'utente destinatario nell'ambito di una data rete. A questo punto la funzione di instradamento può scegliere il percorso migliore verso la località precedentemente individuata.

V.4.1 Generalità sugli algoritmi di instradamento

Con queste premesse, e assumendo risolto il problema della localizzazione dell'utente destinatario (problema che è risolto in modo specifico all'interno di determinate reti di telecomunicazione), ci si concentrerà nel seguito nel problema generale di individuare un percorso da una data origine ad una data destinazione. Si noti che, in una certa situazione, potrebbe esistere più di un percorso che soddisfa i requisiti voluti; in tal caso, tra i percorsi possibili, si può scegliere quello che ottimizza determinati parametri.

La funzione di instradamento è complessa per tre ordini di motivi:

- richiede un coordinamento tra diversi sistemi ed i sistemi coinvolti possono essere molto numerosi;
- deve poter far fronte ad eventuali guasti e malfunzionamenti sia dei nodi che dei rami di rete, eventualmente re-instradando il traffico per cui sono state già prese decisioni di instradamento e tenendo conto delle mutate condizioni per il traffico da instradare in seguito;
- deve, ove possibile e conveniente, tenere in conto lo stato di occupazione delle risorse di rete in modo dinamico e/o adattativo; ciò al fine di scegliere i percorsi evitando porzioni di rete che risultino in congestione, così da ottimizzare sia l'efficienza di utilizzo delle risorse stesse sia le prestazioni percepite dagli utenti.

La funzione di instradamento comporta anche una funzione ancillare e cioè la distribuzione a tutti i nodi di rete delle informazioni necessarie per svolgere tale funzione. In altri termini bisogna che ogni nodo di rete sia reso edotto delle informazioni di cui ha necessità per poter prendere decisioni di instradamento; tali informazioni possono essere:

- dati non preventivamente elaborati e riguardanti, ad esempio, l'occupazione delle risorse, la topologia di rete, i guasti e le riparazioni di elementi di rete; in tal caso ogni nodo elabora localmente le decisioni di instradamento, sulla base dei dati ricevuti. In tal caso si parla di algoritmi di instradamento *distribuiti*;
- informazioni già elaborate da altri sistemi di rete che forniscono quindi direttamente istruzioni sulle scelte da prendere. In tal caso si parla di algoritmi di instradamento *centralizzati*.

Tale funzione può essere considerata un problema aggiuntivo e a se stante ma che nondimeno risulta di considerevole importanza: la rete deve trasportare non solo le informazioni di utente ma anche altre informazioni necessarie ai nodi di rete per svolgere la funzione di instradamento.

Le prestazioni di rete sono fortemente influenzate dalla funzione di instradamento sia in termini della *quantità di servizio* (portata, di precipuo interesse dell'operatore di rete) sia in termini della *qualità di servizio* (ritardo di trasferimento ed eventuale perdita di informazione, parametri direttamente percepiti dagli utenti). La funzione di instradamento interagisce con la funzione di controllo di congestione nel determinare tali misure prestazionali mediante un meccanismo a contro reazione mostrato in Fig. V.4.1.

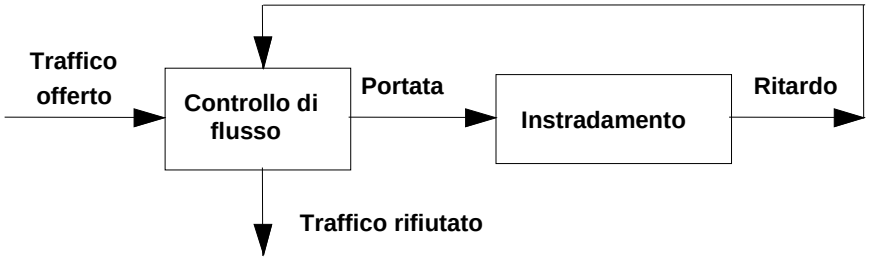


Figura V.4.1 - Interazione tra instradamento e controllo di flusso

Quando il traffico offerto alla rete è considerato eccessivo, il controllo di congestione ne rifiuterà una parte e le prestazioni viste dal traffico accettato saranno influenzate dalla funzione di instradamento. D'altra parte il controllo di congestione decide se accettare o meno il traffico sulla base delle prestazioni che gli utenti percepiranno (ad es. il ritardo di trasferimento) per cui, se la funzione di instradamento opera in modo adeguato, essa riuscirà a mantenere basso il ritardo di trasferimento e il controllo di congestione accetterà una maggiore quantità di traffico. In altri termini lo scopo della funzione di instradamento è anche quello di fare operare il controllo di congestione lungo una curva qualità-portata (ad esempio ritardo-portata) che sia la più favorevole possibile (Fig. V.4.2); ovvero aumentare la portata, a parità di ritardo di trasferimento subito dagli utenti in condizioni di alto traffico, e diminuire il ritardo, in condizioni di basso traffico.

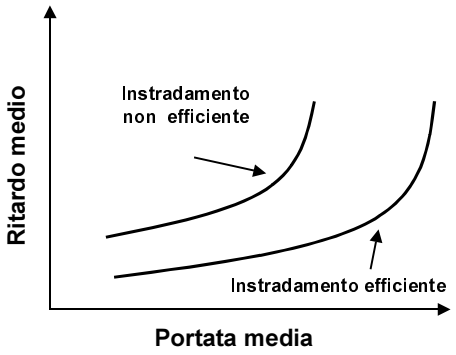


Figura V.4.2 - Interazione tra instradamento e controllo di flusso

Entrando ora in maggiore dettaglio, si può definire l'instradamento come quella funzione decisionale, svolta in ogni nodo di rete, ed avente lo scopo di stabilire il ramo di uscita verso cui deve essere inoltrata una unità informativa (UI) che perviene da un dato ramo di ingresso.

Per approfondire questa definizione, è opportuno distinguere i due casi di servizio di trasferimento con connessione e senza connessione. Nel secondo caso (che si verifica, ad esempio, in un modo di trasferimento a pacchetto con servizio a datagramma) ogni UI viene considerata come entità a se stante

e viene instradata nella rete in maniera indipendente dalle altre UI che hanno uguali origine e destinazione. Ciò equivale a dire che la funzione di instradamento viene operata per ogni singola UI che attraversa il nodo.

Nel caso invece di servizio di trasferimento con connessione (quale si ha in un modo di trasferimento a circuito o in uno a pacchetto con servizio a chiamata virtuale) la funzione di instradamento viene attivata solo durante la fase di instaurazione della connessione e definisce, nodo per nodo, il *percorso di rete* che devono seguire *tutte* le UI trasferite nell'ambito della comunicazione stessa.

In ambedue i casi, una procedura di instradamento consiste nel fissare le regole che rendono possibile il soddisfacimento delle *richieste di traffico*, facendo uso delle *risorse di rete* in modo da garantire opportune *prestazioni di trasferimento*. L'efficacia di una procedura di instradamento è quindi legata alla sua capacità di fare fronte a situazioni mutevoli nel tempo sia a causa delle variazioni delle richieste provenienti dagli utenti della rete, sia a causa della vulnerabilità (ad es. per effetto di guasti o di altre cause di fuori-servizio) delle risorse di cui la rete dispone.

Dal punto di vista di questa caratteristica, si possono allora distinguere procedure di instradamento *statiche*, ovvero *dinamiche*, ovvero infine *adattative*.

Le procedure di instradamento statiche sono basate su regole che, in generale, assumono come invarianti nel tempo sia la topologia della rete (e cioè le sue risorse), sia le caratteristiche delle richieste di traffico di accesso. In base a queste informazioni, in ogni nodo della rete vengono predisposte opportune tabelle (*tabelle di instradamento fisse*), che consentono di operare, di volta in volta, la decisione di instradamento. In queste tabelle è possibile prevedere, oltre a un instradamento-base, anche instradamenti alternativi da scegliere nel caso in cui si verifichino modifiche dello stato della rete tali da rendere non conveniente o impossibile la scelta dell'instradamento-base.

Le procedure di instradamento dinamiche utilizzano invece tabelle che variano nel tempo, mentre quelle adattative dipendono da certe stime sullo stato della rete nell'istante in cui viene presa la decisione di instradamento. Notiamo che una procedura di instradamento dinamica non è necessariamente adattativa, dato che la variabilità nel tempo delle tabelle di instradamento può essere di tipo programmato (con variazioni a livello giornaliero, settimanale, mensile, ecc.) senza l'effettuazione di stime sullo stato della rete. Invece un instradamento adattativo è normalmente dinamico, a meno che lo stato della rete rimanga immutato nel tempo.

Nelle procedure di instradamento adattative è possibile distinguere vari gradi di adattività a seconda della rapidità con cui le relative regole vengono aggiornate per fare fronte alle mutate situazioni in cui si opera. In generale, è conveniente limitare questa rapidità di aggiornamento, in quanto la raccolta dei dati sulle modifiche dello stato della rete implica un traffico addizionale che può limitare gli altri flussi informativi di utente e/o di segnalazione attraverso la rete. Anche per le procedure di instradamento adattative sono impiegate tabelle, una per ogni nodo della rete, che, a differenza di quelle fisse nell'instradamento statico, possono modificarsi nel tempo.

Come è evidente, le procedure di instradamento adattative sono più evolute rispetto a quelle statiche, in quanto consentono di conseguire migliori prestazioni globali. Sono però anche più complesse e quindi più costose e ciò ne ha finora limitato l'impiego alle reti di concezione più recente, quali, ad esempio, le reti che operano con modo di trasferimento orientato al pacchetto (nel seguito indicate per brevità come reti a pacchetto). Nelle reti di impostazione sistemistica più tradizionale, quali ad esempio le reti telefoniche, si sono fino ad oggi impiegate procedure di instradamento statiche. E tuttavia iniziato, anche sulla base dell'aumentata capacità elaborativa presente negli autocommutatori di tecnica più recente, l'impiego di procedure di instradamento dinamiche.

Le procedure di instradamento da adottare in reti a pacchetto dovrebbero idealmente possedere i seguenti attributi:

- *semplicità computazionale*; cioè l'algoritmo di instradamento dovrebbe impegnare in misura minima la capacità di elaborazione dei nodi della rete;
- *robustezza*; cioè dovrebbe essere garantita la possibilità dell'algoritmo ad adattarsi ai cambiamenti dei livelli di flusso di traffico attraverso la rete e a trovare percorsi di rete alternativi quando i nodi e/o i rami si guastano o ritornano in servizio dopo la riparazione;
- *stabilità*; cioè l'algoritmo dovrebbe essere in grado di convergere verso una soluzione accettabile senza oscillazioni eccessive quando si adatta ai cambiamenti di traffico e di topologia;
- *equità*; cioè le prestazioni di qualità di servizio assicurate dall'algoritmo dovrebbero essere fornite senza discriminazioni tra gli utenti, entro i possibili vincoli di assegnate priorità;
- *ottimalità*; cioè l'algoritmo dovrebbe individuare il percorso di rete più conveniente per minimizzare il valore medio del ritardo di trasferimento di un pacchetto e per massimizzare la portata della rete; particolare attenzione deve essere rivolta anche alla continuità di servizio e ai connessi problemi di affidabilità del trasferimento.

Per rispondere a questo insieme di esigenze, le procedure di instradamento che sono state proposte e, in alcuni casi, realizzate per reti a pacchetto sono normalmente di tipo adattativo. In questo ambito si distinguono procedure con controllo *centralizzato* da quelle con controllo *distribuito*.

Le procedure di instradamento con controllo centralizzato prevedono l'esistenza di un *centro di controllo di rete* (CCR), al quale confluiscono le informazioni riguardanti sia il livello del traffico di ingresso alla rete, sia l'integrità e il grado di occupazione dei rami e dei nodi. Il CCR elabora queste informazioni secondo l'algoritmo di instradamento che è stato prescelto, aggiorna di conseguenza le tabelle relative ai vari nodi della rete e provvede a trasferirne i contenuti a questi ultimi.

Nel caso di soluzione centralizzata, si può conseguire più agevolmente la condizione di ottimalità della procedura di instradamento, sollevando al tempo stesso i nodi di rete dai compiti di elaborazione del relativo algoritmo. Questo può allora essere di maggiore complessità computazionale. Lo svantaggio principale di un controllo centralizzato risiede nella vulnerabilità del sistema, dato che le conseguenze di un guasto del CCR possono essere disastrose ai fini della continuità del servizio: la rete può infatti continuare a funzionare basandosi sull'ultimo aggiornamento fornito dal CCR prima del suo guasto, ma questa informazione diventa rapidamente priva di utilità, con la conseguenza che le prestazioni del trasferimento diventano inaccettabili.

Nel caso di procedure di instradamento adattative con controllo distribuito, ogni nodo provvede a raccogliere le informazioni per lui accessibili sullo stato della rete e provvede a elaborarle per aggiornare le proprie tabelle; conseguentemente ogni decisione di aggiornamento viene presa localmente. Per queste procedure, si possono distinguere i casi in cui ogni nodo:

- opera senza interazioni con gli altri nodi, limitandosi a utilizzare le informazioni sullo stato della rete che possono essere raccolte localmente (*algoritmi isolati*);
- aggiorna la propria tabella di instradamento sulla base di dati scambiati con gli altri nodi e, più in particolare, con quelli ad esso adiacenti (*algoritmi cooperativi*).

In ambedue i casi di controllo centralizzato o distribuito (con algoritmo cooperativo) sorge, come già detto, la necessità di limitare il traffico di aggiornamento. Una soluzione è rappresentata da procedure quasi-statiche, che prevedono un aggiornamento dello stato della rete a intervalli regolari abbastanza lunghi (*aggiornamento sincrono*) o addirittura solo in occasioni in cui si verificano variazioni di un certo interesse (*aggiornamento asincrono*). Nel primo caso un intervallo di aggiornamento tipico può essere dell'ordine di una decina di secondi.

Gli algoritmi di instradamento attualmente in uso o semplicemente proposti sono molto numerosi e sono caratterizzati da diversi livelli di sofisticazione e di efficienza. La ragione di una tale varietà è dovuta sia a ragioni storiche che alle diverse esigenze di specifiche reti. Nel seguito della sezione si esamineranno in modo qualitativo alcune classi di procedure di instradamento adottate nelle reti per dati a pacchetto.

V.4.2 Procedure basate sulla ricerca del percorso più breve

Molti algoritmi di instradamento usati in situazioni reali sono basati sulla nozione di *percorso più breve* tra due nodi (shortest path routing). La rete logica viene dapprima rappresentata mediante un grafo orientato. Ad ogni ramo è assegnata una lunghezza (o peso). La scelta di questi pesi dipende dalla metrica che si intende utilizzare per valutare la distanza tra due generici nodi, definita come somma dei pesi dei rami che occorre percorrere per passare da un nodo all'altro. L'opzione più semplice consiste nel definire pari all'unità il peso di ogni ramo; in tal modo il percorso più breve è semplicemente il percorso che attraversa il numero minimo di rami (percorso a numero di salti minimo). Il peso di un ramo può essere però determinato anche con metriche più complesse quali, ad esempio, la distanza fisica tra le sue estremità, la sua capacità trasmissiva, il ritardo di trasferimento di un pacchetto-tipo attraverso il nodo a monte e il ramo considerato, la lunghezza della coda a monte, e così via.

Si vede quindi che, in funzione della metrica utilizzata, un simile algoritmo può essere sia statico che dinamico che adattativo. Inoltre questa procedura è stata realizzata con soluzioni a controllo sia centralizzato che distribuito. Una volta definita la metrica esistono diversi algoritmi che consentono di identificare il percorso tra due nodi tale che la distanza tra questi (misurata secondo la metrica prescelta) sia minima. A titolo di chiarificazione illustriamo un algoritmo di questo tipo.

L'algoritmo di instradamento del percorso più breve è sviluppabile in una molteplicità di passi, con complessità crescente all'aumentare del numero di nodi della rete. Per chiarirne lo sviluppo facciamo riferimento alla Fig. V.4.3, dove il punto di partenza dell'algoritmo è definito in (a). In questo grafo, ai vari rami componenti sono stati attribuiti i relativi pesi. L'applicazione dell'algoritmo è riferita a una coppia di nodi di origine e di destinazione, che, nel caso di Fig. V.4.3, si identificano con *A* e *D*. I passi di sviluppo dell'algoritmo, che nel caso specifico di Fig. V.4.3, sono cinque, sono illustrati ordinatamente in (b), (c), (d), (e) e (f).

Si parte dal nodo di origine (*A* nel caso specifico) e si individua, tra i nodi adiacenti, quello a distanza minima. Questo è assunto come *nodo di lavoro*. Il risultato del primo passo (Fig. V.4.3) individua il nodo *B* come primo nodo di lavoro, in quanto la sua distanza dal nodo di origine è inferiore a quella del nodo *G*. I nodi *B* e *G* sono etichettati (tra parentesi) con la loro distanza dal nodo di origine e, accanto a tale distanza, si indica anche il nodo *A* da cui si è partiti per valutarne la distanza. Dato che, in questo primo passo, non è stato esplorato alcun percorso, al di fuori di quelli corrispondenti ai rami *AB* e *AG*, gli altri nodi del grafo sono etichettati con distanza infinita.

Man mano che l'algoritmo procede e i percorsi possibili sono esplorati, le etichette di ogni nodo sono variate in relazione alla distanza di questo dal nodo di origine lungo il percorso più breve identificato nel passo precedente. L'etichetta di uno specifico nodo *X*, sempre comprendente la distanza di *X* dal nodo di origine e l'indicazione del nodo attraverso cui può essere raggiunto *X* con quella distanza, può quindi essere *di tentativo* o *permanente*. Inizialmente tutte le etichette sono di tentativo. Quando si accerta che l'etichetta di un nodo *X* è relativa al percorso più breve possibile per raggiungere *X* partendo dal nodo di origine, essa è resa permanente e quindi non più modificabile nei passi successivi dell'algoritmo.

Ogni passo è caratterizzato dal proprio nodo di lavoro (indicato con una freccia in Fig. V.4.3), nel senso che vengono presi in considerazione tutti i nodi ad esso adiacenti e per questi vengono modificate, se possibile, le relative etichette. A questo punto si esamina l'intero grafo per cercare il nodo con etichetta di tentativo contenente il minimo valore di distanza dal nodo di origine. Il nodo così individuato assume etichetta permanente e diventa il nodo di lavoro del passo successivo. E’ immediato verificare l’applicazione di queste regole nei contenuti di Fig. V.4.3, da cui emerge che il percorso più breve tra i nodi A e D è quello che attraversa i nodi B,E,F e H.

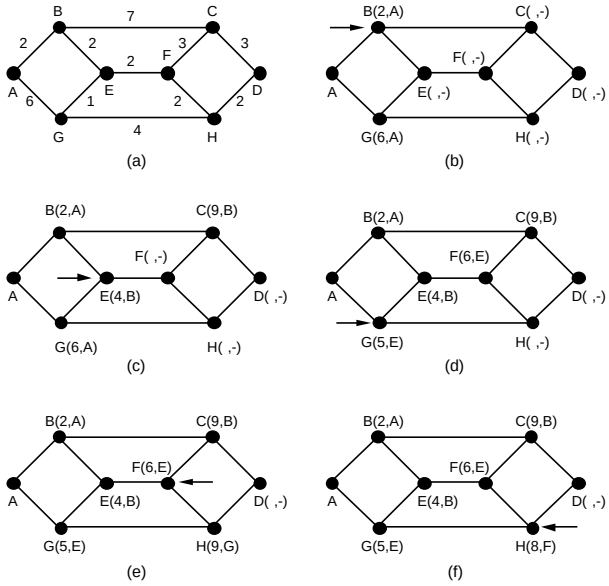


Figura V.4.3 - Applicazione dell'algoritmo di instradamento del percorso più breve (A e D sono i nodi di origine e di destinazione).

Per giustificare come l'algoritmo ora descritto consenta di risolvere il problema della ricerca del percorso più breve si consideri la Fig.V.4.3c, corrispondente al secondo passo dell'algoritmo, in corrispondenza del quale il nodo *E* viene designato come nodo di lavoro. Occorre dimostrare che non può esistere un percorso più breve di *ABE* per andare da *A* ad *E*.

Supponiamo allora che esista un percorso, ad esempio *AXYZE*, che sia più breve di *ABE*. Esistono due possibilità: l'etichetta di Z è stata già resa permanente o non lo è stata. Se lo è stata (prima possibilità) allora *E* è già stato preso in considerazione in un passo seguente a quello in cui l'etichetta di *Z* è stata resa permanente. In queste condizioni il percorso *AXYZE* non è sfuggito alle iterazioni dell'algoritmo.

Se invece *Z* ha ancora una etichetta di tentativo (seconda possibilità) si possono verificare due casi:

- l'etichetta di *Z* è maggiore o uguale a quella di *E*, nel qual caso il percorso *AXYZE* non può essere più breve di *ABE*;
- l'etichetta di *Z* è minore di quella di *E*, nel qual caso *Z* e non *E* deve assumere etichetta permanente; conseguentemente è garantito che *E* sarà preso in considerazione da *Z* e vale quindi la stessa conclusione a cui si è giunti relativamente alla prima possibilità.

V.4.3 Instradamento ottimo

L'instradamento basato sulla ricerca del percorso più breve è abbastanza diffuso, grazie anche alla sua relativa semplicità realizzativa, ma presenta due limitazioni principali:

- la portata che riesce ad offrire è potenzialmente limitata poiché tale algoritmo prevede un solo cammino per ogni coppia origine-destinazione;
- la sua capacità di adattarsi a mutate condizioni di traffico o di disponibilità di risorse di rete è ridotta a causa di fenomeni di oscillazione.

Per chiarire questi due punti ci si avvale delle seguenti considerazioni. Si consideri la rete (avente scopo esclusivamente didattico) mostrata Fig.V.4.4 e composta da 5 nodi e da 5 rami. Si supponga che:

- la rete trasferisca pacchetti di lunghezza costante;

- ogni ramo possa trasportare 10 pacchetti nell'unità di tempo;
- al nodo 1 siano offerti 15 pacchetti destinati al nodo 5;
- alla rete non sia offerto altro traffico.

Nel caso in cui si adotti un algoritmo di instradamento che prevede un solo cammino per ogni coppia origine destinazione, solo 10 pacchetti possono essere trasportati: ad esempio lungo il percorso 1->2->5, che risulta essere il percorso più breve nel caso in cui si scelga pari ad uno il peso di ogni ramo; i restanti 5 pacchetti saranno rifiutati. Se invece per ogni coppia origine-destinazione l'algoritmo di instradamento potesse usare più di una solo percorso, i 15 pacchetti potrebbero essere divise, inviando ad esempio 8 pacchetti lungo il percorso 1->2->5 e i restanti 7 lungo il percorso 1->3->4->5, evitando così di rifiutare del traffico.

Si noti però che in quest'ultima alternativa non è più garantita la consegna in sequenza dei pacchetti. Quindi il prezzo da pagare per l'aumento di efficienza così ottenuto è la necessità di prevedere una operazione di ri-sequenziamento dei pacchetti giunti a destinazione.

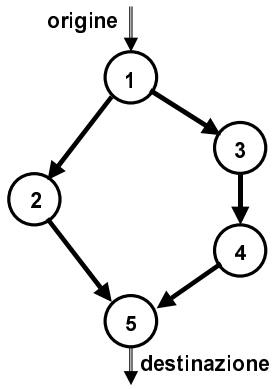


Figura V.4.4 - Limiti dell'algoritmo di instradamento del percorso più breve (nell'esempio ogni ramo può trasportare 10 unità di traffico)

Per quanto riguarda il secondo svantaggio dell'algoritmo di instradamento del percorso più breve e cioè la presenza di eventuali oscillazioni si consideri sempre la rete di Fig. V.3.4. Se la metrica sui cui è basato l'algoritmo utilizza pesi legati allo stato di occupazione dei rami (come è naturalmente il caso, a meno di ridursi a considerare solo il numero e le capacità statiche dei rami attraversati), allora, assumendo che il peso di un ramo sia legato al ritardo di attraversamento, si può verificare la seguente catena di eventi:

- l'algoritmo sceglie un percorso (ad esempio 1->2->5) perché in quel momento risulta non occupato da altro traffico e quindi è caratterizzato da un basso valore del ritardo di attraversamento;
- a causa del traffico appena instradato lungo tale percorso il relativo ritardo di attraversamento aumenta, superando quello del percorso 1->3->4->5;
- di conseguenza l'algoritmo re-instrada il traffico che aveva appena diretto lungo il percorso 1->2->5, spostandolo lungo il percorso 1->3->4->5;
- a causa di questa scelta, il ritardo di attraversamento lungo il percorso 1->3->4->5 supera quello relativo al percorso 1->2->5 e l'algoritmo riporta il traffico sul precedente percorso 1->2->5.

E' immediato vedere che questo modo di operare conduce ad un andamento oscillatorio in cui il traffico è continuamente spostato da un percorso all'altro.

L'instabilità tipica di un tale fenomeno a contro-reazione può essere ridotta: i) limitando la velocità con cui l'algoritmo può cambiare strade; ii) aggiungendo una costante positiva di elevato valore al peso di ogni singolo ramo, così da diminuire l'effetto della variazione del peso dei rami; iii)

rendendo asincroni i messaggi che i nodi si scambiano al fine di comunicarsi le variazioni dei pesi dei rami, in modo tale che la modifica dello stato di un dato percorso non si propaghi in un breve intervallo di tempo in tutta la rete, ma avvenga in modo graduale. Purtroppo l'introduzione di simili accorgimenti limita anche la sensibilità dell'algoritmo alle variazioni delle condizioni di traffico o di disponibilità di risorse di rete.

L'instradamento ottimo, invece, si pone come fine l'ottimizzazione globale su tutta la rete di una data misura prestazionale, ad esempio il ritardo medio di trasferimento. In altre parole non si limita a trovare quale sia il percorso migliore, in un dato momento, per ogni singola relazione di traffico alla volta, ma opera in modo tale da cercare un stato di instradamento che sia ottimo per tutta la rete nel suo insieme. E' facile vedere infatti come cercare la strada migliore per una data relazione di traffico non equivalga a rendere migliori le prestazioni dell'intera rete. In un dato momento ci si può trovare di fronte alla seguente alternativa:;

- scegliere la strada migliore per una certa coppia origine-destinazione anche se tale scelta va a peggiorare le prestazioni viste da altre relazioni di traffico (e queste ultime dovranno quindi adeguarsi alle mutate situazioni);
- scegliere per una certa coppia origine-destinazione una strada che non sia la migliore possibile ma tale che le prestazioni medie globali siano migliori rispetto all'alternativa precedente (possibilmente le migliori in assoluto).

Nel primo caso si sta applicando il principio del percorso più breve, nel secondo il principio dell'instradamento ottimo.

L'instradamento ottimo può dividere il traffico di una data relazione lungo più strade e allocare gradualmente il traffico lungo cammini alternativi; i corrispondenti algoritmi sono ovviamente più sofisticati e complessi.

V.4.4 Instradamento a deflessione

Con il nome di *instradamento a deflessione* (o della patata bollente) si designa un algoritmo d'instradamento molto semplice e robusto, che può essere classificato come adattativo di tipo isolato. E' applicato nei casi in cui le capacità di memorizzazione dei nodi siano molto limitate e si basa sul principio di rilanciare un pacchetto immediatamente, senza immagazzinarlo, eventualmente indirizzandolo verso un'uscita diversa da quella preferita (nel senso che non è quella che lo porterebbe più vicino alla sua destinazione) se quest'ultima è già impegnata.

Supponiamo che più di un pacchetto "preferisca" andare verso una certa destinazione nello stesso momento: solo uno di questi vi sarà inviato mentre gli altri saranno inoltrati verso altre destinazioni "meno gradite" da cui poi saranno rilanciati verso la destinazioni finali se vi saranno risorse disponibili, altrimenti saranno ancora "deflessi" verso ulteriori destinazioni sub-ottimali. Prendendo opportune precauzioni, tra cui quella di evitare che un pacchetto sia inserito in un ciclo chiuso da cui non riesce ad uscire, si può garantire che la totalità dei pacchetti (o una grande maggioranza di essi) arrivi a destinazione. Lo svantaggio di tale schema è che non preserva la sequenza dei pacchetti e che introduce un ritardo di trasferimento che può essere notevolmente variabile.

V.4.5 Instradamento diffusivo

Come accennato in precedenza, durante il funzionamento di un algoritmo di instradamento può essere necessario diffondere alcune informazioni che consentono all'algoritmo stesso di poter operare (ad es. lo stato di occupazione delle risorse di rete). In molte situazioni, tali informazioni devono essere inviate da un nodo di origine a *tutti* gli altri nodi: come esempio di nodo di origine si pensi a

- un centro di controllo di rete, nel caso di algoritmi centralizzati, che deve inviare i dati da esso elaborati a tutti i nodi di commutazione;
- un nodo di commutazione che si trovi in stato di congestione e che deve rendere nota tale sua situazione a tutti gli altri nodi (nel caso di algoritmi distribuiti).

In altri casi può essere necessario inviare in modo diffusivo anche l'informazione di utente: si pensi a servizi di telecomunicazione di tipo appunto diffusivo (come potrebbe essere il servizio televisivo).

Una metodologia molto usata per diffondere a tutti i nodi di rete una certa informazione è nota come "*flooding*" (inondazione) e opera come segue. Il nodo di origine invia l'informazione in oggetto a tutti i suoi "vicini" (cioè ai nodi a cui è direttamente connesso). Ogni vicino la rilancia verso i suoi vicini e così via finché l'informazione in tal modo arriva a tutti i nodi della rete. Al fine di limitare il numero di pacchetti trasmessi si osservano due regole addizionali:

- un nodo non deve rilanciare un pacchetto verso il nodo da cui lo ha ricevuto;
- un nodo deve trasmettere un pacchetto a ogni vicino una sola volta.

Ciò può essere attuato includendo in ogni pacchetto un identificativo del nodo di origine ed un numero di sequenza che è incrementato con ogni pacchetto emesso dal nodo di origine. Memorizzando il numero di sequenza più alto ricevuto per ogni nodo di origine e non rilanciando pacchetti con numero di sequenza più basso di quello memorizzato ogni nodo può evitare di trasmettere lo stesso pacchetto più di una volta.

Questo modo di operare è illustrato nella Fig. V.4.5 che illustra anche come il numero totale di pacchetti trasmessi possa variare tra L e $2L$, dove L è il numero di rami bi-direzionali della rete.

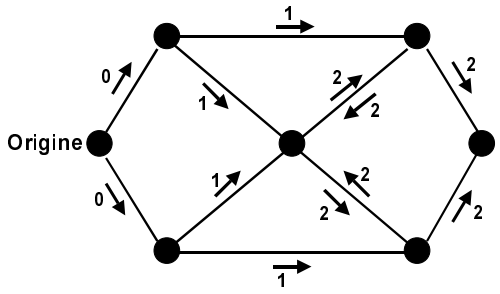


Figura V.4.5 - Esempio di algoritmo diffusivo

Un altro metodo di instradamento diffusivo è basato sulla nozione di "*spanning tree"* (albero pervasivo).

Secondo la usuale nomenclatura usata nei grafi geometrici, uno spanning tree associato a un grafo $\mathcal{G}$ è uno sotto-grafo di $\mathcal{G}$(cioè con rami che sono un sotto-insieme di quelli di $\mathcal{G}$che include il minor numero di rami necessario per fornire completa connettività tra tutti i nodi di $\mathcal{G}$(senza formare percorsi chiusi)

Nell'inoltro a spanning tree ogni nodo della rete logica conserva informazione su quali dei suoi rami uscenti appartengano allo spanning tree che è associato al grafo della rete. I pacchetti che sono instradati in modo diffusivo (*pacchetti diffusivi*) sono distinti da uno speciale indirizzo di destinazione (formato ad es. da tutti "1"). Quando un nodo riceve un pacchetto di tipo diffusivo, ne inoltra una copia su tutti i suoi rami uscenti che appartengono allo spanning tree, con l'eccezione di quello che ha la stessa estremità del ramo entrante. Il nodo fornisce anche una copia del pacchetto a tutti i calcolatori che ad esso fanno capo.

A titolo di esempio, si consideri la Fig. V.4.6 ove è mostrato lo spanning tree associato al grafo di Fig. V.4.5. Grazie all'algoritmo ora descritto è possibile effettuare una comunicazione diffusiva che

richiede un numero totale di pacchetti trasmessi pari a *N-1* , se *N* è il numero di nodi del grafo corrispondente alla rete. Un tale aumento di efficienza rispetto all'algoritmo di inondazione va pagato con la necessità di memorizzare e di aggiornare in ogni nodo interessato lo spanning tree della rete, anche in seguito a variazioni topologiche.

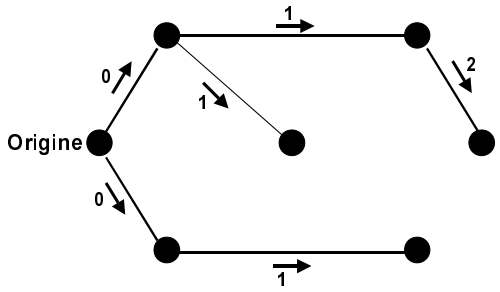


Figura V.4.6 - Esempio di spanning tree

V.5 Instradamento e indirizzamento in una inter-rete

Come già si è visto nel par. II.8, in una inter-rete è necessario fare uso di *sistemi di interconnessione* (SINT), che fungano da interfaccia tra una sotto-rete e l'altra, convertendo i protocolli di una nei corrispondenti protocolli dell'altra, e che consentano ad una unità informativa proveniente da una data sotto-rete di giungere ad un altra sotto-rete, dove si trova il sistema terminale verso cui quell'unità informativa è destinata.

Tra le funzioni che tali dispositivi devono svolgere, vi è, in generale, la funzione di instradamento. Supponendo che l'inter-rete in questione adotti nello strato di rete una modalità di trasferimento senza connessione, un sistema di interconnessione deve svolgere, fra gli altri, i seguenti compiti:

- per ogni unità informativa che prende in consegna, deve determinare dove si trova, nell'ambito della inter-rete, la destinazione verso la quale quella unità informativa è diretta (compito di natura decisionale);
- per ciascuna unità informativa deve individuare una strada che porti quest'ultima verso la destinazione voluta (compito di natura decisionale);
- deve inoltrare fisicamente ogni unità informativa verso la strada prescelta (compito di natura attuativa).

Nel caso di modalità di trasferimento con connessione i compiti di natura decisionale sono svolti solo durante la fase di instaurazione mentre quello di natura attuativa è svolto per ogni unità informativa trattata.

V.5.1 Instradamenti gerarchico e non gerarchico

Dal punto di vista della funzione di instradamento, una inter-rete può essere vista in due modi alternativi:

- in una visione ad *un* livello, in cui i dispositivi di interconnessione hanno il ruolo di nodi addizionali rispetto a quelli delle sotto-reti; ogni sistema di interconnessione è connesso a due o più sotto-reti e quindi a due o più nodi appartenenti alle sotto-reti che interconnette; l'intera inter-rete è vista come un insieme di nodi di numerosità pari alla somma di tutti i nodi delle sotto-reti componenti e di tutti i sistemi di interconnessione;

- in una visione a *due* livelli; al primo di questi livelli ogni sotto-rete è vista come entità a se stante e costituita da un certo numero di nodi; tra tali nodi ve ne sono alcuni di natura particolare (i sistemi di interconnessione) che la mettono in comunicazione con il mondo esterno ad essa.

Al secondo livello, ogni sotto-rete è vista come un macro-nodo e la struttura interna di ogni sotto-rete non è visibile dall'esterno. L'intera inter-rete è quindi vista, a questo secondo livello, come costituita da un insieme di nodi: alcuni di questi sono macro-nodi, rappresentanti le sotto-reti componenti, mentre i rimanenti sono i sistemi di interconnessione che interconnettono i macro-nodi. A questo secondo livello la numerosità di "nodi equivalenti" è quindi ben minore di quella nella visione ad un livello, poiché la struttura interna delle sotto-reti non è visibile.

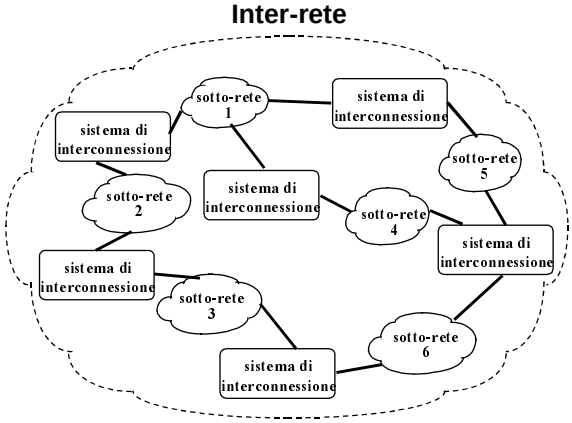


Figura V.4.7 - Esempio di inter-rete

L'instradamento basato sul primo di questi due modelli è chiamato *non gerarchico* e può appartenere a una qualunque delle tipologie discusse nei paragrafi precedenti; ad esempio può essere del tipo a ricerca del percorso più breve. In tal caso l'algoritmo di instradamento dovrà cercare il percorso più breve tra *tutti* i nodi che costituiscono l'inter-rete.

L'instradamento basato sul secondo dei modelli di cui sopra è detto *gerarchico*. Nel caso di instradamento gerarchico, si ha una procedura a due livelli. Al livello più basso vi è un algoritmo di instradamento, separato per ogni sotto-rete, il quale:

- gestisce fino a destinazione il traffico locale (interno alla sotto-rete);
- instrada il traffico diretto ad altre sotto-reti verso i dispositivi di interconnessione, limitandosi a farlo arrivare ad uno di questi ultimi e non ponendosi il problema di quale strada fargli percorrere fino al sistema terminale di destinazione finale.

Al livello più alto vi è un algoritmo di instradamento che determina la sequenza di sotto-reti e dei relativi dispositivi di interconnessione che ogni unità informativa deve attraversare se vuole, partendo da un sistema di interconnessione di una sotto-rete, arrivare ad un sistema di interconnessione appartenente ad una diversa sotto-rete. In altre parole l'algoritmo di basso livello gestisce il traffico "locale" e l'algoritmo di alto livello gestisce il traffico di "lunga distanza", guidandolo attraverso diverse sotto-reti per consegnarlo infine all'algoritmo di basso livello della sotto-rete di destinazione, il quale lo farà pervenire al sistema terminale di destinazione.

L'instradamento di tipo gerarchico è tipicamente preferibile in reti di grandi dimensioni (ad esempio è quello usato in Internet e nella rete telefonica mondiale). Per meglio chiarire quanto appena esposto si ricorre ad un esempio.

Si consideri una nazione con diverse città, ognuna delle quali ha una sua rete cittadina connessa alla rete nazionale tramite dispositivi di interconnessione. Supponendo di usare l'algoritmo del percorso più breve, se si usa un instradamento non gerarchico, non vi è distinzione tra le reti cittadine: ogni nodo deve calcolare il percorso più breve verso tutti i nodi di tutta la nazione; ogni nodo deve conoscere la topologia di tutta la rete e la dimensione delle tabelle di instradamento contenute in ogni nodo (ovvero una serie di informazioni, elaborate durante la funzione decisionale, che specificano per ogni destinazione quale strada far seguire all'informazione di utente) è pari al numero di tutti i nodi della rete nazionale.

Se invece si usa un approccio gerarchico, esiste un algoritmo di instradamento locale di basso livello per ogni città, le cui tabelle di instradamento devono contenere informazioni solo per i nodi appartenenti alla stessa città. L'algoritmo di alto livello si occupa dell'instradamento quando una comunicazione non è limitata all'interno di una sola città e opera nei dispositivi di interconnessione. Ogni dispositivo di interconnessione deve conoscere la topologia della nazione come insieme di città, senza entrare nel merito di come è costituita ogni singola città, e ha una tabella di instradamento la cui dimensione è pari al numero delle città.

Per vedere come l'instradamento gerarchico complessivo lavori, si consideri una unità informativa generata da un sistema appartenente a una data città; l'algoritmo di instradamento locale, di basso livello, prende in consegna l'unità informativa e determina se essa è destinata alla stessa città da cui è stata generata o meno. Nel primo caso è capace di portarla a destinazione, altrimenti la consegna ad un dispositivo di interconnessione; qui è presa in consegna dall'algoritmo di alto livello che ne determina la strada attraverso diversi dispositivi di interconnessione (e quindi attraverso città) fino ad arrivare alla città di destinazione; qui l'unità informativa è di nuovo presa in consegna dall'algoritmo di basso livello che la consegnerà al sistema terminale di destinazione.

E' facile vedere come nel caso di instradamento gerarchico la dimensione delle tabelle di instradamento ed i calcoli da effettuare per determinare la strada più opportuna sono considerevolmente minori che nel caso di un algoritmo non gerarchico. D'altra parte un instradamento gerarchico risulta tipicamente meno ottimale di uno non gerarchico. Bisogna quindi scegliere un opportuno compromesso tra questi due approcci a seconda delle specifiche esigenze.

Si noti infine che l'instradamento gerarchico può essere organizzato in più di due livelli (ad esempio una rete internazionale costituita da più reti nazionali ognuna delle quali comprende diverse città).

V.5.2 Il problema dell'indirizzamento in una inter-rete

È usuale che varie sotto-reti tra quelle interconnesse operino secondo un proprio *Piano di Numerazione* (PN), con il quale un apparecchio terminale H1 può indirizzare in modo univoco un qualunque apparecchio terminale H2 purchè appartenente alla stessa sotto-rete. Il problema dell'indirizzamento sorge quando H1 e H2 appartengono a sotto-reti diverse come in Fig. IV.8.2 e queste sotto-reti operano secondo PN diversi.

Per questo problema partiamo da una prima ipotesi di soluzione, che è descritta con riferimento al caso semplice di Fig. II.8.1. Siano PN1 e PN2 i piani di numerazione delle sotto-reti 1 e 2 rispettivamente. SINT, che ha il compito di interconnettere queste sotto-reti, presenta (nei loro confronti) due interfacce fisiche, ognuna delle quali è indirizzabile nella modalità prevista dal PN della sotto-rete verso cui SINT si interfaccia. Invece i sistemi H1 e H2 sono indirizzabili in accordo a PN1 e PN2, rispettivamente.

Allora, se l'utente facente capo ad H1 conosce, oltre all'indirizzo di H2 secondo PN2, anche l'indirizzo di SINT secondo PN1, il trasferimento del segmento informativo *IS* da H1 a H2 può avvenire nel seguente modo:

- l'utente in H1 invia *IS* a SINT, come descritto dal modello di Fig. II.8.2; per questo scopo utilizza l'indirizzo di SINT secondo PN1; nell'invio comunica a SINT anche l'indirizzo di H2 secondo PN2;
- SINT, una volta ricevuto *IS* da H1, lo rilancia verso H2 utilizzando l'indirizzo di H2 secondo PN2, che gli è stato comunicato da H1.

Questa ipotesi di soluzione, se è praticabile in un ambiente semplice come quello descritto in Fig. II.8.1, non lo è più se si fa riferimento ad un inter-rete costituita da varie sotto-reti e da una molteplicità di SINT. Infatti per ogni coppia H1-H2 di origine e di destinazione, comunque collocata nell'ambito dell'inter-rete, è necessario che l'origine H1 conosca

- l'indirizzo della destinazione H2 con il vincolo che questo indirizzo sia secondo il *PN* della sotto-rete a cui fa capo H2; questo è l'*indirizzo locale* di H2;
- la sequenza *S* dei SINT che portano dalla sotto-rete di origine a quella di destinazione qualora il percorso da H1 a H2 debba attraversare due o più SINT, ognuno dei quali mette in corrispondenza una sotto-rete a monte con una a valle;
- l'indirizzo con cui ogni SINT di *S* è indirizzabile localmente nella sotto-rete a monte per stabilire una corrispondenza con la sotto-rete a valle.

In altre parole H1, oltre a conoscere tutti gli indirizzi locali dei possibili destinatari H2 e dei SINT da attraversare, deve essere anche in grado di svolgere la funzione di instradamento: lo scopo è individuare la sequenza *S* di SINT da attraversare per raggiungere H2. Se queste condizioni sono soddisfatte, H1 trasferisce *IS* a H2 attraverso i vari SINT di *S*. Congiuntamente H1 invia al primo SINT di *S* anche gli indirizzi locali di tutti i SINT a valle. Inoltre ogni SINT di *S* rilancia verso il SINT che gli è a valle, assieme al segmento *IS*, anche l'indirizzo locale dei SINT di *S* ancora da percorrere.

È evidente che questa ipotesi di soluzione presenta inconvenienti non eludibili. In primo luogo la individuazione della sequenza *S* di SINT per ogni coppia H1-H2 richiede che ogni apparecchio terminale della inter-rete sia in grado di svolgere una funzione di instradamento per ogni stato della rete, e quindi con la conoscenza completa delle risorse di trasferimento di volta in volta disponibili. Inoltre, al crescere della dimensione della inter-rete, l'ipotesi di soluzione sopra prospettata, anche se valida in linea di principio, diventa rapidamente troppo complessa e quindi impraticabile, da un punto di vista gestionale: si pensi ad esempio alle difficoltà connesse all'aggiornamento delle memorie degli apparecchi terminali ogni qualvolta viene inserito un nuovo sistema terminale o di interconnessione nella inter-rete.

Una seconda ipotesi di soluzione cerca di superare queste difficoltà aumentando le funzionalità dei SINT e rivedendo la modalità di indirizzamento nell'ambito della inter-rete.

Circa l'arricchimento delle funzionalità dei SINT, si attribuiscono a questi sistemi *capacità di instradamento*: l'obiettivo è determinare, in relazione all'indirizzo di destinazione, la sequenza di SINT che porti dall'origine alla destinazione.

Relativamente alle modalità di indirizzamento, si arricchisce il riferimento ai soli indirizzi locali : si aggiunge infatti, per ogni coppia origine-destinazione, un *indirizzo globale* (cfr. § II.2.4) che ha significatività e univocità nella intera inter-rete. Il piano di numerazione globale riguarda anche i SINT ognuno dei quali è caratterizzato da tanti indirizzi quante sono le sue interfacce fisiche.

Con queste scelte sistemistiche e architetturali, un H1 di origine, quando vuole indirizzare un H2 di destinazione, individua quest'ultimo con il suo indirizzo globale e consegna il suo segmento informativo a un SINT con interfaccia fisica verso la sua sotto-rete. Questo SINT, individuato con un indirizzo globale o con quello locale, sceglie la sotto-rete a valle su cui inoltrare il segmento *IS* lungo il percorso verso H2. Una volta raggiunta la sotto-rete di destinazione, la consegna di *IS* a H2 può avvenire convertendo il numero globale in quello locale.

Da un punto di vista architetturale, quanto ora detto si attua introducendo, al di sopra dello strato di rango 3, uno *strato aggiuntivo* SA (con un relativo protocollo PA), comune a tutti i sistemi coinvolti nella sotto-rete (Fig. V.5.2). Lo scopo di questo strato è svolgere funzioni di *instradamento*. Compito del protocollo PA è anche quello di gestire il *piano di numerazione globale* in aggiunta a quello di significatività locale. In particolare sarà necessario che PA o altri protocolli di supporto riescano a determinare quale indirizzo locale corrisponde ad un dato indirizzo globale e viceversa.

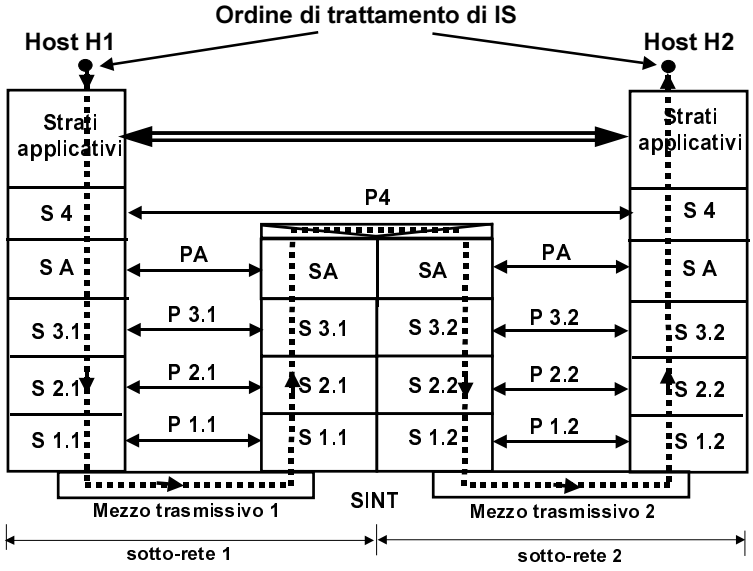


Figura V.5.2 – Architettura di sistema di interconnessione con adozione di uno strato aggiuntivo

In conclusione si osserva che questa soluzione, con il doppio indirizzamento e con l'instradamento affidato ai sistemi di interconnessione, elimina completamente gli inconvenienti della prima ipotesi di soluzione; la sua praticabilità è dimostrata dall'uso che se ne fa in Internet fin dai suoi primi sviluppi.

V.5.3 Modo di servizio di rete in una inter-rete

Un'ultima considerazione riguarda la scelta del modo di servizio di rete da adottare per lo strato SA: modo con connessione o senza?

Ambedue le scelte presentano pro e contro. Si riprenda in esame la Fig. IV.8.3. Se SA è nel modo *con connessione*, le funzioni che i SINT devono svolgere sono più complesse. Infatti quando H1 vuole instaurare una connessione nello strato SA, deve farne richiesta prima a SINT e non direttamente a H2. SINT deve elaborare tale richiesta, eventualmente accettarla, per il tratto che va da H1 a SINT, e chiedere a H2 di accettare una connessione da SINT a H2. La connessione da H1 a H2 risulta quindi dalla concatenazione di diverse connessioni componenti. La connessione da H1 a SINT deve soddisfare i requisiti prestazionali posti da H1 in termini, ad esempio, di portata media e di ritardo di trasferimento massimo. Dovrà quindi essere cura di SINT far si che anche la tratta da SINT a H2 soddisfi gli stessi requisiti, affinché sia poi l'intera connessione da estremo a estremo a godere delle prestazioni volute da H1. Si noti che SINT potrebbe non riuscire a soddisfare i requisiti posti da H1 (anche solo nel tratto da SINT a H2) e quindi potrebbe essere forzato a rifiutare la richiesta di H1.

Se, viceversa, SA opera in modalità *senza connessione*, il compito di SINT è più semplice. SINT tratta le unità informative in arrivo da H1, una per volta, in modo indipendente l'una dall'altra, e non deve porsi il problema di soddisfare determinati requisiti relativi ad una connessione, giacché

quest'ultima non esiste. Se la sotto-rete 2 diviene indisponibile a causa di un guasto, SINT può cercare una strada alternativa per raggiungere H2, senza preoccuparsi di re-instradare le connessioni già instaurate, cosa che avrebbe dovuto fare nell'ipotesi precedente.

L'ipotesi senza connessione è di più semplice realizzazione, ma ha due svantaggi significativi:

- *non* è possibile *garantire* che lo scambio informativo tra H1 e H2 avvenga in accordo a determinate caratteristiche prestazionali (portata, ritardo, perdita di informazione);
- la funzione decisionale relativa all'instradamento deve essere svolta per ogni segmento informativo che attraversa SINT; al contrario, nell'alternativa con connessione, la decisione su quale strada far percorrere all'informazione verrebbe presa solo una volta, durante la fase di instaurazione; poiché la funzione di instradamento è abbastanza complessa ed è tipicamente realizzata in software, invece che in hardware, la modalità senza connessione impone ai SINT limitazioni in termini di portata.

In altri termini, adottando la modalità senza connessione, un SINT deve svolgere compiti più semplici ma deve eseguirli un numero più elevato di volte, rispetto all'alternativa con connessione.

Uno strato che realizza le funzioni sinora genericamente attribuite a SA è l'*Internet Protocol* (IP), adottato in Internet. La modalità senza connessione di IP ne ha determinato la semplicità implementativa ed ha significativamente contribuito allo sviluppo di Internet; è però la causa degli inconvenienti attuali di questa rete: limitazioni in termini di portata ed assenza di garanzie sulle prestazioni percepite dagli utenti.

V.5.4 Modalità di instradamento in una inter-rete

Per concludere la descrizione delle modalità generali di funzionamento di una inter-rete, si fa notare che ogni SINT deve disporre di opportune informazioni riguardanti la topologia dell'inter-rete. Tali informazioni sono quelle necessarie a svolgere la *funzione di instradamento* e cioè a determinare quale strada far seguire ad un dato segmento informativo *IS* affinché questo giunga alla destinazione voluta. A titolo di esempio, si consideri la inter-rete rappresentata in Fig. V.5.3. Con riferimento agli *IS* avente origine nella sotto-rete 1, il SINT 1 deve trasferire alla sotto-rete 2 sia gli *IS* aventi destinazione nella sotto-rete 2 stessa, sia quelli che hanno la sotto-rete 3 come loro destinazione. Il SINT 1 deve quindi essere "a conoscenza" anche dell'esistenza della sotto-rete 3, a cui non è direttamente connesso.

Figura V.5.3 – Interconnessione di tre sotto-reti

La complessità dei SINT può essere limitata adottando la seguente modalità di instradamento che è applicata in Internet: i SINT instradano gli *IS* solo verso la sotto-rete di destinazione e non verso il singolo apparecchio terminale a questa appartenente; una volta che un *IS* arriva alla sotto-rete di destinazione sono i protocolli propri di questa sotto-rete ad instradarlo verso lo specifico apparecchio terminale di destinazione. L'algoritmo di instradamento operante nei SINT determina solo la sequenza dei SINT da attraversare e non quella di *tutti* i sistemi di *ogni* sotto-rete.

In tal modo la quantità di informazioni, relative alla funzione di instradamento, che SINT deve elaborare per far arrivare un generico messaggio a destinazione è proporzionale al numero di *sotto-reti* che costituiscono l'inter-rete e non al numero degli *apparecchi terminali* ad essa connessi. Tale modalità di funzionamento corrisponde dunque ad un algoritmo di instradamento di tipo gerarchico .

V.6 Controllo del traffico e della congestione

Una fenomenologia tipica che si riscontra in una rete di telecomunicazione è legata agli effetti del traffico sulle prestazioni della rete e delle sue parti componenti. Ciò è legato al largo impiego di risorse condivise e ai conseguenti fenomeni di congestione, che si possono manifestare sia nelle strategie di pre-assegnazione individuale o collettiva (*congestione di pre-assegnazione*), sia nel corso dell'evoluzione di attività di utilizzazione, che agiscono nell'ambito di strategie con assegnazione a domanda o con pre-assegnazione collettiva (*congestione di utilizzazione*).

Nei fenomeni di congestione di pre-assegnazione o di utilizzazione si può osservare che, al di sopra di un carico-limite (*soglia di sovraccarico*), le prestazioni dell'insieme peggiorano bruscamente. Ad esempio, se le contese di pre-assegnazione sono risolte con modalità a perdita, all'aumentare della domanda di pre-assegnazione al di sopra di questa soglia, si verifica un brusco aumento della probabilità di rifiuto. Se si fa invece riferimento a contese di utilizzazione gestite con modalità ad attesa, al crescere del carico offerto, il sistema risponde con un rapido incremento del valore medio dei tempi di attesa e con una diminuzione, altrettanto brusca, della sua portata media.

Per evitare condizioni di questo tipo, che possono condurre rapidamente al completo collasso di intere sezioni di rete, occorre prevedere opportune forme di controllo del traffico, che consentano di filtrare le domande in modo che:

- l'insieme operi con una intensità di traffico offerto che sia sempre inferiore alla soglia di sovraccarico;
- le risorse componenti siano utilizzate in modo efficiente.

Notiamo che questa seconda condizione è in parziale contrasto con la prima, dato che impone di lavorare in prossimità della soglia.

Come è facile intuire, il problema ora delineato è normalmente di soluzione molto complessa. Ci limiteremo quindi a illustrare qui, a titolo d'esempio, solo alcune linee-guida oggi seguite nell'ambito delle reti per dati a pacchetto.

In questo caso sono previste due forme di controllo di traffico, e cioè quelle di tipo *preventivo* e quelle di tipo *reattivo*.

V.6.1 Controllo preventivo

Nel controllo di tipo preventivo, che è applicabile solo a infrastrutture con servizio di rete orientato alla connessione, è previsto il meccanismo di *controllo di accettazione di chiamata*. In base a questo una chiamata virtuale viene accettata solo quando sono disponibili sufficienti risorse da pre-assegnare virtualmente in modo che siano soddisfatte due condizioni:

- la nuova chiamata possa usufruire di una adeguata qualità di servizio;
- le chiamate già instaurate e tuttora in corso continuino ad evolvere con la qualità di servizio concordata al momento della loro accettazione.

La gestione delle situazioni di contesa è effettuata normalmente con modalità a perdita. L'accettazione di chiamata è operata, ad esempio nei nodi a pacchetto (cfr. § IV.7.3), con assegnazione *a domanda media* ovvero *a domanda di picco*.

V.6.2 Controllo reattivo

Nei controlli di tipo reattivo, quali sono applicati almeno nelle reti per dati a pacchetto, sono utilizzati, durante la fase di trasferimento dell'informazione, meccanismi *a retro-pressione,* tra i quali si possono citare le tecniche di *controllo di flusso*. Compito di questo controllo è assicurare che, nel caso in cui il ritmo di arrivo delle UI sia superiore alla capacità di elaborazione della entità ricevente, non si

verifichino perdite di informazione a causa della saturazione delle sue risorse di memorizzazione. Circa la sua attuazione, un controllo di flusso che è normalmente gestito con modalità ad attesa, viene effettuato attraverso il rilascio, a cura della parte ricevente e nei confronti di quella emittente, di opportune autorizzazioni relative al numero massimo di UI che possono essere emesse e/o ricevute. Ciò consente di ridurre o di eliminare gli effetti delle singole cause di congestione di utilizzazione, all'atto in cui si presentano condizioni di carico elevato in certe parti della rete.

Sono possibili vari schemi di controllo di flusso, la maggior parte dei quali utilizza le regole protocollari che sono definite per consentire il recupero di condizioni i trasferimento anomale.

Un esempio di controllo reattivo è offerto dalla tecnica di *controllo di flusso a finestra*. Tale tecnica si realizza, per ciascun verso di trasferimento, attraverso l'adozione di una *finestra in emissione* posta nella estremità emittente e di una *finestra in ricezione* collocata nell'estremità ricevente .

Con riferimento, ad esempio, al flusso di UI scambiate fra due entità di strato *A* e *B* e al verso di trasferimento da *A* a *B*, la finestra in emissione specifica qual è il numero massimo di UI che *A* può emettere in funzione della disponibilità di risorse in *B*. La finestra in ricezione (collocata in *B*) precisa invece se il numero massimo di UI ricevute da *B* è effettivamente quello autorizzato ad essere emesso da parte di *A*. Con il meccanismo delle finestre, la disponibilità al trasferimento dall'entità *A* a quella *B* viene aggiornata esclusivamente da *B* e cioè da chi agisce come ricevente: ciò avviene agendo direttamente sulla finestra in ricezione (se presente) e inviando ad *A* opportune informazioni per il controllo della finestra in emissione (se presenti).

E' tuttavia da notare che questa tecnica, come del resto qualunque altra a retro-pressione, può essere impiegata quando il ritmo binario di trasferimento non è troppo elevato in relazione al valore che assume il ritardo di propagazione sulla connessione fisica tra *A* e *B*. Infatti un eventuale ritardo nel rivelare la condizione di congestione e nel prendere le opportune azioni correttive può rendere inefficace l'applicazione di meccanismi del tipo ora descritto.

V.7 Il trattamento della segnalazione

In quanto precede ci siamo interessati solo dei modi di trasferimento relativi all'informazione d'utente. Verrà ora considerata anche l'informazione di segnalazione. Questa, come si è già detto in § I.3.1, ha lo scopo di controllare una comunicazione. È quindi impiegata, oltre che per instaurare e per abbattere una connessione fisica o virtuale al servizio della comunicazione, anche per modificarne le caratteristiche quando la comunicazione è stata già inizializzata.

Nel seguito del paragrafo viene data, dapprima (§V.7.1), una classificazione dei possibili modi di segnalazione adottati nelle reti per telefonia e per dati e di cui si prevede l'impiego per la fornitura di servizi integrati a banda stretta o larga. Successivamente si parlerà brevemente della segnalazione a canale comune (§V.7.2).

V.7.1 Classificazione dei modi di segnalazione

È opportuno distinguere preliminarmente due tipi di informazione di segnalazione, e cioè quella che è scambiata, nella rete di accesso, tra apparato terminale e autocommutatore locale (*segnalazione di utente*) e quella che viene trasferita tra i nodi della rete di trasporto (*segnalazione inter-nodale*).

Consideriamo poi separatamente le modalità di trattamento di questa informazione nelle reti a circuito, in quelle a pacchetto con servizio a chiamata virtuale e in quelle (*reti ATM*) operanti con modo di trasferimento asincrono (cfr. § IV.5.3).

Nelle reti a circuito per telefonia, le segnalazioni di utente e inter-nodale sono state in passato, e sono tuttora, trattate con le modalità più varie, che si differenziano per una molteplicità di aspetti, quali

la codifica di queste informazioni e il protocollo di comunicazione per l'interazione degli elementi di rete preposti al controllo della chiamata.

Un ulteriore aspetto distintivo riguarda l'associazione tra il modo di trasferimento dell'informazione di segnalazione preposta al trattamento di una data chiamata e quello dell'informazione d'utente scambiata nell'ambito della stessa chiamata. A tale riguardo esistono due alternative principali, e cioè la *segnalazione associata al canale* (SAC) e la *segnalazione a canale comune* (SCC).

Nell'alternativa SAC, l'informazione di segnalazione preposta al controllo di una chiamata viene scambiata su un canale fisico distinto rispetto ad altri canali che trasferiscono informazioni dello stesso tipo, ma preposte al controllo di chiamate differenti. In queste condizioni si stabilisce una corrispondenza biunivoca tra due canali: quello su cui è trasferita l'informazione di utente relativa alla chiamata considerata (*canale controllato*) e quello su cui viaggia l'informazione di segnalazione che tratta quella chiamata (*canale controllante*). Per ambedue i tipi di informazione si attua lo stesso modo di trasferimento a circuito.

I canali controllante e controllato (Fig. V.7.1) possono coincidere (*soluzione a canale unico*) o essere distinti (*soluzione a canali separati*); nel primo caso si parla di *segnalazione in banda*, nel secondo di *segnalazione fuori banda*.

Nella soluzione SAC a canale unico, le informazioni di utente e di segnalazione corrispondenti alla stessa chiamata condividono lo stesso canale con una utilizzazione in intervalli di tempo disgiunti. Le attuazioni di questa soluzione dipendono dal tipo di rete a circuito che si considera e dalle conseguenti modalità di codifica dell'informazione di segnalazione. Essa ha trovato impiego in ambedue le sezioni di rete, come nel caso delle reti telefoniche analogiche, ove la segnalazione è codificata da impulsi o da toni. Ma si incontra anche nella rete di accesso, come nelle reti a circuito per dati in tecnica numerica, ove la segnalazione d'utente è presentata in un formato a caratteri di 8 cifre binarie, secondo l'*Alfabeto Internazionale no.5* (IA5) e come è specificato nella Racc. X.21.

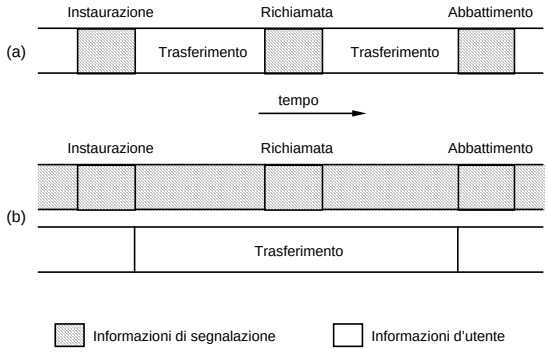


Figura V.7.1- Modi di trasferimento dell'informazione di utente e di quella di segnalazione su un canale unico (a) e su canali separati (b).

La soluzione SAC a canali separati è utilizzata, nelle reti telefoniche, sia per la segnalazione d'utente che per quella inter-nodale. I modi di trasferimento sono a circuito per ambedue i canali. Un esempio tipico è offerto dalla modalità di trasferimento dell'informazione di segnalazione in uno schema di multiplazione statica PCM, attuata secondo la versione primaria europea. In questo caso l'informazione d'utente è trasportata da un canale scelto tra i 30 disponibili con capacità di 64 kbit/s, mentre l'informazione di segnalazione usufruisce di un canale, con capacità di 2 kbit/s, ottenuta sotto-multiplando l'IT no.16, con modalità a multitrama.

Nell'alternativa SCC, come nella soluzione SAC a canali separati, il trasferimento dell'informazione di segnalazione avviene su un canale distinto rispetto a quelli che trasportano

l'informazione di utente, ma con due differenze sostanziali: il canale della segnalazione agisce come controllante per una pluralità di chiamate e il modo di trasferimento che viene impiegato è del tipo orientato al pacchetto, con servizio di rete senza connessione.

L'alternativa SCC è oggi la soluzione più avanzata per il trasferimento della segnalazione inter-nodale in una rete a circuito in tecnica numerica. Ma viene impiegata anche nel caso della N-ISDN, sia nella rete di trasporto, che in quella di accesso.

Per concludere, rimangono da esaminare i casi delle reti a pacchetto con servizio a chiamata virtuale e di quelle ATM. Nel primo caso, se i protocolli di accesso e inter-nodali sono del tipo descritto dalle Racc.X.25 e X.75, l'informazione di segnalazione (sia di utente, che inter-nodale) è strutturata in UI specializzate (*pacchetti di segnalazione*), che condividono, in uno schema di multiplazione dinamica, lo stesso canale logico utilizzato dalle UI contenenti l'informazione di utente (*pacchetti di dati*). Invece, nel caso delle reti ATM, le informazioni di utente e di segnalazione sono trasferite su canali logici separati. Anche in questi due casi si parla di segnalazione in banda e fuori banda, rispettivamente.

V.7.2 La segnalazione a canale comune

Il trasferimento della segnalazione a canale comune nella sezione interna di una rete a circuito utilizza come supporto una rete apposita, che è sovrapposta alla precedente e che è chiamata *rete di segnalazione a canale comune* (rete SCC). Il trasferimento su questa rete avviene secondo le regole di protocolli, che, nel loro insieme, costituiscono il *sistema di segnalazione no.7* (SS no.7), che è normalizzato dal ITU-T nelle Raccomandazioni della serie Q.700.

Nella rete SCC si distinguono i punti terminali e i nodi di commutazione. Un punto terminale, chiamato *punto di segnalazione* (SP - Signalling Point), rappresenta la sorgente e il collettore dell'informazione di segnalazione ed è in corrispondenza con l'unità di comando di un autocommutatore (AUT) della rete a circuito; nel seguito, per indicare questa corrispondenza, si dirà che SP e AUT costituiscono coppia.

Un nodo di commutazione della rete SCC è denominato *punto di trasferimento della segnalazione* (STP - Signalling Transfer Point), mentre i rami della rete SCC sono i *collegamenti di segnalazione*. Con riferimento a una specifica coppia (SP, AUT), un collegamento di segnalazione uscente dall'SP di questa coppia è al servizio di uno o più tra i fasci di linee di giunzione (*vie di giunzione*), che escono dall'AUT della stessa coppia verso altri autocommutatori della rete a circuito.

L'instaurazione di un circuito fisico nella sezione interna della rete a circuito fra un autocommutatore di origine e uno di destinazione avviene in accordo a una ricerca di percorso e a una operazione di impegno effettuate con il metodo *sezione per sezione*. Perciò, in una connessione che coinvolge uno o più nodi della rete a circuito, le informazioni di segnalazione relative a tale connessione devono essere ricevute ed elaborate in ogni nodo intermedio tra quelli di origine e di destinazione.

Una rete SCC è una rete di calcolatori tra gli elaboratori preposti al comando degli autocommutatori della rete a circuito. L'informazione è scambiata sotto forma di messaggi. La metodologia dell'architettura a strati ben si presta a descrivere i protocolli della rete SCC. L'SS no.7 è organizzato in quattro *livelli*. I tre livelli più bassi, con i numeri d'ordine 1, 2 e 3, sono, almeno orientativamente, in corrispondenza con i primi tre strati (1, 2 e 3) del modello OSI e costituiscono, nel loro insieme, la cosiddetta *parte di trasferimento di messaggio* (MTP - Message Transfer Part).

Il livello 4 dell'SS no. 7 comprende funzioni, sempre orientativamente, in corrispondenza con gli ultimi quattro strati (4, 5, 6 e 7) del modello OSI. Tali funzioni sono mostrate in Fig. V.7.2 e costituiscono le *parti di utilizzazione* (UP - User Part) nell'SS no.7; nella stessa figura è sottolineata la corrispondenza tra gli strati OSI e i quattro livelli di questo sistema di segnalazione.

Tra le parti di utilizzazione si distinguono innanzitutto la *Parte di controllo della connessione di segnalazione* (SCCP - Signalling Connection Control Part) e le *Potenzialità di transazione* (Transaction capabilities). In particolare, la funzione SCCP fornisce i mezzi per controllare connessioni logiche in una rete SCC; essa quindi modifica il servizio offerto dall'MTP, a favore di quelle UP che richiedano un servizio orientato alla connessione, per trasferire segnalazione o altre informazioni a questa relative, quali sono utilizzate in applicazioni di esercizio e manutenzione o con lo scopo di controllare specifiche transazioni.

Le Potenzialità di transazione sono composte da due elementi: la *Parte di applicazione delle potenzialità di transazione* (TCAP - Transaction Capability Application Part) e la *Parte di servizio intermedio* (ISP - Intermediate Service Part). Tali potenzialità sono utilizzate per gestire l'interazione tra nodi di commutazione della rete a circuito e centri di servizio, nell'ambito di specifiche applicazioni, quali la gestione centralizzata dei gruppi chiusi di utente (cfr. § IV.3.3) e l'autenticazione di carte di credito.

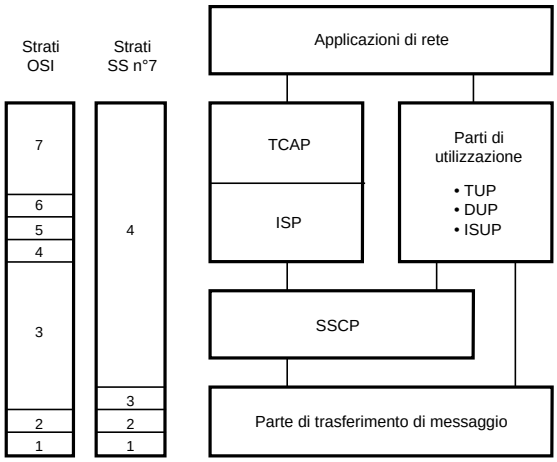


Figura V.7.2- Struttura generale delle funzioni dell'SS no. 7.

Altre parti di utilizzazione, che utilizzano i servizi forniti dall'MTP (direttamente o indirettamente tramite l'SCCP) sono definite sia per trattare il controllo di specifiche connessioni, sia per supportare altre applicazioni legate, ad esempio, all'esercizio e alla manutenzione della rete SCC. Si hanno:

- la *UP per il servizio telefonico* (*TUP* - Telephony User Part), che è preposta alla segnalazione necessaria per controllare le chiamate telefoniche;
- la *TUP migliorata* (*TUP-E* - TUP Enhanced), che è stata definita per fronteggiare le prime realizzazioni di una ISDN;
- la *UP per i servizi di comunicazione di dati* in reti dedicate a circuito (*DUP* - Data User Part);
- la UP, che definisce le procedure di chiamata in una ISDN (*ISUP* - ISDN User Part);
- la *parte di applicazione per l'esercizio e la manutenzione* (*OMAP* - Operation and Maintenance Application Part), che è preposta alle procedure di sorveglianza della rete SCC svolte da un SP preposto a funzioni di controllo (*CSP* - Control Signalling Point).

La Fig. V.7.3 mostra poi in dettaglio le funzioni principali dell'MTP. Nel livello 1, abbiamo le funzioni del *circuito di dati di segnalazione* (Signalling Data Circuit), comprendendo in queste la trasmissione fino a 64 kbit/s e lo scambio di protezione di linea. Il livello 2 include le funzioni di *collegamento di dati di segnalazione* (Signalling Data Link), quali la strutturazione in opportune unità informative, la sequenzializzazione e il recupero d'errore.

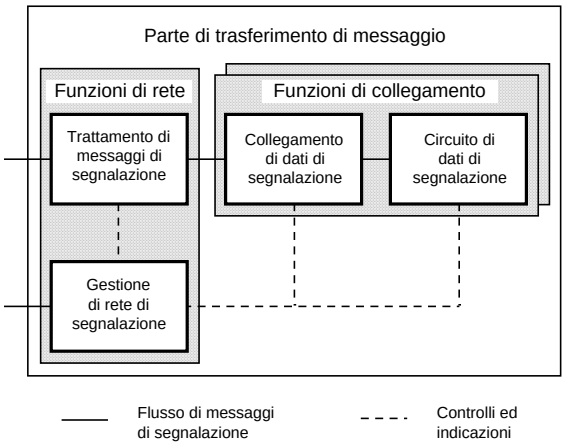


Figura V.7.3 - Struttura della parte di trasferimento di messaggio (MTP).

Nel livello 3, si distinguono le funzioni di *trattamento di messaggio* (Message Handling) e di *gestione di rete* (Network Management). La prima di queste comprende l'instradamento in accordo a un servizio di rete senza connessione. La seconda riguarda le procedure addizionali per il funzionamento della rete SCC in condizioni anormali, quali guasti e sovraccarichi.

La Fig. V.7.4 illustra infine i protocolli di comunicazione tra due SP per il tramite di un STP intermedio. I protocolli di livello 4 dell'SS no. 7 sono gestiti da estremo a estremo attraverso la rete SCC, dato che gli STP trattano funzioni solo fino al livello 3.

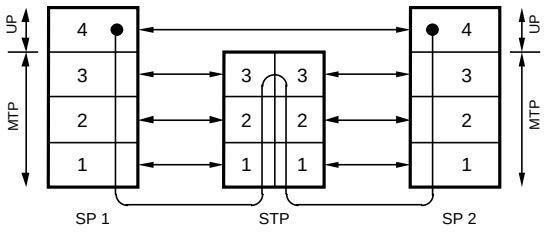


Figura V.7.4 - Architettura protocollare dell'SS no. 7.