INDICE TESI

Introduzione

L'obiettivo che intende essere raggiunto dallo studio illustrato nelle pagine seguenti, consiste nella progettazione e implementazione su calcolatore di un protocollo di accesso al mezzo.

Il presente lavoro,infatti , è nato con lo scopo di realizzare un protocollo MAC per l'accesso al mezzo radio condiviso, per sistemi OFDM (Orthogonal Frequency Division Multiplexing) in una Wireless LAN indoor.

La tratta considerata è quella di downlink nella quale è stato scelto di considerare il sistema di modulazione OFDM (come secondo specifica del progetto), affiancato da un protocollo di accesso al mezzo di tipo CDMA.

Principali requisiti da rispettare sono stati:

- una gestione del traffico che fosse la più realistica possibile e, quindi, tenesse conto non solo delle differenti tipologie di traffico ma anche dei processi di nascita e morte all'interno della rete;
- un'allocazione delle risorse ottima sotto determinate condizioni di sistema e rispettando o comunque tenendo presenti le differenti strategie di scheduling note in letteratura.

La tesi è strutturata in sette capitoli.

Il primo capitolo fa una rapida panoramica del progetto da cui è nata la tesi. Si descrive chi partecipa al progetto, le attività svolte dalle varie unità e le linee guida di quanto affrontato invece dall'unità di Roma, per poi soffermarsi sulla descrizione di cosa è una Wireless LAN, sulle sue applicazioni, sulla tecnologia alla base di queste e sugli standard ora esistenti: IEEE 802.11 ed HIPERLAN 2

Il secondo capitolo tratta in maniera abbastanza approfondita la tecnica di modulazione OFDM e il suo utilizzo se affiancata con tecniche di accesso multiplo quali TDMA, FDMA e CDMA. Vengono presentate brevemente anche tecniche più avanzate, da considerarsi come alternative future alla tecnica OFDM-CDMA, considerata nell’ambito del sistema sotto esame.

Il terzo capitolo illustra la struttura del sistema, lo scenario di riferimento, il sistema di trasmissione scelto per la tratta presa in esame ed i vari strati considerati, dallo strato di rete a quello fisico, descrivendo le caratteristiche di ognuno.
Il quarto capitolo comincia ad entrare nel cuore della tesi, trattando in maniera piuttosto approfondita da un lato la caratterizzazione delle sorgenti di traffico e la loro modellizzazione e dall'altro le classi di servizio ed il modo in cui può essere gestita la risorsa radio disponibile a strato MAC.
Nel quinto capitolo vengono delineate le varie strategie di scheduling note in letteratura .
Vengono poi affrontati i due tipi di scheduling, utilizzati per la simulazione del protocollo MAC, uno di natura statica, l'altro di natura dinamica. Entrambe i meccanismi di scheduling si basano sul protocollo di assegnazione delle risorse RSVP, ma mentre il primo concede la gran parte della risorsa alla classe GB, il secondo tenta di garantire un'assegnazione equa sempre rispettando le priorità delle diverse classi di servizio garantendo se necessario un minimo di risorsa anche alla classe BE e permettendo allo scheduling di transitare attraverso varie fasi.
Come parallelo al capitolo 5, il sesto capitolo mostra l'evoluzione del simulatore da un ambiente statico, ad un ambiente caratterizzato da processi di nascita e morte e da un algoritmo di scheduling adattativo.
Infine il settimo capitolo è il capitolo dedicato ai risultati, ai grafici ed ai confronti tra i due meccanismi di scheduling implementati

CAPITOLO 1

INTRODUZIONE ALLE RETI WLAN E PROGETTO "SISTEMI OFDM CON APPLICAZIONE ALLE RETI WLAN"

1.1 INTRODUZIONE ALLE RETI WIRELESS E MODELLI DI UNA WLAN

Le reti wireless si pongono come obiettivo quello di liberare l'utente dai vincoli e dai costi imposti dal cablaggio. Raggiungere un obiettivo di tale portata, necessita della definizione di standard che riassumano le tecnologie e rendano interoperabili i vari apparati tra loro.

Quando si parla di Wireless LAN si considera un tipo di rete in area locale, LAN (Local Area Network), che utilizza onde radio ad alta frequenza piuttosto che "fili" per comunicare e trasmettere dati attraverso nodi. E' un sistema di comunicazione flessibile ai dati implementato come estensione, o a volte come alternativa, ad una Wired LAN all'interno di un edificio (*in building*) o di un *ambiente di campus.*.

	WIRELESS LOCAL AREA NETWORK (WLAN)	LAN-LAN BRIDGE	WIRELESS WIDE AREA NETWORK (WWAN)	WIRELESS METROPOLITAN AREA NETWORK (WMAN)	WIRELESS PERSONAL AREA NETWORK (WPAN)
COVERAGE AREA	IN BUILDING OR CAMPUS	BUILDING TO BUILDING	NATIONAL	METROPOLITAN AREA	A FEW FEET
FUNCTION	EXTENSION OR ALTERNATIVE TO WIRED LAN	ALTERNATIVE TO WIRED CONNECTION	EXTENSION OF LAN	EXTENSION OF WIRED LAN	ALTERNATIVE TO CABLE
USER FEE	NO	NO	YES	YES	NO
TYPICAL THROUGHPUT	1-11 Mbps	2-200 Mbps	1-32 Kbps	10-100 Kbps	0.1-4 Mbps

Fig.1.1 - Differenze ad alto livello tra le WLANs ed altre tecnologie

Per quanto riguarda la copertura di una WLAN, quando la collocazione delle stazioni all'interno di un edificio non varia o varia lentamente, si parla di ambiente "*in building tethered"*. Questo segmento di mercato copre i vecchi edifici dove, come accennato precedentemente, diventa spesso difficile e in ogni caso troppo costoso installare nuove reti cablate. In un ambiente di tipo "*in building non tethered"*, viene sfruttata, invece, la caratteristica di mobilità delle reti Wireless. Si fornisce cioè una connessione tra un PC portatile ed i servizi di una LAN, mentre l'utente può liberamente spostarsi all'interno dell'edificio.
Si parla di ambiente di campus quando vi sono più edifici vicini situati all'interno di un'area limitata. Anche in tal caso le reti wireless rispondono alle esigenze di connessione tra gli edifici e di mobilità delle stazioni all'interno del campus.

1.1.1 Applicazioni per Wireless LANs e loro benefici

Le WLANs accrescono, piuttosto che sostituire, le potenzialità delle reti Wired LAN, fornendo il cosiddetto collegamento "dell'ultimo miglio" tra una rete dorsale e l'utente mobile in building od in ambiente di campus. La lista che mi accingo a presentare descrive alcune delle tante applicazioni possibili di wireless LANs:

- in ambito ospedaliero, medici ed infermieri possono cooperare in maniera più produttiva perché mediante *handheld computers* o notebooks collegati mediante WLAN possono scambiare informazioni sui pazienti in tempo reale;
- gli amministratori di rete in ambienti dinamici minimizzano gli overhead di gestione causati da spostamenti, estensioni di rete e da cambiamenti in generale; inoltre installare reti di computer in vecchi edifici senza il bisogno di creare infrastrutture costose
- studenti possono accedere ad Internet o all'Intranet della facoltà per consultare i cataloghi delle biblioteche da qualsiasi punto nel campus universitario;

- negozianti usano WLAN per scambiare informazioni con database centrali e accrescere la loro produttività;
- gli amministratori di rete implementano WLAN per provvedere al backup di applicazioni importanti che girano su reti wired;
- dirigenti d'azienda in sede di conferenze prendono decisioni più rapide grazie alle informazioni real-time che ottengono tramite WLAN.

Come si può constatare da alcune delle applicazioni citate, i benefici risultanti sono molteplici. La crescita smodata di Internet e dei servizi online sono testimonianze forti di informazioni e risorse condivise. Con una WLAN, gli utenti possono aver accesso alle informazioni senza cercare un posto dove inserire la spina, ed gli amministratori di rete possono progettare e potenziare la rete senza dover installare o muovere fili.

La *mobilità* aumenta la *produttività* ed il *servizio,* ed i sistemi WLAN possono fornire agli utenti accesso ad informazioni real-time dovunque nella loro organizzazione. Questa mobilità supporta produttività e opportunità di lavoro non possibile con le reti wired.

La *velocità di installazione e la semplicità* rendono molto più semplice l'installazione di una WLAN ed eliminano la necessità di far passare cavi attraverso muri e soffitti.

La *flessibilità di installazione* permette alla tecnologia wireless di far andare la rete laddove i cavi non possono.

I *ridotti costi di "possesso"* sono tali da poter dire che mentre l'investimento iniziale richiesto per la parte hardware di una wireless LAN può risultare maggiore di quello della parte hardware di una wired LAN, l'intera spesa di installazione ed i costi del ciclo di vita possono essere significativamente più bassi.

I benefici di costo a lungo termine si riscontrano maggiormente in ambienti dinamici dove, cioè, cambiamenti e spostamenti risultano molto frequenti.

La *scalabilità* è tale che le WLAN possono essere configurate in varie topologie per incontrare le necessità di specifiche applicazioni ed installazioni. Le configurazioni

possono variare da reti indipendenti adatte per un esiguo numero di utenti, ad infrastrutture di rete di mille o più utenti che accettano roaming su una grande area.

1.1.2 Come funzionano le WLANs e loro configurazioni

Le WLANs usano onde elettromagnetiche (radio o a infrarossi) per comunicare informazioni da un punto ad un altro senza supporto di alcuna connessione fisica. I dati da trasmettere sono modulati su portanti radio così da poter essere estratti accuratamente nel punto di ricezione. Una volta che i dati sono stati modulati sulla portante radio, il segnale radio occupa un intervallo di frequenze, dal momento che la frequenza o il bit rate del segnale modulato si sovrappone alla portante. Portanti radio multiple possono sussistere nello stesso spazio e nel medesimo tempo senza interferire l'una con l'altra se le onde radio vengono trasmesse su frequenze diverse. Per estrarre dati, un ricevitore radio si sintonizza su una frequenza radio mentre rigetta tutti i segnali radio di frequenze diverse.
In una tipica configurazione WLAN si evidenzia un dispositivo trasmettitore-ricevitore, il transceiver, chiamato *Access Point (AP)*, che si connette alla rete cablata da una locazione fissa usando cavi standard Ethernet. Esso deve poter ricevere, bufferizzare e trasmettere dati tra la WLAN e l'infrastruttura della rete cablata[Pahlavan K. et altri,1996-1997].
Esistono due tipi di AP (come mostrato nelle figure 1.2.1 e 1.2.2):

***Dedicated Hardware Access Points (HAP)*:** sono dei dispositivi stand alone che si connettono direttamente alla rete cablata e fungono da interfaccia tra questa e i dispositivi wireless nella propria area di copertura.

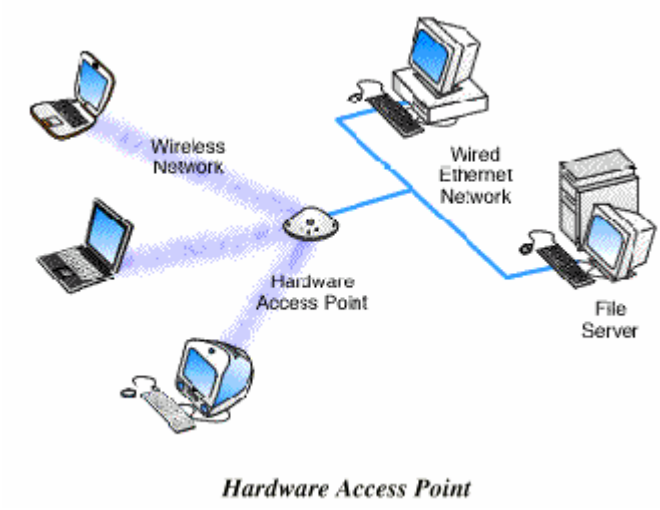


Figura 1.2.1- Hardware Access Point

***Software Access Point (SAP)*:** poiché lo standard utilizzato prevede soltanto la possibilità di collegare reti Ethernet alla WLAN, con un HAP risulterebbe impossibile implementare estensioni wireless per altri tipi di rete (Token Bus, Token Ring, …).

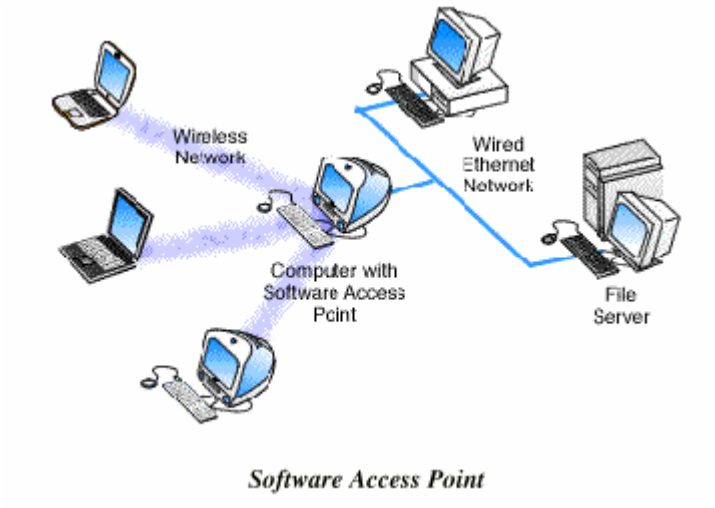


Figura 1.2.2 – Software Access Point

Un singolo access point è in grado di supportare un piccolo gruppo di utenti e può funzionare all'interno di un range che varia tra i 30 ed i 150 metri. L'access point (o l'antenna connessa all'access point) è usualmente posta in alto ma può essere montata in qualsiasi posizione, fino al raggiungimento della copertura radio desiderata. Gli utenti accedono alla WLAN mediante appositi adattatori che

possono essere implementati come schede PCMCIA per notebooks o computer palmtop, oppure come schede in computer desktop o integrati in computer handheld.

Tali adattatori forniscono un'interfaccia tra il sistema operativo di rete (NOS) e le onde elettromagnetiche mediante l'antenna. Il tipo di collegamento senza fili è del tutto trasparente al NOS.

La più semplice configurazione WLAN è quella detta "*peer to peer*" (o "*ad hoc*",come si può vedere in fig.1.3) e consiste in una WLAN indipendente che connette un insieme di PCs con adattatori wireless. In qualsiasi momento due o più adattatori wireless si trovano all'interno del range di ogni altro e possono organizzare una rete indipendente. Queste reti tipicamente non richiedono nessuna amministrazione o configurazione. Ciascun utente avrà accesso solamente alle risorse di un altro utente e non anche ad un sistema centrale. Le stazioni, quindi, comunicano direttamente l'una con l'altra e non è necessaria l'installazione di alcuna infrastruttura. Unico svantaggio di tale configurazione l'area di copertura limitata.

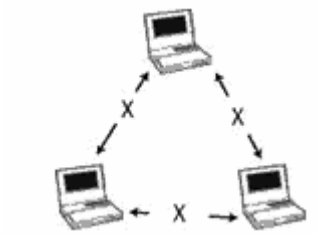


Fig.1.3 - WLAN indipendente o "ad hoc" (peer to peer)

Per ovviare, almeno in modo parziale, all'area di copertura limitata della configurazione precedente gli access point possono estendere il range delle WLANs indipendenti agendo come ripetitori, raddoppiando così la distanza tra PC wireless. Tale strategia è mostrata in fig.1.4.

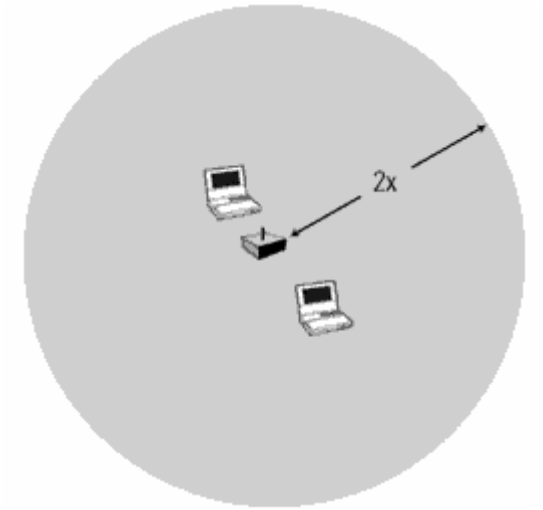


Fig.1.4- WLAN “ad hoc” di range esteso usando l’access point come ripetitore

Nelle WLAN “*ad infrastruttura*” (un esempio è riportato in fig.1.5), più access point collegano la WLAN alla rete wired e permettono agli utenti di condividere efficientemente le risorse di rete. Gli access points non solo provvedono alla comunicazione con la rete wired ma mediano anche il traffico della rete wireless nelle immediate vicinanze. Più access points possono provvedere coperture di aree per un intero edificio o campus.

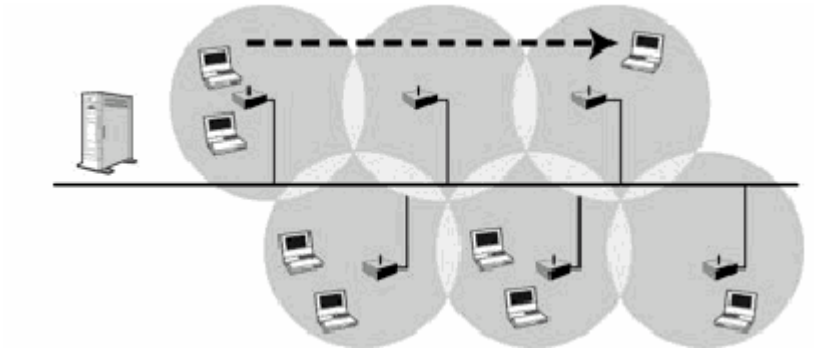


Fig.1.5- WLAN a infrastruttura

Le comunicazioni wireless sono limitate per le potenze di uscita in gioco a dei limiti massimi di distanza raggiungibile dal segnale. Le WLAN usano celle, dette microcelle, simili a quelle del sistema GSM per estendere il range della connettività wireless. Da ciò si desume che un PC mobile completo di adattatore WLAN è

associato con un singolo punto di accesso e la sua microcella, o area di copertura. Microcelle singole si sovrappongono per consentire comunicazioni continue all'interno della rete cablata. In questo modo viene realizzata una gestione efficiente di segnali a bassa potenza e handoff di utenti che attraversino una data area geografica. Una topologia che porta alla procedura di hand-off è mostrata in fig.1.6.

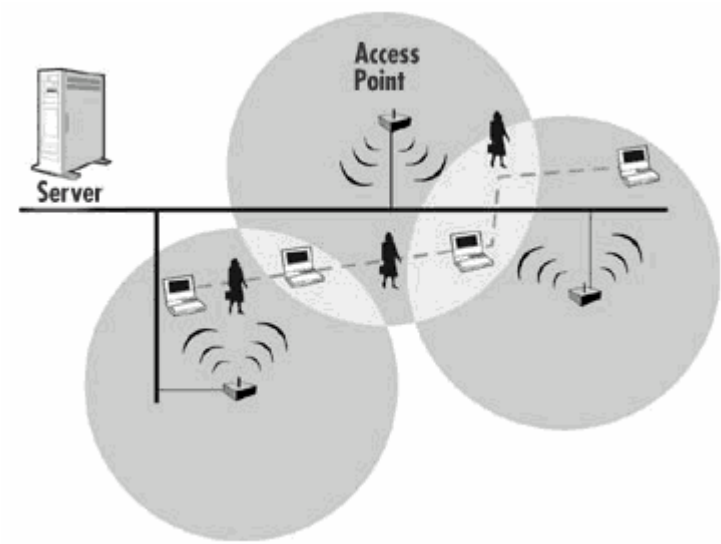


Fig.1.6 - Gestione dell'handoff in una connessione WLAN tra access point

Le proposte standardizzate ed in via di standardizzazione di nuove architetture WLAN (IEEE 802.11 e ETSI HIPERLAN II), rispondono ad esigenze basilari dell'utenza, garantendo bit-rate fino a 54 Mbit/s con trasmissione a pacchetto. Di seguito, ho riportato la descrizione e le caratteristiche fondamentali degli standard.

1.2 STANDARD ESISTENTI

1.2.1 Lo Standard IEEE 802.11

L'IEEE 802.11 è un gruppo di lavoro che si occupa della standardizzazione del livello MAC e del livello fisico delle reti locali wireless. Il gruppo di lavoro 802.11 si è prodigato nello sviluppo di uno standard globale per sistemi radio e reti operanti

nella banda di frequenza “unlicensed” dei 2.4GHz con data-rates dell’ordine di 1 e 2Mbps. Lo standard non specifica tecnologie o implementazione ma semplicemente definisce le specifiche per gli strati fisico e MAC (Medium Access Control) per connessioni wireless per stazioni fisse, portatili ed in movimento all’interno di un’area locale (in building o campus) in grado di supportare velocità trasmissive multiple, scelte a seconda dello stato del mezzo e della capacità delle stazioni, e comunque superiori ad 1 Mb/s.[IEEE Std 802.11-1997 Information Technology]. Uno degli scopi principali di questo standard è che un singolo MAC possa supportare più livelli fisici, anche se questi fanno uso di tecnologie diverse. Il wireless MAC supporta sia servizi connectionless a velocità comprese tra 1 e 20 Mb/s, sia servizi di tipo isocrono (time bounded) per controllo di processi, voce e video.

Lo strato fisico in qualsiasi rete definisce le caratteristiche di modulazione e trasmissione delle informazioni. A livello di strato fisico sono state definiti due metodi di trasmissione a RF ed uno ad infrarossi.

Gli standards di trasmissione a RF sono:

- Frequency Hopping Spread Spectrum (FHSS)
- Direct Sequence Spread Spectrum (DSSS)

Il metodo ad infrarossi opera nella banda base, mentre i due metodi a RF operano nella banda tra i 2.4GHz e i 2.483GHz.

I sistemi ad infrarossi utilizzano le stesse frequenze usate nelle fibre ottiche e sono denotati dalla caratteristica di “scoprire” solo l’ampiezza del segnale, in tal modo l’interferenza è notevolmente ridotta. Essendo sistemi non limitati in banda possono raggiungere velocità maggiori rispetto ad altri sistemi. La tecnologia IR era inizialmente molto popolare per il fatto di essere caratterizzata da elevati data rates e costi relativamente bassi. Lo svantaggio dei sistemi IR è quello che lo spettro di trasmissione è condiviso con sole ed altre luci fluorescenti. Se c’è molta interferenza da altre sorgenti la LAN può diventare inutile. I sistemi IR richiedono una LOS (line of sight) libera da ostacoli e si deve anche ricordare che i segnali IR

non sono in grado di penetrare oggetto opachi. Ciò vuol dire che oggetti come muri, divisori, cortine o anche nebbia possono ostruire il segnale InfraLAN è un tipico esempio di WLAN che usa tecnologia ad infrarossi.

I sistemi a radio frequenza utilizzano la tecnologia Spread Spectrum Frequency Hopping e la Direct Sequenze Spread Spectrum. A causa di valori di overhead più elevati rispetto ai sistemi IR, i data rates supportati sono inferiori.

Con la DSSS la trasmissione del segnale è allargata su tutta la banda permessa (per esempio 25MHz). Una sequenza binaria, il codice di spreading, è usata per modulare il segnale da trasmettere. I bit di informazione sono mappati in una struttura di "chips" a loro volta mappati in un bit a destinazione. Il numero di chips rappresentanti un bit è il cosiddetto rapporto di spreading. Maggiore è quest'ultimo valore, più il segnale risulterà resistente all'interferenza. Più basso è tale rapporto, maggiore sarà la quantità di banda disponibile all'utente. L'F.C.C. (Federal Communication Commission) ha imposto che il rapporto di spreading deve essere maggiore di 10, lo standard 802.11 ha imposto un valore pari ad 11. trasmettitore e ricevitore devono essere sincronizzati sul medesimo codice di spreading. Se vengono utilizzati codici di spreading ortogonali, allora più di una LAN può condividere la stessa banda.

La tecnica FHSS suddivide la banda in tanti piccoli sottocanali (1 MHz). Il segnale salta allora da un sottocanale all'altro per trasmettere piccoli burst di dati su ogni canale per un certo periodo di tempo, detto "*dwell time*". La sequenza di hopping deve essere sincronizzata al trasmettitore e al ricevitore altrimenti l'informazione è persa. L'FCC richiede che la banda sia suddivisa in almeno 75 sottocanali e che il "dwell time" non sia maggiore di 400ms. Il FHSS è meno sensibile all'interferenza poiché la frequenza viene continuamente "shiftata". Ciò rende i sistemi FHSS difficili da intercettare, garantendo un elevato grado di sicurezza. Non a caso tali sistemi vengono usati in ambito militare. Per disturbare un sistema di tale tipo è necessario che tutta la banda sia bloccata. Molte LAN FHSS possono essere situate vicino se sono usate frequenze di hopping ortogonali. Poiché i sottocanali sono più piccoli di quelli dei sistemi DSSS, il numero di LAN collocate nella stessa area

geografica può essere più grande con i sistemi FHSS [ISO-IEC 8802.11;ANSI/IEEE Std 802.11, 1999 edn, 20 Aug.1999 Information Technology].

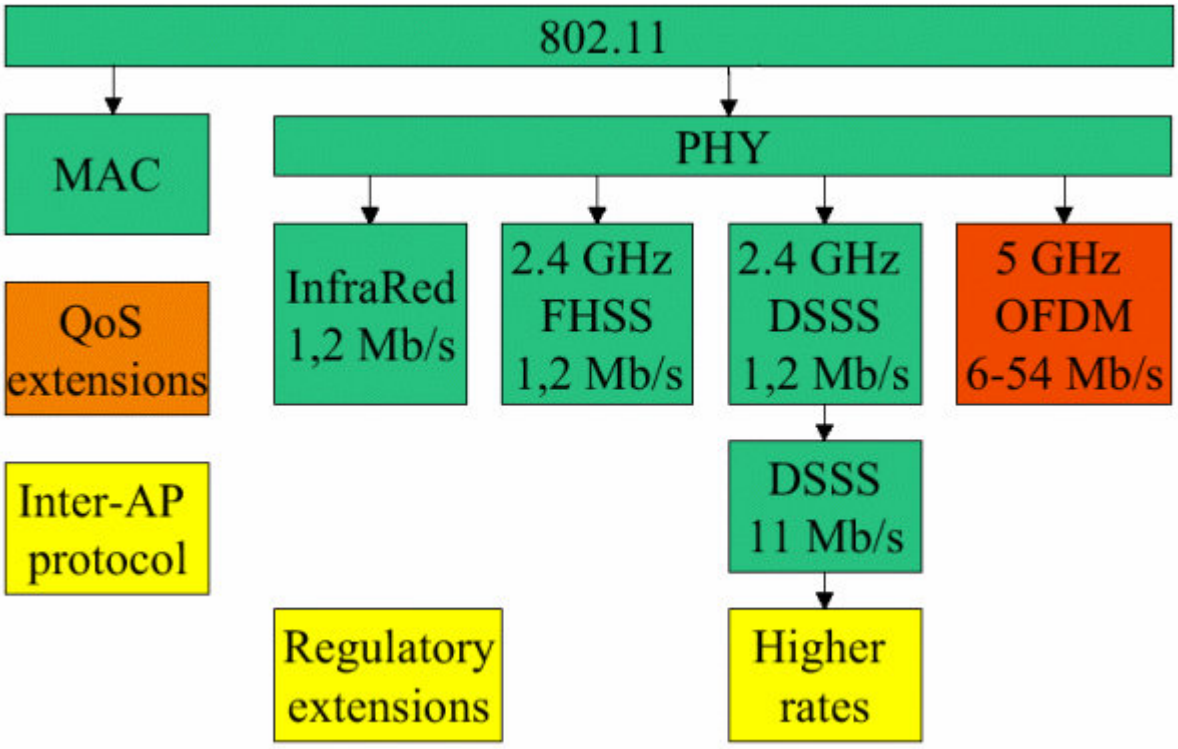


Fig.1 7– Struttura dello standard 802.11

Per quanto riguarda lo strato MAC, questo risulta essere caratterizzato da un insieme di protocolli responsabili dell'utilizzo intelligente del mezzo condiviso. Lo standard 802.11 specifica il protocollo CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance). Esso è utilizzato per la trasmissione asincrona, e può essere affiancato da una funzione di coordinamento centralizzata (PCF - Point Coordination Function) con priorità superiore per i servizi tempo limitati. In questo protocollo, quando una stazione deve trasmettere un pacchetto, per prima cosa deve verificare che nessun'altra stazione stia trasmettendo, ascoltando il mezzo (Carrier Sense). Se è riconosciuta la presenza di trasmissioni si mette in attesa. Quando il mezzo si libera attende che rimanga tale per un intervallo di tempo minimo (DIFS - Distributed InterFrame Space) e successivamente inizia una fase di contesa (Contention Window): la stazione sceglie un intervallo casuale (backoff) al termine del quale, se il mezzo è ancora libero, la stazione può trasmettere il pacchetto. Dal

momento che la probabilità che due nodi scelgano lo stesso fattore di backoff è piccola, le collisioni tra pacchetti saranno minimizzate. L'intervallo di backoff è scelto tenendo conto di un parametro che oscilla tra un valore massimo ed uno minimo, raddoppiando ogni volta che si deve ripetere la trasmissione di un frame. In tal modo si allunga la finestra di contesa riducendo la probabilità di collisione nel caso di carico elevato della rete. Quando una stazione, in attesa che termini il backoff, sente che il mezzo non è più libero, congela il tempo di backoff rimasto. Quando rileva il mezzo libero per un tempo pari a DIFS, non sceglie un nuovo tempo di attesa ma termina il precedente. La detezione di collisione (Collision Detection) non può, nello standard 802.11, essere utilizzata come in Ethernet, in quanto una stazione che trasmette non può contemporaneamente ascoltare un'altra stazione nel sistema che potrebbe essere in trasmissione, dal momento che, il proprio segnale copre qualsiasi altro segnale arrivi alla stazione. Se un pacchetto sta per essere trasmesso, la stazione trasmittente invia prima un pacchetto corto RTS (ready-to-send) contenente l'informazione della lunghezza del pacchetto. Se la stazione ricevente ascolta l'RTS, risponde con un altro pacchetto corto CTS (clear-to-send). Dopo questo scambio, la stazione trasmittente manda il suo pacchetto. Quando il pacchetto è ricevuto con successo, come determinato dal CRC (Cyclic Redundancy Check), la stazione ricevente trasmette un pacchetto di acknoledgment (ACK).

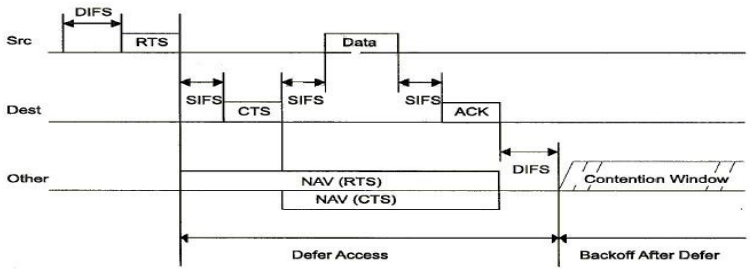


Fig.1.8 – Handshake tra TX e RX

Questo scambio avanti ed indietro è necessario per evitare il problema del terminale nascosto (hidden node).

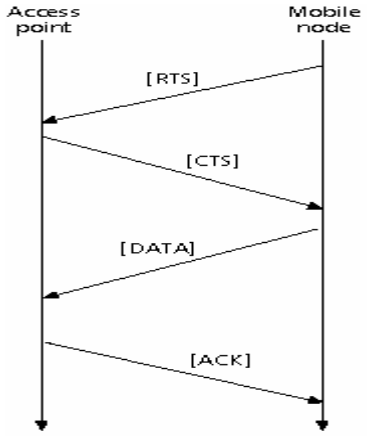


Fig.1.9 – Scambio messaggi RTS-CTS-DATA-ACK

L' "Hidden Node" è un terminale che si trova al limite tra il terminale di destinazione e il terminale di sorgente.Consideriamo la fig.1.10.

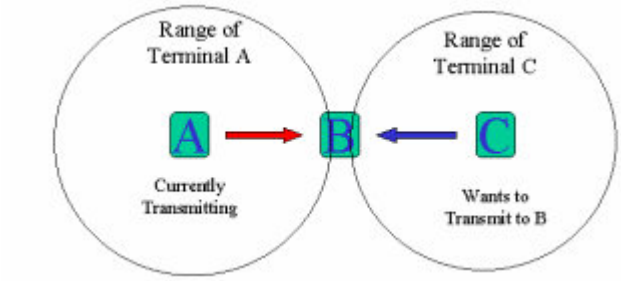


Fig.1.10 - Hidden Node

Il nodo A trasmette al nodo B. Il nodo C non può sentire la trasmissione di A. Durante questa trasmissione, quando C sente il canale, pensa erroneamente che il canale sia "idle". Se, a questo punto, il nodo C inizia una trasmissione verso B, lui stesso interferirà con la ricezione dei dati del nodo B. In questo caso il nodo C è un "hidden node" (terminale nascosto) per A e, come tale, può causare collisioni sulla trasmissione dei dati. Il meccanismo di Handshake serve a far fronte al problema.
Da chiarire che per evitare che durante i messaggi di protocollo si entri in una nuova fase di contention window, il tempo di attesa per i messaggi di risposta e per l'invio dei dati dopo il CTS è più corto del DIFS, tale tempo è detto SIFS (Short InterFrame Space). Per la bassa affidabilità della trasmissione, una stazione potrebbe ricevere i messaggi e iniziare una trasmissione generando una collisione.

Per evitare questo, il protocollo realizza una sorta di "carrier sense virtuale". I messaggi RTS e CTS contengono informazioni sulla durata della trasmissione successiva, che le stazioni non interessate alla ricezione caricano in un registro detto NAV (Net Allocation Vector). Questo registro viene via via decrementato ed ogni stazione ne attende l'azzeramento prima di cominciare la procedura di trasmissione.

Per quanto riguarda la funzione di coordinamento centralizzata precedentemente accennata, questa può essere gestita solo da alcune stazioni, come gli Access Point. La PCF usa una struttura a superframe in cui si alternano periodi di contesa in cui è attivo il DCF a periodi senza contesa dove è attiva la PCF. Il point coordination diventa padrone del mezzo trasmissivo mediante un accesso prioritario. Uno schema indicativo del meccanismo appena descritto è rappresentato in fig.1.8.

I servizi offerti dal MAC dello standard 802.11 sono riassunti nella tabella sotto

Categoria di servizi	Servizio	Scopo
Servizi forniti da ogni stazione	Autenticazione	Utilizzato per verificare l'identità delle stazioni che vogliono stabilire fra loro un link diretto di comunicazione. Non si tratta di autenticazione user to user o end to end. L'802.11 fornisce il supporto e lascia la possibilità di implementare protocolli di autenticazione diversi.
	Associazione	Servizio mediante il quale una stazione entra a far parte di un BSS. Nel caso di rete a infrastruttura tale servizio è fornito unicamente dall'AP. Così il Distributed System sa a quale AP far riferimento per trasmettere un frame alla stazione
	Dissassociazione	Servizio mediante il quale si termina una precedente Associazione. E' una notifica, non può essere rifiutata.
	Privacy	Utilizzato per criptare i messaggi
Servizi forniti dal Distribution System (DS)	Distribuzione	Servizio mediante il quale, usando le informazioni di associazione, le MSDU vengono distribuite all'interno di un DS.
	Integrazione	Permette lo scambio di MSDU tra DS e una rete esistente.
	Riassociazione	Permette il trasferimento di una stazione da un BSS ad un altro, tramite passaggio da associazione della stazione con AP del vecchio BSS a quella con AP del nuovo. Servizio necessario per permettere la mobilità delle stazioni al di fuori del BSS

1.2.2 Lo standard ETSI – HIPERLAN2

HiperLAN, disponibile in varie versioni, è una tecnologia LAN wireless ad alta velocità e di nuova generazione che offre all'utente finale un throughput fino a 25 Mbps.

L'ETSI (European Telecommunications Standards Institute) sta sviluppando gli standard HIPERLAN all'interno del BRAN (Broadband Radio Access Network) che comprende quattro standard: HiperLAN1, HiperLAN2, HiperLink progettato per i backbone radio all'interno degli edifici, HiperAccess progettato per l'uso fisso all'esterno in modo da fornire accesso ad una struttura cablata.

Lo schema HIPERLAN2 è stato promosso da un gruppo industriale chiamato HiperLAN2 Global Forum, che conta tra i suoi componenti alcune grandi aziende come Bosch, Dell Computer, Nokia, Telia, Ti e Xircom. La funzionalità più attraente di HiperLAN2 è l'elevata velocità che a volte viene erroneamente posta pari a 54Mbps. La velocità in aria è 54Mbps, mentre il throughput continuo per le applicazioni è prossimo a 20 Mbps. Un'altra caratteristica di notevole importanza è il supporto della qualità di servizio (QoS), che risulta fondamentale per applicazioni come il video e la voce. L'architettura HiperLAN2 mette a disposizione connessioni con vari tipi di infrastrutture di rete, tra cui Ethernet, IP, ATM, PPP. Le funzionalità di sicurezza comprendono l'autenticazione e la crittografia. Una caratteristica significativa è data dalla gestione automatica della frequenza.

Lo standard HiperLAN2 definisce uno strato Fisico ed uno strato Data-Link. Al di sopra di questi si trova uno strato di convergenza che accetta pacchetti o celle dai sistemi di networking già esistenti e li formatta per la distribuzione sul mezzo wireless. Lo schema di modulazione supportato da HiperLAN2 è l'OFDM (Ortogonal Frequency Division Multiplexing), estremamente efficace in ambiente dispersivo nei confronti del tempo, dove i segnali possono seguire vari percorsi per raggiungere la destinazione, portando alla nascita di ritardi variabili. In relazione all'elevata velocità dei dati, tali ritardi possono raggiungere una proporzione significativa del simbolo trasmesso, portando così a quella che viene definita

“interferenza intersimbolo”. Lo schema OFDM combatte tale eventualità come si vedrà in dettaglio nel capitolo successivo, trasmettendo i dati in parallelo su una molteplicità di sottoportanti.

Nell’allocazione delle frequenze per l’Europa, i canali HiperLAN2 saranno intervallati da una distanza di 20 MHz - per un totale di 19 canali. Ogni canale è suddiviso in 52 sottoportanti, 48 per i dati e 4 come pilota per la sincronizzazione.

La maggior parte delle LAN wireless a bassa velocità non adotta la correzione del FEC, mentre HiperLAN2 offre diversi livelli, ciascuno dei quali è in grado di offrire una protezione nei confronti di una certa percentuale di errori dei bit.

Lo strato Data-Link è orientato alla connessione e questo lo differenzia dalle altre tecnologie wireless. Prima che un terminale mobile trasmetta i dati, lo strato Data-Link comunica con il punto di accesso nel piano di segnalazione in modo da impostare una connessione temporanea. Questo approccio di connessione consente la negoziazione dei parametri QoS, come ritardo e larghezza di banda. Ciò assicura anche che altri terminali non vadano ad interferire con la trasmissione successiva. D’altra parte un terminale mobile conforme allo standard 802.11 inizierà a comunicare quando il canale radio diventa disponibile e potrebbe sperimentare delle collisioni di pacchetti provenienti da altri terminali. Bisogna però sottolineare il fatto che lo standard 802.11, offre un meccanismo separato per le applicazioni sincrone come quelle vocali.

L’HiperLAN2 implementa la QoS tramite i time slot. I parametri Qos comprendono la larghezza di banda, la quantità di errori a livello di bit, la latenza ed il jitter. L’access point accorda l’accesso allocando dei time slot specifici per una durata ben definita, in quelli che vengono chiamati canali di trasporto. Il terminale mobile invia quindi i dati senza avere interruzioni da parte di altri terminali che lavorano sulla medesima frequenza.

Un canale di controllo invia un riscontro al mittente, indicando se i dati sono stati ricevuti con errori e se devono essere ritrasmessi.

Sopra il Data –Link c’è lo strato di convergenza che risponde alle richieste di servizio da parte degli strati più alti e formatta i dati come richiesto.

Questo strato supporta comunicazioni basate su pacchetti (Ethernet) e su celle (ATM). L'HiperLAN2 viene inoltre fornito con l'allocazione AFA (Automatic Frequency Allocation). Per garantire una copertura continua, i punti di accesso devono avere aree di copertura che si sovrappongono. Di solito la copertura si estende dai 30 m in ambienti chiusi ai 150 in ambienti all'aperto privi di ostacoli. I punti di accesso tengono sotto controllo i canali radio HiperLAN attorno a queste aree e scelgono automaticamente un canale non utilizzato. Ciò elimina il problema della pianificazione delle frequenze. Quando un terminale mobile si sposta dall'area di copertura di un punto di accesso a quella di un altro, avvia un handoff con il nuovo punto di accesso dopo aver rivelato un segnale migliore su un nuovo canale radio. Il nuovo punto di accesso ottiene dal precedente le informazioni sulla connessione del terminale mobile, quindi le comunicazioni continuano in modo omogeneo. Lo schema HiperLAN2 rende sicure le comunicazioni da un terminale mobile, creando una sessione chiamata associazione con un punto di accesso usando in primo luogo lo scambio di chiavi per negoziare una chiave di sessione segreta, e successivamente un processo di reciproca autenticazione a chiave segreta o pubblica. Il traffico dati viene, infine, crittografato usando lo schema DES (Data Encryption Standard) o Triple DES.

1.3 PROGETTO "SISTEMI OFDM CON APPLICAZIONE ALLE RETI WLAN"

La scelta della tecnologia per la realizzazione di una rete wireless è strettamente legata alla topologia ed alla tipologia della rete stessa.

Negli ultimi anni sono aumentate notevolmente le richieste riguardanti la disponibilità di sistemi di trasmissione radio (sistemi wireless) per ambienti interni ad elevato bit-rate, WLAN, che permettano il collegamento di stazioni di lavoro fisse o portatili ai servizi di rete disponibili sui posti di lavoro. I dispositivi attualmente sul mercato per WLAN non soddisfano completamente le esigenze dell'utenza, poiché i bit-rate supportati si limitano al Mbit/s (intorno al 1.6 Mbit/s).

E’ in questo contesto che si è sviluppata l’idea del progetto ”Sistemi OFDM con applicazione alle reti WLAN”.
La tecnica di modulazione, supportata da entrambe gli standard, è basata sulla tecnica di modulazione OFDM. Il progetto, facendo riferimento agli standard esistenti, fonda le sue basi proprio su questo tipo di modulazione.
HIPERLAN II ed una modalità di IEEE 802.11 prevede che la WLAN serva come accesso radio ad una rete di trasporto fissa a larga banda, per questo la struttura considerata è simile a quella di un sistema cellulare con una stazione radio base e dei terminali mobili ad essa collegati. E’ a questo tipo di WLAN che si fa riferimento nel programma di ricerca. Studi recenti hanno portato all’approfondimento di sistemi di trasmissione e protocolli di accesso di tipo OFDM-CDMA. Ciò vuol dire e nei capitoli seguenti sarà poi chiarito, come, continuando a considerare una tecnica di modulazione di tipo OFDM, si riescano ad ottenere migliori risultati, affiancando a tale tecnica un tipo di accesso al canale basato su FDMA e CDMA e permettendo in tal modo una gestione delle trasmissioni multimediali con qualità di servizio garantita per alcuni flussi di informazione come voce e video. In quest’ambito il progetto si rivolge allo studio più approfondito della modulazione OFDM-CDMA e alla valutazione dell’effetto dei protocolli di accesso radio sulle prestazioni complessive del sistema di rice-trasmissione perseguendo due obiettivi:

- Valutare le prestazioni derivanti da determinati schemi di modulazione, protocolli di accesso alla rete, algoritmi di ricezione, protocolli per il trasporto di alcuni servizi IP a larga banda in modo da riuscire a garantire un utilizzo ottimo delle risorse disponibili
- Realizzare un dimostratore che consenta di vedere se le prestazioni teoriche sono confermate anche dalla pratica

Al fine di raggiungere tutti gli obiettivi, il progetto è stato suddiviso nei seguenti temi di ricerca:

TEMA 1: Elaborazione numerica del segnale al trasmettitore;
TEMA 2: Elaborazione numerica del segnale al ricevitore;
TEMA 3: *Tecniche di assegnazione della risorsa radio;*
TEMA 4: Piattaforma DSP per il dimostratore.

Il TEMA 3 mira alla definizione di un protocollo MAC, che garantisca un sistema di accesso alla risorsa radio efficiente per la fornitura di servizi IP classici e a qualità garantita in una WLAN basata su OFDM-CDMA. Questa tesi rientra all'interno del tema 3 e mira all'approfondimento di due aspetti per la realizzazione del protocollo MAC: uno teso alla gestione di un traffico che risulti di tipo realistico, l'altro alla caratterizzazione di differenti discipline di scheduling per garantire QoS supportando le classi di traffico definite in precedenza.
Partendo dapprima da una gestione statica del traffico, in cui le sorgenti nel sistema erano fisse, l'evoluzione considerata si è basata sulla razionalizzazione del sistema, che in una ottica più realistica mostra l'evoluzione del traffico attraverso l'introduzione dei processi di nascita e morte. Grazie a questa rappresentazione del sistema è stato possibile realizzare diversi algoritmi di scheduling che mirassero ad una gestione intelligente della risorsa disponibile. Come verrà mostrato nei capitoli successivi, la gestione delle varie classi di servizio è stata realizzata tenendo presente non solo i vincoli di priorità legati ad ogni classe ma anche una sorta di equità di trattamento delle varie sorgenti, in modo da non penalizzare troppo le risorse a qualità non garantita.

CAPITOLO 2

LA MODULAZIONE OFDM E LE TECNICHE DI ACCESSO MULTIPLO

2.1 LA TECNICA DI MODULAZIONE OFDM

La rapida crescita dei servizi Internet ed il crescente interesse per una totale mobilità e connettività per i computer portatili hanno verosimilmente creato una forte domanda per servizi dati wireless ad alta velocità.

Una delle tecniche più promettenti per raggiungere trasmissioni con elevato data-rate in un ambiente mobile è l'*OFDM* (Orthogonal Frequency Division Multiplexing) che, in linea di principio, divide una banda frequenziale larga in molti sottocanali a banda stretta, attraverso i quali la trasmissione avviene in parallelo. La banda del sottocanale è tipicamente scelta stretta in modo da eliminare gli effetti del *delay spread*[1], ovvero della selettività in frequenza dovuta al fading.

La tecnica OFDM fornisce dunque un interessante approccio nelle comunicazioni mobili per raggiungere un'alta efficienza spettrale[2] e combattere la selettività in frequenza del canale. Tali vantaggi giustificano il perchè gli standard per le WLAN abbiano scelto proprio questa tecnica di modulazione come supporto per i loro sistemi. Prima di mostrare lo schema di modulazione MC (Multi Carrier) e successivamente lo schema OFDM facente uso di IFFT-FFT, è utile chiarire il suo funzionamento.Il canale di comunicazione mobile soffre della propagazione dovuta a multipath, tempo varianza (movimento dei veicoli e del ricevitore), variazioni dell'ambiente e rumore.

---

[1] Il delay spread fornisce una misura della dispersione temporale del segnale. In letteratura è usato per dare una corretta indicazione del massimo data-rate supportato dal canale, quando non vengono prese in considerazione equalizzazione e tecniche di diversità.

[2] $\rho = M (T_s B)^{-1}$, con M pari al numero di sottoportanti, B uguale alla banda disponibile e Ts tempo di simbolo

La funzione di trasferimento del canale $H(f,t)$ può essere modellata con un modello di canale selettivo in frequenza ed in tempo. Il massimo delay spread dovrebbe essere più piccolo del tempo di simbolo mentre il tempo di coerenza (l'inverso della massima frequenza doppler) dovrebbe essere maggiore del tempo di simbolo. L'OFDM fornisce un buona proposta per combattere la selettività in frequenza del canale.

Consideriamo una banda di $B\,[Hz]$ ed un data-rate di $R\,[bit/s]$. Sia t_s il periodo di informazione di simbolo, f_d la massima frequenza doppler, e τ il massimo delay spread del canale. Il principio di OFDM si basa sul suddividere la banda B in N sotto-portanti, spaziate di $B/N\ Hz$. In ogni sottoportante, l'informazione, multiplata, di rate $R/N\ bit/s$ è modulata con portanti ortogonali. Lo spettro delle differenti sottoportanti si sovrappone reciprocamente, sfruttando l'ortogonalità, a differenza di quanto accade in un sistema FDM (fig.2.1) in cui le sottoportanti sono rigorosamente non sovrapposte, dando una efficienza ottima nell'occupazione della banda.

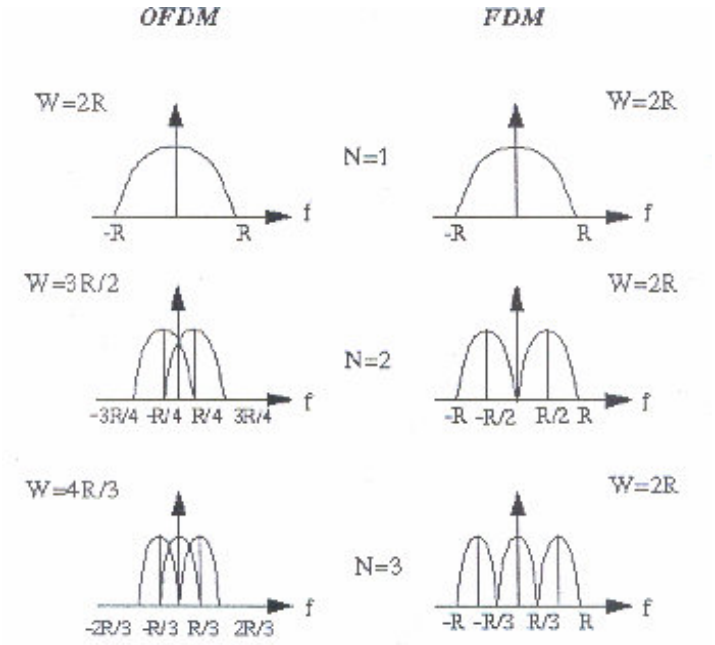


Fig.2.1-OFDM vs FDM

Durante la propagazione, a causa del multipath del canale, la condizione di perfetta ortogonalità tra le sottoportanti non è mantenuta a lungo, di conseguenza deve essere inserito un intervallo di guardia per ogni simbolo prima della trasmissione, in modo da contrastare l'interferenza intersimbolica (ISI). Se l'intervallo di guardia è più grande del delay spread, e se il tempo di simbolo OFDM che è pari a $T_s = N * t_s$, su ogni sottoportante è più piccolo del tempo di coerenza $1/f_d$, allora $H(f,t)$ sarà costante nel tempo e nella frequenza per un singolo simbolo: il canale apparirà come un canale con fading piatto in tempo ed in frequenza.
Un primo schema rappresentativo di una modulazione di questo tipo compare nella fig.2 2:

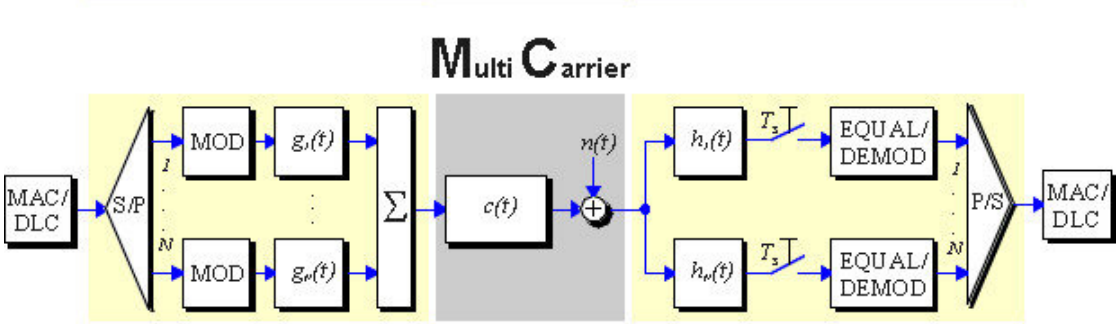


Fig.2.2 – Multi Carrier

Come spiegato precedentemente, la trasmissione dei dati avviene in parallelo ripartendo il contenuto informativo del segnale seriale a larga banda in pacchetti di simboli, appartenenti alle singole sottobande multiplate di frequenza, dove l'ampiezza associata ad ogni simbolo modula una portante differente. Il passaggio da una trasmissione serie ad una in parallelo permette di ridurre il problema dell'equalizzazione (che nel caso di trasmissione a singola portante risultava costosa per valori elevati di data-rates che causavano durate di simbolo brevi e ritardi di canale lunghi), dal momento che ogni sottocanale diventa non selettivo, sotto l'ipotesi di sotto-banda sufficientemente stretta, e risulta, pertanto, affetto da una attenuazione ed una rotazione di fase *costante*.
La durata di simbolo più lunga di un fattore N nel caso Multi Carrier (dove N è il

numero di sottoportanti) consente di far fronte all'interferenza intersimbolica in modo più semplice.

Nel modulare un flusso di dati parallelo con N sottocanali, ogni canale è modulato con un singolo simbolo separatamente. Le prestazioni sono direttamente relazionate al numero di sottobande che è possibile allocare. L'utilizzo della modulazione con portanti ortogonali, idealmente, evita il problema dell'interferenza tra le sottoportanti, consentendo la sovrapposizione parziale dei sottocanali senza pregiudicare la loro ricostruzione.

L'ortogonalità è ottenuta scegliendo delle opportune forme di impulsi di dati che sagomano il pacchetto di simboli in TX ed in RX.

Dal punto di vista di mo-demodulazione è possibile ricostruire i simboli senza interferenza tra loro. L'interferenza tra simboli di uno stesso pacchetto e di pacchetti differenti, dovuta alla presenza di echi può essere controllata introducendo l'intervallo di guardia temporale (fig.2.3), citato prima, in cui viene ripetuta una parte del pacchetto trasmesso

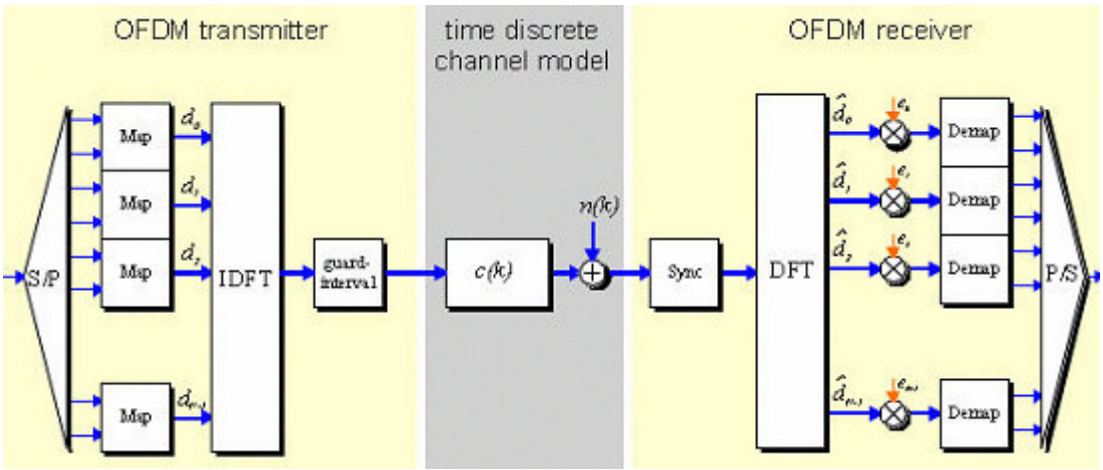


Fig.2.3 – Sistema OFDM

Questo accorgimento, a meno di una attenuazione, consente il recupero dei simboli se la durata della risposta impulsiva del canale si mantiene all'interno dell'intervallo di guardia prescelto. L'utilizzo della IFFT, mostrato in fig.2.3, dopo il blocco S/P per modulare le portanti, restituisce i campioni del segnale da trasmettere.In

ricezione vengono affrontate le operazioni inverse ovvero FFT e conversione parallelo/serie. La complessità degli algoritmi di mo-demodulazione è contenuta proprio grazie all'uso di efficienti FFT-chips.Qui sotto, viene mostrato un sistema discreto OFDM.

I blocchi funzionali del trasmettitore e ricevitore digitale OFDM sono illustrati in Fig.2.3. La sequenza di dati modulanti in ingresso viene segmentata in blocchi di dimensione M dal convertitore serie-parallelo (S/P) per poi calcolare le IFFT (IDFT). Gli ultimi L punti della IFFT (IDFT) vengono aggiunti in testa al vettore formando il cosiddetto *prefisso ciclico*. Il convertitore parallelo serie mette in sequenza i dati delle successive IFFT (IDFT) ed i dati vengono trasmessi digitalmente con una modulazione lineare ad un rate circa pari a B.

Al ricevitore i dati vengono demodulati linearmente ed i campioni x(n) vengono convertiti da serie a parallelo, il prefisso ciclico viene scartato e la FFT (DFT) dei dati restituisce la sequenza di simboli moltiplicata per la funzione di trasferimento del canale sulla corrispondente sottoportante.

2.2 STRATEGIE DI ACCESSO MULTIPLO

2.2.1 FDMA

L'avvento della modulazione di radiofrequenza ha reso possibile la coesistenza di diverse trasmissioni radio nel tempo e nello spazio senza che queste interferiscano tra loro, usando frequenze portanti diverse. *Frequency-Division Multi-plexing o Frequency-Division Multiple Access* (FDMA) è un sistema di accesso multiplo utilizzato soprattutto nei sistemi analogici di comunicazione cellulare. La banda allocata dal sistema viene divisa in un certo numero di sottobande ed ogni sottobanda è assegnata ad un singolo utente, come mostrato in fig.2.4. Un canale corrisponde ad una sottobanda. Poiché ogni utente usa una banda differente, tutti gli utenti possono trasmettere contemporaneamente.

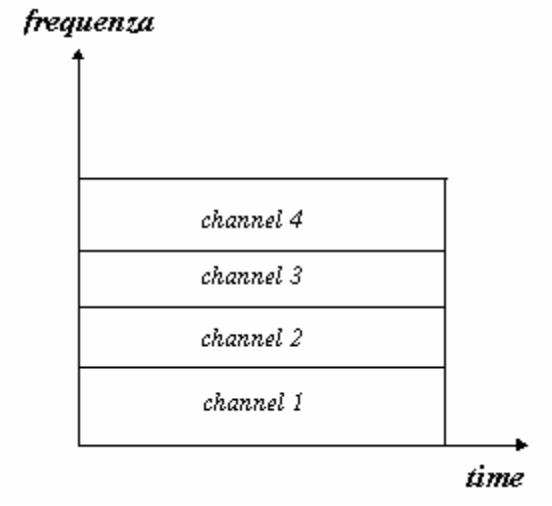


Fig.2.4 - FDMA

Un filtraggio di tipo passabanda o eterodina abilita la distinzione di ogni sottocanale mediante demodulazione, unico inconveniente è dato dalla generazione di prodotti di intermodulazione, originati da non linearità, tra le portanti contemporaneamente presenti.

2.2.2 TDMA

Un secondo tipo di accesso multiplo è l'accesso TDM *(Time-Division Multiplexing) o*, TDMA (Time Division Multiple Access). Le risorse in questo caso vengono assegnate su base temporale:

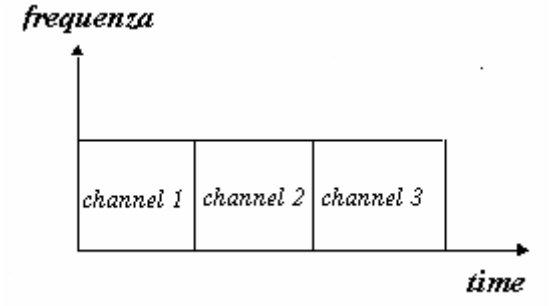


Fig.2.5 - TDMA

Nel *Time-Division Multiplexing*, ogni utente accede al sistema in un intervallo temporale limitato (slot) ed è abilitato a farlo periodicamente. L'insieme di slot che

si susseguono in tale periodo viene chiamato trama. TDMA lavora dividendo una frequenza radio in TS e allocando su questi più chiamate. Un canale consiste quindi in uno slot temporale. Gli utenti trasmettono ad alta velocità (circa N*fb, dove fb è il ritmo binario d'utente ed N il numero di slot nella trama) pacchetti realizzati mediante compressione temporale di flussi numerici provenienti da sorgenti (ad esempio la fonia numerica) a ritmo di emissione costante. La banda occupata è proporzionale quindi a N*fb. Il TDMA può usare un singolo sistema ricevente per tutti i canali che utilizzano la stessa portante. E' un sistema di accesso molto adatto ad essere gestito in modo dinamico con un accesso su richiesta per trasmissioni a pacchetto. L'inconveniente è che richiede requisiti stringenti di sincronizzazione con conseguente perdita di capacità per overhead destinati al monitoraggio del sistema, cosa che invece non accade con la tecnica FDMA dove le trasmissioni non vengono assolutamente coordinate nel tempo in quanto non è necessaria alcuna sincronizzazione tra gli utenti.
L'effetto di un canale o ricevitore non ideali possono richiedere l'inserimento di intervalli di guardia nel TDMA e bande di guardia nel FDMA per evitare interferenza co-canale.

2.2.3 CDMA

L'ultima tecnica di accesso multiplo presentata è la tecnica CDMA (Code Division Multiple Access) che consente la contemporanea trasmissione degli utenti sull'intera banda mediante l'utilizzo di opportuni codici.

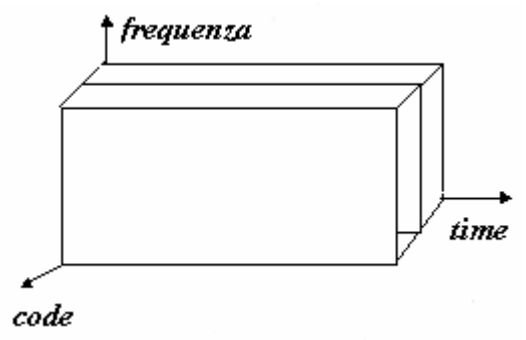


Fig.2.6 - CDMA

I flussi informativi lato ricevitore risultano separabili mediante l'assegnazione ad ogni utente di un differente codice. Tale tecnica si basa su trasmissioni di tipo Spread Spectrum (a Spettro Espanso) del tipo DS-CDMA (Direct Sequence-CDMA), in cui è utilizzata una unica portante per tutti gli utenti. Ad ogni utente viene associato un particolare codice che consente, in sede di ricezione, di estrarre il segnale utile dai segnali degli altri utenti , nonostante siano sovrapposti in banda. Un canale quindi consiste in un codice. La separazione del segnale utile dall'interferenza da accesso multiplo è possibile grazie a proprietà di ortogonalità di cui godono i codici.
Descritte le diverse tecniche di accesso multiplo, si può ora passare a vedere come queste possano affiancare la tecnica di modulazione OFDM.
Le risorse che debbono essere distribuite possono essere rappresentate all'interno di una griglia Banda- Tempo e a seconda della tecnica utilizzata, le risorse, ovvero i tasselli assegnati, saranno disposti in modo diverso. Il simbolo OFDM fissa la durata dello slot temporale, M sono i sottocanali disponibili in frequenza durante ogni T_s.

2.2.4 OFDM-TDMA

In OFDM-TDMA più simboli OFDM (slots) sono raggruppati in frames ed i frames sono riuniti in superframes. Gli utenti possono essere multiplati in tempo tramite un accesso di tipo random oppure tramite un controllo centralizzato che assegna dinamicamente gli slots. In tal caso, un certo numero di slots sono destinati alla trasmissione broadcast di informazione di controllo, da parte della BS. L'assegnazione dinamica degli slots consente di multiplare statisticamente il canale, e come in un tipico sistema TDMA, l'interferenza multiutente è limitata all'interferenza dovuta a celle adiacenti che riutilizzano la stessa banda trasmissiva. Base Stations dotate di antenne multiple possono notevolmente mitigare questi effetti tramite algoritmi di *beamforming* o più sofisticate tecniche di *space-time coding*.[Rohling H., Grunheid R,1996]. In ogni slot la banda può essere usata

efficientemente perché tutte le sottoportanti sono sincrone tra loro. Tuttavia, interferenza intersimbolica, sotto forma di interferenza tra portanti adiacenti (ICI), è presente a causa della non perfetta sincronizzazione di portante del ricevitore, del rumore di fase degli oscillatori in trasmissione e ricezione e delle variazioni della risposta del canale in generale.

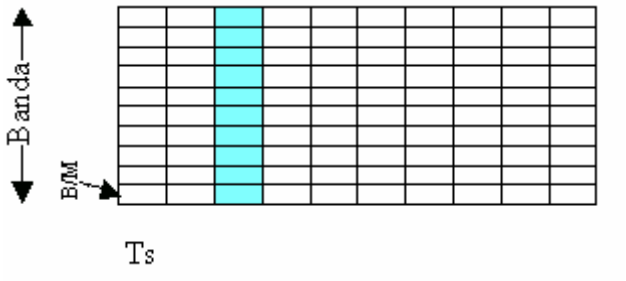


Fig.2.7 - Suddivisione del canale radio nel OFDM-TDMA

Come mostrato in figura 7 un accesso OFDM-TDMA è caratterizzato dalla peculiarità che ciascun utente, fissato lo slot dedicatogli, ha a sua disposizione tutta la banda.

2.2.5 OFDM-FDMA

Lo schema OFDM-FDMA è il duale dello schema OFDM-TDMA, in quanto all'utente viene assegnata una sottobanda per una durata che comprende più simboli OFDM.

La durata dello slot si allunga corrispondentemente e più utenti possono contemporaneamente usare le risorse essendo suddivisi su diverse sottoportanti.

A causa dell' impossibilità di garantire perfetto sincronismo di portante tra gli utenti, le sottoportanti adiacenti non sono esattamente ortogonali e ciò provoca l'insorgere di interferenza multiutente (MUI). Pertanto OFDM-FDMA richiede l'uso di guardie in frequenza (non tutte le sottoportanti possono essere utilizzate) e richiede tecniche di controllo di potenza, poiché l'interferenza dovuta alla non esatta sincronizzazione di portante è interferenza multiutente: utenti il cui segnale sia ricevuto a livelli di potenza relativamente elevati possono mascherare utenti i cui

segnali sono deboli in frequenza (l'effetto *near far*), con esiti ben peggiori della ICI presente nei sistemi OFDM-TDMA.

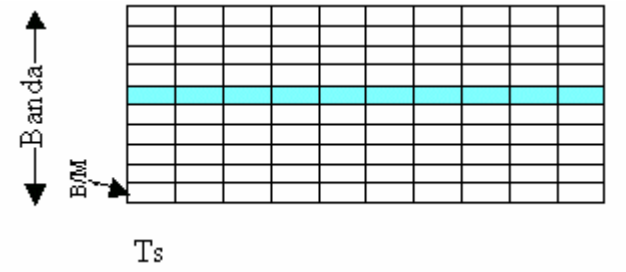


Fig.2.8 - Suddivisione del canale radio nel OFDM-FDMA

Come mostrato in figura 2.8, un accesso OFDM-FDMA è caratterizzato dalla peculiarità che ciascun utente, fissata la banda a lui dedicata, ha a sua disposizione più simboli OFDM.
Una variante del OFDM-FDMA è l'OFDMA. Questa tecnica di accesso multiplo assegna le varie sottoportanti agli utenti come in un sistema FDMA, con la differenza (sostanziale in temini di costo implementativo) che la mo/demodulazione è effettuata digitalmente.

OFDMA realmente utilizza tutti i tasselli della griglia come sottocanali separati ed ha la massima flessibilità nell'assegnazione delle risorse. Sfortunatamente come OFDM-FDMA non può essere efficiente. Controllo di potenza ed intervalli di guardia tra utenti sono richiesti per mitigare l'effetto della MUI.

2.2.6 OFDM-CDMA

Per mitigare l'effetto dell'interferenza multiutente e dell'ICI lo schema OFDM può essere combinato con il CDMA. In OFDM-CDMA, anche detto Multi-Carrier CDMA (MC-CDMA) le sottoportanti possono essere ritenute equivalenti ai chips di un sistema spread spectrum. Lo schema di base opera come segue: la sequenza di simboli da trasmettere viene espansa moltiplicando ciascun simbolo s(n) per un

codice **c**=(c_0,…, c_{K-1}), che tipicamente consiste in una sequenza binaria pseudo-random di lunghezza K<=M (M: numero di punti della FFT):

x(n)= **c**s(n)

In pratica è come se trasmettessimo in parallelo lo stesso simbolo sulle K sottoportanti pesando ognuna con il relativo chip del codice.

La sequenza trasmessa è la successione dei vettori di dati **x**(n), cioè (….., x_0(n-2),…, x_{K-1}(n-1), x_0(n),…, x_{K+1}(n), x_0(n+1),…). La sequenza da trasmettere ha un rate (chip rate) che è K volte maggiore di quello di simbolo. Un insieme di utenti diversi trasmette sullo stesso gruppo di sottoportanti contemporaneamente usando parole di codice diverse, ortogonali ai codici utilizzati da tutti gli altri utenti. Nulla vieta, ovviamente, l'uso di più codici per un singolo utente su uguali o diverse sottobande. Questa tecnica offre rispetto alle precedenti uno *spreading gain* del codice, che mitiga l'effetto sia della MUI che della ICI [Yee N., Linnartz,1993]. Usando codici a bassa correlazione più utenti simultaneamente possono usare le stesse sottoportanti ed essere separati grazie a questi, traendo vantaggio anche del guadagno dovuto alla diversità di frequenza. Se il canale fosse ideale sarebbe possibile ricostruire esattamente il segnale trasmesso da ciascun utente semplicemente moltiplicando ciascuna sottoportante per il rispettivo elemento di codice in questione e sommando i termini risultanti. La perfetta sincronizzazione di portante non può essere imposta in caso di utenti multipli, pertanto le bande assegnate a utenti diversi dovranno essere opportunamente separate, per mitigare l'interferenza multiutente (MUI). D'altro canto bande adiacenti assegnate allo stesso utente (nel caso estremo, tutte le sottoportanti di uno slot) possono essere utilizzate più efficientemente e/o risentono di minore interferenza. Questo aspetto va tenuto in considerazione nella definizione della strategia di suddivisione delle risorse e nel progetto del MAC.

Il livello di ortogonalità è doppio:

1. le sottoportanti sono ortogonali;
2. i codici di espansione sono ortogonali.

Questa è la ragione per cui questa tecnica è robusta al fading in frequenza.

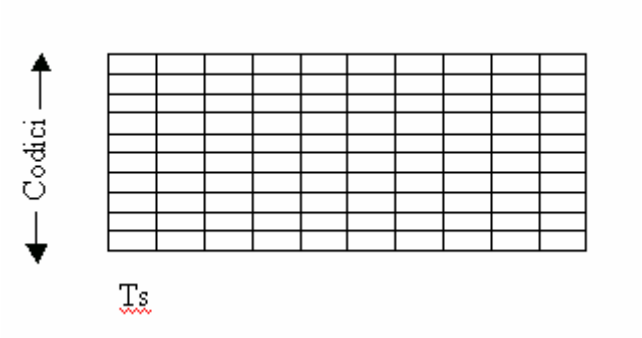


Fig.2.9 – Suddivisione del canale radio per OFDM-CDMA

A differenza delle altre due rappresentazioni, l'OFDM-CDMA può essere schematizzato come una griglia codici-tempo (fig.2.9).

2.2.7 Cenni a sistemi alternativi: AMOUR

Una nuova classe di codici sono stati proposti recentemente per annullare completamente l'interferenza multiutente (MUI) nei sistemi di tipo CDMA. Questi codici, detti *codici Lagrange Vandermonde(LV),* offrono una eliminazione deterministica della MUI senza necessità di stima del canale, e convertono canali selettivi in frequenza in canali di tipo "*flat-fading*" (con fading piatto).
Una delle maggiori limitazioni dei sistemi CDMA a larga banda, come già più volte detto, è l'interferenza multiutente che si verifica in presenza di propagazione da multipath e affligge in particolar modo la tratta di uplink.
Su questa tratta infatti, codici che sono ortogonali in partenza, non lo sono più al RN. I segnali provenienti da tutti i RT vengono moltiplicati tutti per la stessa funzione di trasferimento di canale e sommati al ricevitore. Una operazione di equalizzazione non è sufficiente a recuperare il segnale di ognuno, se non quello degradato dagli altri utenti. L'ortogonalità dei codici non può più essere utilizzata. Per distinguere gli utenti è necessario separarli in banda.

Il modello di trasmissione è caratterizzato da sottocampionatori e sovracampionatori che hanno lo scopo di fare spreading e despreading per un fattore pari a P. Ogni utente allarga la sua sequenza di informazione $s_m(n)$ tramite il sovracampionatore e la codifica usando un codice $c_m(n)$ di lunghezza P, prima di trasmetterla attraverso il j-esimo canale "non conosciuto" $h_m(j)$ che, oltre al multipath, aggiunge l'impulso (transmit spectral-shaping) e l'm-esimo asincronismo d'utente nella forma di fattori di ritardo. I dati multiplati $v(n)$ sono filtrati e campionati al chip rate. Il passaggio attraverso il filtro di ricezione $g_m(k)$ è seguito dal despreading (sottocampionamento) di un fattore P e dalla decisione per ottenere la sequenza stimata $\hat{s}_m(n)$. Si assume che:

a. $P - M \geq L \quad e \quad M > L$, dove L rappresenta il massimo numero di canali $\{h_m(j)\}_{m=1}^{M}$ tali da ottenere ISI al rate di chip ma non al rate di simbolo;
b. i codici hanno un numero di bit di prefisso ridondante (prefisso ciclico) pari a $L_g \geq L$.

Basandosi sui punti a e b, i campioni ricevuti sono:

$$x(n) = \sum_{m=1}^{M} \sum_{i=-\infty}^{+\infty} s_m(i) \cdot \sum_{j=0}^{L} h_m(j) \cdot c_m(n - j - iP) + v(n) \quad \textit{o equivalentemente}$$

$$x(nP+k) = \sum_{m=1}^{M} \sum_{i=-\infty}^{+\infty} s_m(i) \cdot \sum_{j=0}^{L} h_m(j) \cdot c_m((n-i)P + k - j) + v(nP+k)$$

Dal momento che $c_m(i) = 0$ per i<0 e i>P-1, il solo termine diverso da zero nella somma relativa ad i corrisponde all'indice i=n, e quindi si ha:

$$x(nP+k) = \sum_{m=1}^{M} s_m(n) \cdot \sum_{j=0}^{L} h_m(j) \cdot c_m(k - j) + v(nP+k)$$

Il problema risolto dal sistema AMOUR si è basato sul progetto di codificatori e decodificatori in grado di eliminare l'interferenza multiutente in maniera completa [Giannakis G.B. et altri,1999-2000], consentendo una assegnazione della risorsa all'interno della griglia tempo- frequenza in cui uno slot e l'intera banda può essere condivisa da più utenti facendo uso dei codici ora menzionati .

2.3 OFDM combinato con strategie di accesso multiplo: vantaggi e svantaggi

La relativa semplicità ed il controllo dell'interferenza (MAI) garantiti dall'OFDM-TDMA rappresentano il maggiore vantaggio di questa soluzione. Un'accurata progettazione di trasmettitore e ricevitore, infatti, può mitigare gli effetti dell'ICI. Il maggiore problema di questa soluzione è la rigidità della suddivisione delle risorse: il MAC di fatto non utilizza l'OFDM per aumentare la granularità delle risorse. Pertanto, non considerando i vantaggi dell'OFDM come tecnica di modulazione in se, l'OFDM potrebbe essere sostituito da un sistema single carrier con rate di simbolo pari ad M/Ts, rimanendo il MAC inalterato. Dunque l'OFDM-TDMA è sostanzialmente un sistema TDMA con una specifica tecnica di modulazione e ciò chiaramente fa comprendere che ci sono elementi di OFDM di cui l'OFDMA-TDMA non beneficia e che il MAC può sfruttare meglio per aumentare il throughput e la qualità di servizio.

Simili conclusioni possono essere tratte a proposito dell'OFDM-FDMA. In aggiunta, l'OFDM-FDMA con un'allocazione statica e non ottimizzata delle sottobande dedicate agli utenti è affetto da fading come un sistema di comunicazione a banda stretta, a meno che non vengano risolti adattativamente complessi problemi di ottimizzazione nell'allocare le sottobande. Una contromisura per il fading può essere l'uso dell'OFDM-FDMA in combinazione con tecniche di Frequency Hopping. L'OFDM-FDMA è affetto da ICI ma anche da interferenza multiutente (MAI), dovuta all'asincronismo di portante tra i vari utenti.

L'OFDMA è certamente la soluzione più efficiente e flessibile, perché sfrutta ogni tassello della griglia, come una risorsa assegnabile per trasmettere informazione, lasciando la massima flessibilità al MAC. Il MAC può combinare o dividere le risorse in modo da soddisfare al meglio le necessità dell'utenza. Due sono i problemi di OFDMA:

1) Nello scenario ideale di perfetto sincronismo tra utenti, il fatto che ogni WTIU (Wireless Terminal Interface Units)[3] abbia una diversa funzione di trasferimento

[3] WTIU sono i terminali fissi e mobili con interfaccia radio presenti nel sistema considerato

del canale fa si che ciascun sottocanale ha un diverso rapporto segnale a rumore a seconda dell'utente. Pertanto, il problema di suddividere le risorse in modo da soddisfare le richieste delle WITU è un complicato problema di ottimizzazione non lineare. Tuttavia l'uso di tecniche di Frequency Hopping può parzialmente compensare questo inconveniente.

2) L'OFDMA è afflitto da MAI ed ICI, per combattere le quali occorre fare affidamento su una complicazione della struttura del ricevitore notevole e che è in contrasto con la semplificazione del ricevitore a cui dovrebbe portare la scelta dell'OFDM come tecnica di modulazione e multiplazione. Tuttavia, occorre tenere presente che l'uso di antenne multiple e beamforming potrebbe considerevolmente ridurre la MUI.

Rimanendo nello spirito OFDMA ed assumendo che lo strato fisico non sia in grado di separare altrimenti i segnali provenienti da più utenti, la soluzione auspicabile è di non utilizzare alcuni sottocanali, ponendo delle guardie tra utenti. L'interrogativo che ovviamente sorge è se invece di "spegnere" tali sottocanali non sia più opportuno usarli come fonte di diversità, come viene fatto nell'OFDM-CDMA.

Il sistema OFDM-CDMA ha due svantaggi rispetto ai precedenti due:

i) richiede una complicazione dello strato fisico (lo spreading dei dati);
ii) se il sistema viene usato senza sovrapporre gli utenti nelle sottobande il sistema è inefficiente; viceversa, se si sovrappongono gli utenti sulle varie sottobande usando codici diversi e non si usano tecniche di joint detection o decorrelation receivers, l'interferenza degrada le prestazioni e richiede tecniche di power control, difficili da far convergere in caso di comunicazioni a pacchetto.

L'OFDM-CDMA è anche affetto da MUI ed ICI, anche se, a parità di condizioni, ne risente in misura minore di entrambi i metodi citati sopra, il che significa che la durata T_s può essere incrementata maggiormente, aumentando il numero si sottoportanti disponibili.

Altro vantaggio dell'OFDM-CDMA è che può chiaramente degenerare nei due casi precedenti, pertanto può essere usato in modalità che sono una soluzione di

compromesso tra OFDM-TDMA e OFDMA. I nuovi gradi di libertà che OFDM-CDMA aggiunge, possono essere sfruttati per raggiungere buoni compromessi tra l'efficienza e la robustezza della tecnica trasmissiva che, accoppiate ad una efficace strategia MAC, determinano la capacità del sistema di fornire differenti parametri di qualità di servizio.

2.3.1 Confronto degli schemi di accesso

Un rapido confronto tra questi schemi deve essere fatto in termini di flessibilità e overhead di segnalazione, aspetto di notevole importanza nel dimensionamento della velocità di trasmissione e del ritardo.
Un efficiente schema di accesso dovrebbe garantire una elevata flessibilità nella distribuzione della risorsa disponibile tempo-banda tra tutti gli utenti. Da un lato deve essere considerato il comportamento selettivo in frequenza del canale e dall'altro le richieste degli utenti per data rate differenti (dipendenti dal tipo di servizio richiesto).
I tre schemi mostrano all'incirca lo stesso grado di flessibilità. Il numero delle sottoportanti per utente (OFDM-FDMA), o il numero di time slots (OFDM-TDMA) per utente, o il numero di codici per utente (OFDM-CDMA) possono essere adottati in accordo alle richieste correnti. L'ultima tecnica presuppone che tutti i codici siano conosciuti ad ogni utente.
Dal momento che l'OFDM-CDMA sfrutta la diversità in frequenza per mezzo di varie sottoportanti per ogni simbolo, l'adattamento della sottoportante alla funzione di trasferimento del canale (equalizzazione delle sottoportanti) è possibile e necessaria. Questa è la differenza di questo approccio rispetto agli altri due schemi di accesso.
Adottando una tecnica di tipo OFDM-TDMA l'accesso multiplo è garantito suddividendo l'asse temporale in trame e time-slots (TS). Ad ogni utente viene assegnato un TS a cadenza di trama e quando è il suo turno, l'utente può trasmettere i propri pacchetti sull'intera banda disponibile (su tutte le sottoportanti), all'interno

di una singola trama TDMA, che copre alcuni simboli OFDM. Il numero di simboli OFDM per trama può essere variato, in accordo alle richieste di ogni utente. Le risorse vengono così assegnate in maniera dinamica sulla base delle esigenze di ogni singolo utente ottenendo uno sfruttamento omogeneo della banda da parte di tutti gli utenti, anche se l'impiego della risorsa nel tempo potrebbe non essere ottimale, se il canale è un canale rapidamente variante nel tempo.

Nell'OFDM-FDMA tutti gli utenti accedono contemporaneamente al sistema, ma ogni utente usa un sottoinsieme di sottoportanti che gli viene assegnato in modo esclusivo. Un'allocazione efficiente deve garantire che ogni utente venga servito al meglio. La tecnica OFDM-FDMA può sfruttare il fatto che ogni utente prova un differente canale radio allocando soltanto sottoportanti "buone" con elevato rapporto segnale/rumore e aumentando con ciò la capacità di canale. Questo tipo di tecnica è però anche quella che richiede più overhead di segnalazione poiché deve essere trasmesso il numero di utenti per ogni sottoportante (vedi Tabella I).

Passiamo ora all'ultimo schema di accesso (OFDM-CDMA).

La struttura rappresentativa, allo strato MAC, di questo schema di accesso è la matrice dei codici. La disponibilità dei codici fornisce un grado di libertà in più nell' accesso radio.

Nell'OFDM-CDMA, tutti gli utenti condividono sempre tutta la banda a disposizione (tutte le sottoportanti) usando differenti codici ortogonali. In tal modo vengono sfruttate al massimo le caratteristiche di diversità in frequenza, le proprietà Spread Spectrum e le eventuali variazioni nel tempo del canale. Con tale metodo di accesso viene assegnato un numero di parole di codice tale da poter garantire la risorsa richiesta. Periodicamente queste parole di codice dovranno essere riallocate per far fronte alla variabilità di traffico nel tempo. Un intervallo di tempo troppo elevato tra due successive riallocazioni, potrebbe comportare, infatti, un ritardo eccessivo nella consegna della risorsa richiesta da un utente, degradando la qualità del servizio. L'overhead di segnalazione dell'OFDM-CDMA è paragonabile a quello del TDMA. La tecnica CDMA è una buona candidata per supportare servizi multimediali, soprattutto in comunicazioni radiomobili, grazie alla sua capacità di

adattarsi alla natura asincrona del traffico dati multimediale, per fornire una capacità di trasferimento informativo maggiore di quella offerta da TDMA e FDMA. Se l'allocazione delle risorse è gestita dalla base-station, l'informazione circa l'allocazione sottoportante/codice deve essere trasmessa sia per il downlink che per l'uplink.

Inoltre l'OFDM-CDMA permette di mediare statisticamente l'interferenza tra più utenti che utilizzano le stesse sottoportanti con codici diversi;usando codici di spreading di lunghezze diverse, permette di avere una MAC-PDU con un numero fisso di chips che però supporta un numero diverso di dati informativi [Rohling et altri, 1997].

Access scheme	Kind of signalling information	Overhead (bit)
FDMA	Allocation table for all subcarriers	$K \cdot \log_2 U$
TDMA	Start/end of time slot	$2U \cdot \log_2 F$
CDMA	Start/end of code index	$2U \cdot \log_2 K$

Tabella I - Informazione di segnalazione richiesta per i diversi schemi di accesso; U = numero di utenti, K = numero di sottoportanti, F = numero di simboli OFDM per frame (up- e downlink).

Dall'analisi ovvero dal confronto di tali sistemi si nota come la tecnica MC-CDMA sia efficiente solo quando siamo in grado di garantire in ricezione l'ortogonalità (o almeno, la bassa correlazione) tra i codici, cosa che nella realizzazione di una tratta Downlink, che è la tratta da noi sviluppata, data la particolarità che ogni utente riceve su un canale separato è senz'altro assicurata.

CAPITOLO 3

DESCRIZIONE DEL SISTEMA

3.1 SCENARIO DI RIFERIMENTO

Come visto nei due capitoli precedenti, il sistema preso in considerazione è una Wireless-LAN di tipo indoor che sfrutta come tecnica di modulazione l'ibrido tra la modulazione OFDM e la tecnica di accesso CDMA. Per lo scopo dei nostri studi abbiamo ipotizzato, in prima approssimazione, che la nostra WLAN fosse assimilabile ad un ufficio o ad una vasta struttura ad un solo piano. Studi futuri, nell'ambito di questo progetto terranno conto di problematiche anche dovute alla gestione dell'handover all'interno di un edificio, anziché di un ufficio. La configurazione di rete che si assume è quella anche utilizzata nell'ambito delle reti cellulari. Si suppone infatti che l'area analizzata sia condivisa da una molteplicità di celle all'interno delle quali si trovano numerosi terminali fissi e mobili, i quali sono sotto il controllo di una unica stazione radio base, il Radio Node. Lo scenario di riferimento è proprio la cella all'interno della quale si suppone che detti terminali colloquino tra di loro attraverso il RN, che risulta essere la struttura cardine di un protocollo di accesso al mezzo di tipo centralizzato quale è quello sotto studio. La struttura del sistema schematizzabile come mostrato in figura rientra nella classe delle WLAN con topologia a stella.

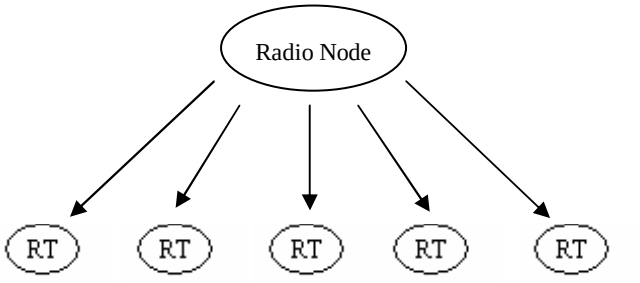


Fig.3.1 – Struttura di una cella in una WLAN con topologia a stella

All’interno di questa classe, è possibile individuare due canali logici per la comunicazione delle varie entità: il canale di downlink (DL) attraverso cui il Radio Node (RN) dialoga con i vari Radio Terminal (RT) e il canale di Uplink (UL) attraverso cui, in modo condiviso, i vari RT dialogano con il RN. Questa tesi si riferisce solamente al downlink, come d’altra parte anche il simulatore progettato per la realizzazione del protocollo MAC.

Lo scenario di assegnazione dinamica delle risorse stabilisce che un codice, tra i K disponibili per un accesso contemporaneo al canale di DL, può non essere assegnato per tutto il tempo ad un singolo RT, ma può essergli dedicato solamente per consentirgli la trasmissione di un certo numero di MAC PDU.

In altre parole, si può pensare l’asse dei tempi suddiviso in unità temporali (*Time Slot*) la cui durata rappresenta il tempo necessario alla trasmissione di una singola MAC PDU e sarà dimensionata nel paragrafo relativo allo strato MAC.

Gli *N* intervalli temporali così definiti sono stato raggruppati in un periodo di *Trama* che si ripete in modo periodico. Questo risulta utile nella gestione delle unità informative della classe di traffico che necessita di un’alta trasparenza temporale (classe GB) in quanto è possibile cadenzare le opportunità trasmissive, adattandosi alla natura periodica delle emissioni da parte dei trasmettitori.

Inoltre, data la disponibilità di K *codici,* c'è un'ulteriore grado di libertà nell'accesso radio, che di fatto da luogo ad una struttura trasmissiva di tipo *“ a matrice”* (detta appunto *"matrice dei codici")* per il trasferimento di informazioni in DL, come mostrato in Figura 3.2, dove K è il numero di codici utilizzabili nella trasmissione di unità informative, mentre N è il numero di TS componenti la trama.

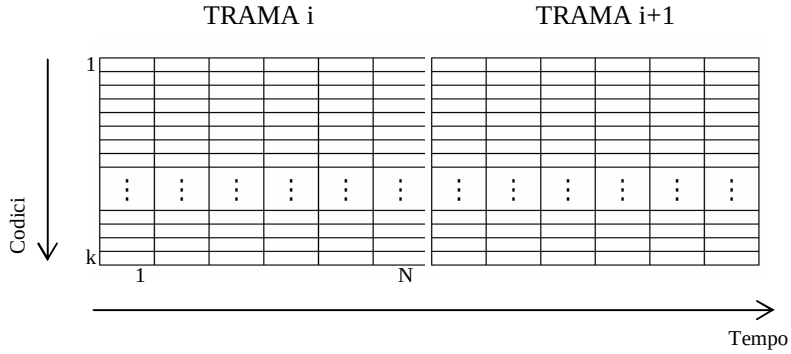


Fig.3.2 – Struttura trasmissiva del sistema

La scelta della lunghezza N della trama dipende da fattori tra loro contrastanti. Da un lato, si spinge verso lunghezze di trama piccole per l'esigenza di minimizzare il ritardo di accesso delle unità informative ed il loro ritardo di trasferimento, specie per quei servizi che necessitano di una alta trasparenza temporale, dall'altro, lunghezze di trama grandi comportano uno spreco percentualmente minore di banda utilizzata dai canali di segnalazione del MAC, descritti sempre nel paragrafo relativo allo strato MAC.

La capacità minima assegnabile ad un singolo utente può essere anche molto piccola e decisa arbitrariamente: è infatti possibile assegnare anche una sola coppia <codice, TS>, che rappresenta la quantità minima di capacità assegnabile (1 MAC-PDU) in una trama ad un utente e poi attendere un numero qualsivoglia di trame prima di rassegnargli nuovamente una opportunità trasmissiva (compatibilmente con i ritardi necessari per soddisfare il trasferimento di informazioni d'utente).

3.2 ARCHITETTURA PROTOCOLLARE

3.2.1 Strato di Rete e strato di Adattamento

Nella definizione del nostro sistema abbiamo ipotizzato un'architettura protocollare in cui gli strati da considerare partendo dall'alto sono lo strato IP, lo strato di Adattamento, lo strato MAC e quello Fisico.

Il protocollo IP è un protocollo di strato 3 caratterizzato dalle seguenti funzionalità:

- definisce lo schema di indirizzamento (diretto o indiretto)
- stabilisce l'unità base per il trasferimento dei dati attraverso internet, indicando il formato dei dati che attraversano l'inter-rete
- definisce il percorso dell'unità dati da sorgente a destinazione
- specifica le regole cui host e router si devono attenere per il processamento delle unità informative
- definisce le modalità di segmentazione e riassemblaggio delle unità dati.

Il principale servizio offerto dall' IP è il trasferimento di unità informative.Il servizio è inaffidabile, senza sessione e basato sul paradigma Best effort.
Il sistema preso in considerazione deve essere in grado di trattare flussi di dati con vincoli sul ritardo di trasferimento ed è per questo che è stato necessario considerare il modello di servizi integrati INTSERV (Integrated Services) che è in grado di fornire sia servizi a qualità garantita sia servizi BE. Tale modello si basa sul protocollo di riservazione delle risorse RSVP, trattato nel capitolo successivo insieme alla definizione dei servizi con qualità garantita e best effort.
Il protocollo IP tratta ciascuna unità informativa come un messaggio indipendente da tutti gli altri; non esistono pertanto, in questo strato i concetti di connessione e circuito logico; il protocollo IP è senza connessione. Il trasferimento dei dati può richiedere una loro segmentazione laddove le dimensioni delle unità informative gestite dalle sottoreti siano inferiori alle dimensioni massime consentite alle stesse unità informative del TCP-IP. A tale scopo l'IP fornisce un meccanismo specifico per la segmentazione ed il riassemblaggio dei propri dati. La scelta della dimensione del pacchetto IP non è casuale. Di seguito si capirà il perché le sotto-reti componenti internet possono avere diverse limitazioni circa la massima lunghezza delle loro unità dati che ad esempio in una LAN Ethernet è di 1500 bytes.
La dimensione massima dell'unità dati di una sottorete è denominata in TCP-IP, Maximum Transfer Unit (MTU). Dovendo scegliere la dimensione di un datagramma IP, una possibile soluzione potrebbe essere quella di adottare un valore pari al minimo delle MTU delle sottoreti da attraversare. Ciò richiederebbe però uno scambio di informazioni di controllo per determinare tale valore minimo e, causerebbe delle inefficienze nel trasporto attraverso sottoreti con dimensioni di MTU maggiori del valore minimo.
Come per altre problematiche, si è scelta una soluzione che sia la più semplice possibile e che non sia legata a particolari tecnologie delle sottoreti componenti internet. Ogni host che emette un datagramma IP può scegliere una qualsiasi dimensione per il datagramma stesso, purchè inferiore alla dimensione massima di un datagramma pari a 65536 ottetti e superiore a quella dell'intestazione, pari a 40

bytes. Tipicamente la dimensione di un datagramma viene scelta pari alla MTU della sottorete a cui è connesso il sistema mittente. Ovviamente se la quantità di dati da trasferire è inferiore alla MTU prescelta, il datagramma avrà una dimensione minore della MTU stessa. Il solo vincolo che IP pone ai sistemi connessi ad internet è che i routers debbano accettare datagrammi di dimensione pari a quelli della MTU delle sottoreti a cui sono connessi e che tutti i sistemi debbono comunque accettare e gestire datagrammi di dimensioni almeno pari a 576 ottetti.

Al di sotto dello strato di rete troviamo lo strato di adattamento.Compito di questo strato è adattare le unità informative provenienti dagli strati superiori (nel nostro caso, datagrammi IP) al tipo di servizio di trasferimento offerto dallo strato MAC.

Nello strato di adattamento usualmente vengono svolte funzionalità che sono comuni a qualunque protocollo di livello superiore che usufruisca dell'AL-servizio, come la segmentazione e la ricostruzione dei pacchetti di strato 3, nonché tecniche di protezione e recupero d'errore differenziate per classi di servizio supportate dalla rete di accesso.

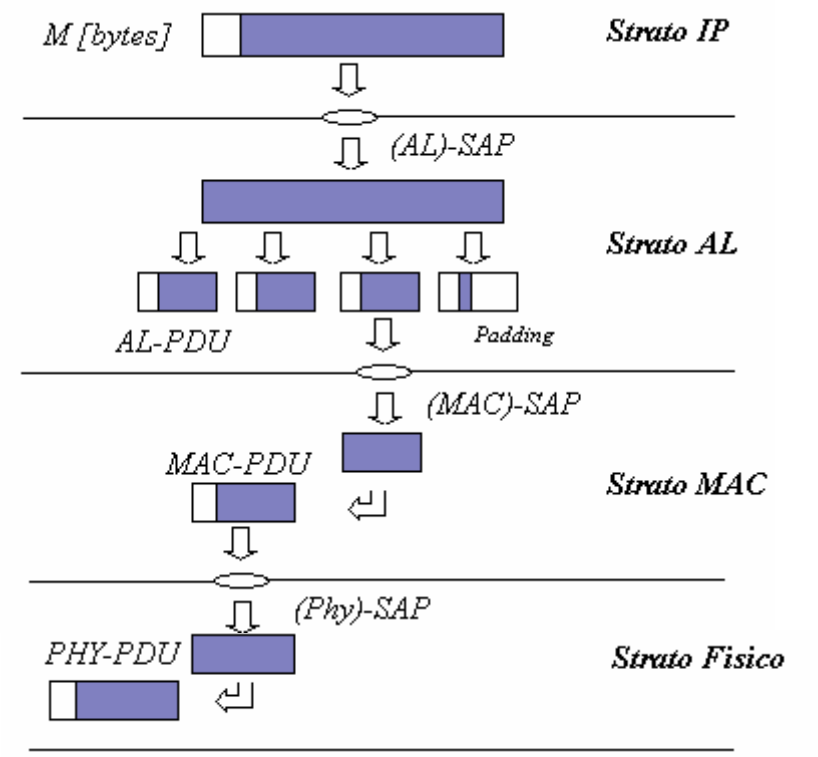


Fig.3.3 – Suddivisione del pacchetto IP attraverso gli strati inferiori

Vengono svolte inoltre funzionalità specifiche del particolare protocollo che si trova sopra ad AL, come la traduzione degli indirizzi IP in indirizzi locali del sistema (indirizzi MAC dei RT su cui sono attestati gli utenti terminali IP), e questo al fin di ridurre la quantità di informazione da trasferire sull'interfaccia radio. Un'altra funzionalità altrettanto importante è la classificazione dei pacchetti in transito, al fine di effettuare un "mapping" dei flussi di datagrammi IP che supportano una propria qualità del servizio, nelle classi di traffico del MAC.

3.2.2 Strato MAC

La principale funzione che lo Strato MAC deve svolgere è quella di fornire un accesso multiplo agli utenti della WLAN, sfruttando in modo efficiente la risorsa radio ed implementando un adeguato meccanismo di allocazione dinamica delle risorse. Il MAC dovrà fornire sia un trasferimento ad alta trasparenza temporale per la classe di servizio a qualità garantita, tipica di applicazioni *time-bounded* come voce e video, sia un trasferimento ad alta integrità informativa per la classe BE, garantendo al tempo stesso equità di trattamento tra i vari terminali radio componenti il sistema. Tuttavia, il protocollo dovrà essere in grado di consentire la assegnazione e l'utilizzo da parte di un singolo RT di tutta la capacità disponibile all'interfaccia radio, permettendo quindi la massima flessibilità di accesso.
E' prevista una gestione differenziata delle classi di traffico a seconda della priorità attribuita a ciascuna.
La prima caratterizzazione del MAC si è basata sulla gestione di due sole code, una relativa alla classe GB, l'altra alla classe BE. Lo sviluppo del protocollo affrontato in un secondo tempo ha consentito l'inserimento di una seconda coda per la classe BE a priorità più bassa, con un conseguente incremento del traffico nel sistema.
Il controllo delle code di ogni classe e l'assegnazione intelligente delle MAC-PDU contenute nelle diverse code sono stati punti di partenza per la realizzazione di un algoritmo di scheduling dinamico che permettesse di raggiungere un migliore utilizzo della banda rispetto ad un primo tipo di scheduling che dedicava troppa

risorsa alla classe a qualità garantita. Tutto questo al fine di ottenere un elevato *throughput* delle unità informative di entrambe le classi considerate, sempre nel rispetto delle priorità attribuite alle varie sorgenti.
Ulteriore obiettivo del MAC è quello di preservare l'ordine dei pacchetti, tentando di garantire l'assenza di fenomeni di "fuori sequenza" soprattutto per applicazioni di tipo real time. Tale problema in questa tesi non ha ricevuto molta attenzione, dal momento che si è supposta una trasmissione ideale priva di errori per quanto concerne la ricezione dei dati e priva di errori relativamente a perdite di pacchetti: Si è supposto che tutto ciò che viene trasmesso arrivi correttamente e intatto a destinazione. Lo scopo del MAC in questione è stato quello di adattarsi ad una situazione di nascita di sorgenti reale e garantire tramite uno scheduling adattativo un buon uso della banda trasmissiva, trascurando problemi relativi alla non idealità del canale radio.

Nella fase di definizione dello strato MAC, allo scopo di poter valutare le scelte di progetto effettuate e di poter realizzare un simulatore, si è reso necessario procedere con un primo dimensionamento di massima delle risorse del sistema. Alcune delle scelte operate non riguardano esclusivamente funzionalità appartenenti allo Strato MAC, la cui definizione e progetto sono affidate all'Unità di Roma, ma anche aspetti tipici di altri strati di un'architettura di comunicazione come la WLAN in considerazione. Le scelte effettuate dall'Unità di Roma nel suo complesso possono essere considerate come una proposta per il dimensionamento di alcuni aspetti del sistema, tra cui:

- tempo di trama
- numero e durata dei Time Slot
- Sotto-portanti utili del sistema OFDM

Il primo elemento che deve essere dimensionato è la MAC PDU utilizzata sulla tratta di *downlink* (dal Radio Node ai Radio Terminal). La lunghezza della MAC

PDU dipende in generale da una serie di fattori in trade-off che richiedono da un lato di avere una lunghezza elevata per l'unità dati del MAC e dall'altro una di dimensione ridotta.

In generale, avere grandi unità informative di Strato MAC consente di ridurre il peso percentuale dell'informazione di *overhead* necessaria a gestire il protocollo, aumentandone così l'efficienza. D'altra parte in questo modo aumentano i ritardi di riempimento dell'unità dati ed i tempi di trasferimento si dilatano. Questo è dannoso per applicazioni con necessità di tempo reale. Ne segue che tale scelta deve risultare da un compromesso fra queste due esigenze.

Come accade in generale, la MAC PDU in fase di definizione deve prevedere una parte di unità dati (*payload*) e una parte di informazione di controllo aggiuntiva (*overhead*).

Il punto di partenza verso la scelta delle dimensioni della MAC PDU è stata la volontà di non segmentare il tipico pacchetto vocale di Strato IP (72 Byte). Fissato in questo caso un payload di 72 Byte, si è passati alla definizione dell'informazione di controllo ed al suo dimensionamento.

A questo scopo, è stata effettuata uno studio dei campi tipicamente presenti nella definizione dell'informazione di controllo degli standard IEEE 802.11 e ETSI Hyperlan 2. L'informazione di controllo aggiunta, dovrà offrire alcune funzionalità.

Parte dell'overhead verrà dedicato all'indirizzamento, in modo che il RN possa identificare in DL il Radio Terminal di destinazione della MAC PDU, oppure una comunicazione *multicast* o *broadcast* attiva nella WLAN.

Un'altra funzionalità sicuramente presente è distinguere le diverse classi di servizio in modo da poter assicurare a ciascuna connessione il rispetto dei requisiti di qualità accordati. Inoltre, le diverse classi di servizio possono avere differenti necessità di protezione (in generale, FEC nel caso di classe GB e ARQ nel caso di classe BE). Ad esempio, considerando la classe GB, se da un lato un BER pari a 10^{-4} è sufficiente a garantire l'integrità informativa per servizi come il trasferimento della voce, dall'altro non è sufficiente a garantire i parametri di qualità per quelle applicazioni ad elevato bit-rate (tipicamente video) che appartengono comunque

nella classe di servizio GB. Un discorso analogo vale anche per applicazioni appartenenti alla classe BE.

Inoltre, risulterà necessario identificare con un opportuno numero di sequenza le MAC PDU derivanti dalla segmentazione di uno stesso pacchetto IP e, utilizzando un ulteriore campo, quando la MAC PDU in questione è l'ultima appartenente ad uno stesso pacchetto IP allo scopo di effettuare correttamente le operazioni di segmentazione e ricostruzione.

Dovrà essere definito un apposito campo per identificare la lunghezza in byte del payload, e implicitamente la lunghezza dei bit di riempimento (*padding*).

Naturalmente, l'informazione di overhead sarà protetta da un apposito campo informativo. L'informazione d'utente sarà invece protetta da un codice a rivelazione di errore (CRC), ad esempio nel caso di applicazioni dati, oppure da una codifica a correzione di errore "in avanti" (FEC), ad esempio nel caso di applicazioni con requisiti di qualità del servizio garantita.

Considerando quindi le grandezze tipiche di questi e altri campi nel caso di trasmissione di un pacchetto IP per la voce (72 Byte) è stato deciso di considerare, per il caso di una MAC PDU di tipo vocale, un'informazione di controllo complessivamente di 18 Byte, raggiungendo di conseguenza, la dimensione complessiva di 90 Byte.

Avendo dimensionato la MAC PDU vocale, poiché il protocollo utilizza una MAC PDU di formato fisso, la sua dimensione è fissata a 90 Byte, indipendentemente dalla particolare connessione. Le dimensioni relative del payload e dell'header non sono però fisse e dipendono dal tipo di sorgente (voce, dati, …), allo scopo, fra gli altri, di garantire una diversa codifica di canale in funzione dei requisiti di qualità accordati alle diverse connessioni.

Fig.3.4 - Rappresentazione MAC-PDU

A questo punto è importante sottolineare che, allo scopo di realizzare il simulatore, si è ipotizzato che la dimensione del payload sia effettivamente di 72 Byte e, di conseguenza, quella dell'overhead di 18 Byte.

Il passo successivo è la definizione del Periodo di Trama. In questo caso la scelta è stata operata introducendo il concetto di *sorgente elementare*. Si definisce sorgente elementare la tipica codifica PCM del segnale vocale che corrisponde a considerare un flusso CBR a 64 Kb/s.

Si vuole che la sorgente elementare generi una MAC PDU vocale (corrispondente al pacchetto IP per la voce di dimensione 72 Byte) per ogni periodo di Trama *T*, che quindi vale:

$$T = \frac{72 \cdot 8bit}{64000bit/\sec} = 9ms$$

A questo punto, si passa a definire il numero globale di risorse da allocare all'interno della singola trama. Il numero di risorse allocabili dipende da quanti Time Slot sono contenuti in una Trama e quanti codici (quante sottoportanti) sono supportati dal sistema di trasmissione. Si è scelto di dimensionare il sistema introducendo 6 Time Slot per Trama e, riferendosi anche alle scelte operate dagli Standard IEEE 802.11 e Hyperlan 2, considerando un sistema che supporta 52 codici differenti (vedi figura 3.5):

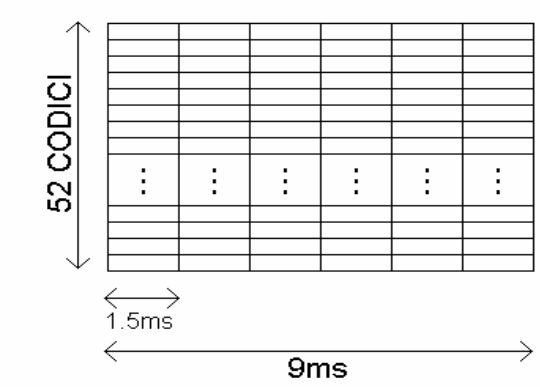


Fig.3.5 - Dimensionamento della struttura a trama

In base alle scelte operate, per ogni trama è possibile trasmettere 312 MAC PDU. Pertanto, per ogni trama è allocabile una capacità lorda equivalente a 312 sorgenti

elementari. Si tratta di una capacità lorda poiché, nell'assegnare le risorse agli utenti MAC, occorre riservare, per ogni trama, alcune PDU per i canali di segnalazione broadcast (ad esempio il Canale di Paging, il canale di Allocazione).
All'interno di una determinata coppia <Time Slot, Codice> viene trasmessa una MAC PDU codificata (l'utilizzo dei codici permette la trasmissione contemporanea di 52 MAC PDU nello stesso Time Slot).
Si può allora ricavare il ritmo binario del segnale aggregato:

$720bit \times 52 = 37440bit$ in 1.5ms

corrispondenti a:

$$R = \frac{37440bit}{0.0015s} = 24.96Mb/s$$

Questo è il ritmo binario del segnale aggregato in aria. Il risultato ottenuto, in realtà, è in difetto poiché non stata considerata l'estensione ciclica di ogni simbolo OFDM, che, d'altra parte, non può essere dimensionata senza una stima del massimo valore di *delay spread.*
A questo punto è possibile ricavare la capacità che lo Strato MAC è in grado di assegnare per la trasmissione di unità informative. Poiché la capacità corrispondente all'assegnazione di un elemento della matrice dei codici è pari a 64 Kb/s, in totale il MAC può allocare:

$C = 64Kb/s \times 52 \times 6 = 19.968Mb/s$

che corrisponde a circa 20 flussi a Mb/s e che si ritiene adeguata al sistema di comunicazione in esame.
A questa capacità lorda va sottratta la capacità utilizzata dal sistema per i canali di segnalazione broadcast, che sono il canale di Paging (Paging Channel, PagCh) ed il canale di allocazione (Allocation Channel, AlCh). Stesso procedimento sarà

riservato anche in UL dove il canale logico di segnalazione è rappresentato dal canale delle richieste (Request Channel, ReqCh).

Il canale di Paging è trasmesso in DL e consente di comunicare al generico RT quali sono le risorse trasmissive all'interno della prossima trama, sulle quali ci sono delle unità dati a lui indirizzate (dal RN al RT).

L'ovvio vantaggio di questo metodo di accesso al mezzo utilizzato per la tratta in DL è che i terminali in ricezione devono "ascoltare" solo in determinati periodi di tempo (in particolare, in determinati TS) e utilizzando solo determinati codici; si è creata cioè quella segnalazione che ha come scopo quello di minimizzare la potenza consumata in ricezione cui si accennava nel Paragrafo 1.

L' utilizzo dei canali ReqCh (in UL) e AlCh (in DL) è legato alla modalità scelta per l'accesso al mezzo (tratta UL) da parte dei RT.

In fase di progettazione si sono confrontate due modalità: una centralizzata, l'altra distribuita.

Nella modalità *centralizzata,* le informazioni sulle richieste di capacità sono comunicate dai RT al RN, il quale decide le allocazioni per poi comunicarle ai RT sull' AlCh.

Nella modalità *distribuita,* le informazioni sulle richieste di capacità sono rese note dal RN a tutti i RT. Ogni terminale attestato sull'interfaccia radio esegue un algoritmo distribuito di allocazione delle risorse basato sulle richieste di tutti gli altri RT e determina in modo autonomo la propria assegnazione di capacità.

Pur riconoscendo una buona dinamicità al secondo meccanismo, ci si è orientati verso la soluzione centralizzata per una maggiore semplicità nella gestione della segnalazione. In tale ottica, il ReqCh è utilizzato dai RT per effettuare le richieste di capacità, mentre l' AlCh è utilizzato dal RN per comunicare ai diversi RT quali sono le risorse loro assegnate per trasmettere sull'interfaccia radio in UL.

Le informazioni d'utente sono trasferite sull'interfaccia radio in modalità "senza collisioni". Queste sono fondamentalmente un problema dell' UL e dovrebbero essere evitate non solo per riuscire a raggiungere un alto throughput sull'interfaccia radio ma soprattutto al fine di rispettare i vincoli sul ritardo di trasferimento per il

traffico GB. Per questo motivo, si è scelto per il sistema in esame un approccio senza collisioni per tutte le classi di traffico considerate.

Così come per i dati anche per la segnalazione emessa dai RT sono possibili due approcci diversi: uno *senza collisione,* l'altro ad *accesso casuale.*

Il primo approccio è quello classico, nel quale la capacità di segnalazione è associata staticamente ad ogni RT, che periodicamente accede o ascolta un canale di segnalazione a lui dedicato, mentre il secondo si adatta bene al caso di segnalazione orientata al messaggio e guidata dagli eventi.

Nella definizione della segnalazione utilizzata dal protocollo MAC, nell'ambito di questa tesi, è stato considerato solamente il primo approccio.

A questo punto bisogna entrare più in dettaglio nella struttura della trama di DL, mostrata in fig. 3.6. Come si può osservare, parte del primo TS è occupata dai canali di segnalazione definiti precedentemente.

L'unico canale di segnalazione considerato all'interno del simulatore è, però, il solo Paging Channel ed è così per due motivi:

1) si è considerata unicamente la tratta DL
2) si è ipotizzato che il sottostante canale radio sia privo di errori.

In fase di progettazione, si è scelto di costruire un canale di Paging a lunghezza fissa, che obbliga il RN a trasmettere l'informazione di segnalazione verso tutti i RT, anche se alcuni di essi sono inattivi. Sebbene questa soluzione sia inefficiente in termini di banda richiesta rispetto ad un canale di segnalazione a lunghezza variabile dove vengono trasmesse solo le informazioni relative ai RT attivi, è stata preferita la semplicità della prima scelta, anche perché scopo di questo lavoro non è valutare l'efficienza della segnalazione relativa alla allocazione delle risorse.

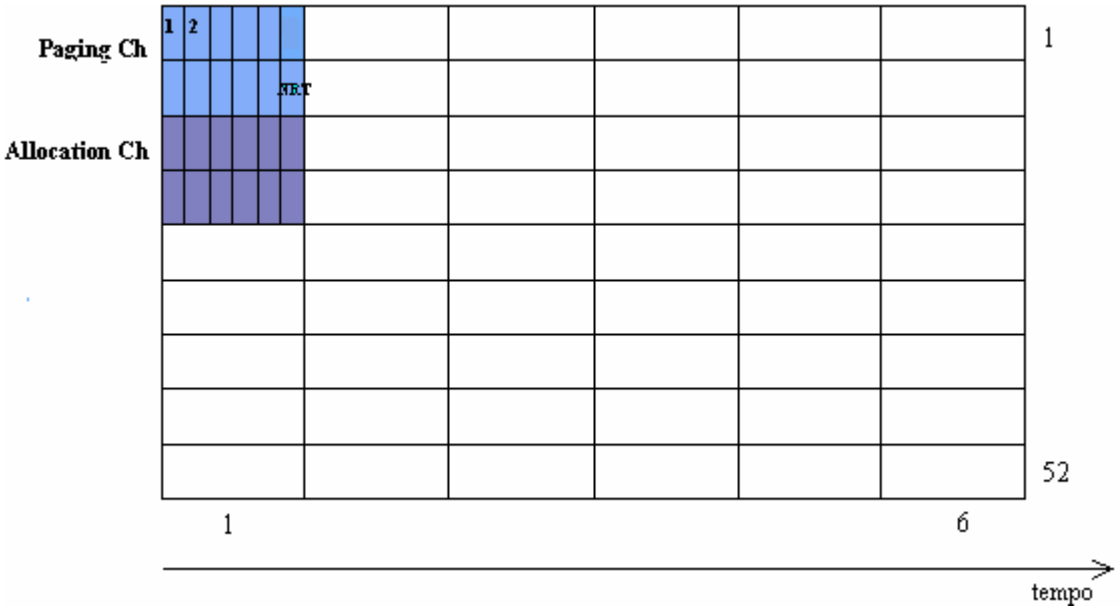


Fig.3.6 - Struttura della Trama di Downlink

Il canale di Paging è distribuito su un solo TS e su un insieme di codici (in figura 3.6 si è immaginato fossero due). Ognuna delle MAC PDU impiegate per la trasmissione del Paging Channel contiene un numero *s* di *minislot* i quali, trasportano l'informazione relativa ad un singolo RT. Il RN, una volta deciso, attraverso un apposito algoritmo di scheduling (vedi cap.5), il numero di MAC PDU che intende dedicare a ciascuna connessione nella prossima matrice di trama, deve decidere come posizionare queste MAC PDU all'interno della trama. Le risorse disponibili in una trama sono "tutte uguali", poiché è stato ipotizzato un canale radio privo di errori. L'assegnazione di una risorsa ad una determinata connessione RN-RT comporta che la MAC PDU trasmessa su quella risorsa venga ricevuta correttamente dal RT, qualunque sia la sua posizione all'interno della trama.

In considerazione di questo, per non dare nessuna preferenza nell'assegnazione delle risorse della matrice dei codici ai vari RT, la disposizione effettuata dal RN delle MAC PDU all'interno della trama è *casuale*; in altre parole, dopo aver stabilito il numero di MAC PDU che saranno dedicate ad ogni connessione, il riempimento della trama è effettuato tramite estrazione di una variabile aleatoria a distribuzione uniforme.

Di conseguenza ogni minislot deve essere in grado di comunicare al RT corrispondente le coordinate delle risorse della prossima trama dove dovrà "ascoltare" MAC PDU a lui dirette.

La soluzione adottata per la trasmissione del Paging Channel "in parallelo", cioè su un solo TS ma utilizzando *H* codici, risulta più efficiente rispetto alla soluzione cosiddetta "in serie", cioè su un solo codice ma su *H* TS, poiché riduce il tempo necessario ad effettuare l'allocazione, a parità di banda utilizzata dall'intero Paging Channel.

In fase di progettazione, è stato considerato che una volta che il RN ha trasmesso il canale di Paging, esso deve essere correttamente ricevuto ed interpretato dai terminali; al fine di permettere ai RT di capire da dove leggere le informazioni a lui dirette, si è considerato di lasciare 2 Time Slot fra il termine del Paging Channel e il primo Time Slot in cui effettivamente potrebbe trovarsi una MAC PDU diretta a uno dei RT attivi e che ha "ascoltato" il Paging Channel. Dunque l'informazione trasportata dal Canale di Paging si riferisce ad una struttura che ha la stessa dimensione di una trama di DL, a meno di una traslazione ciclica di due Time Slot, come si può vedere in figura 3.7:

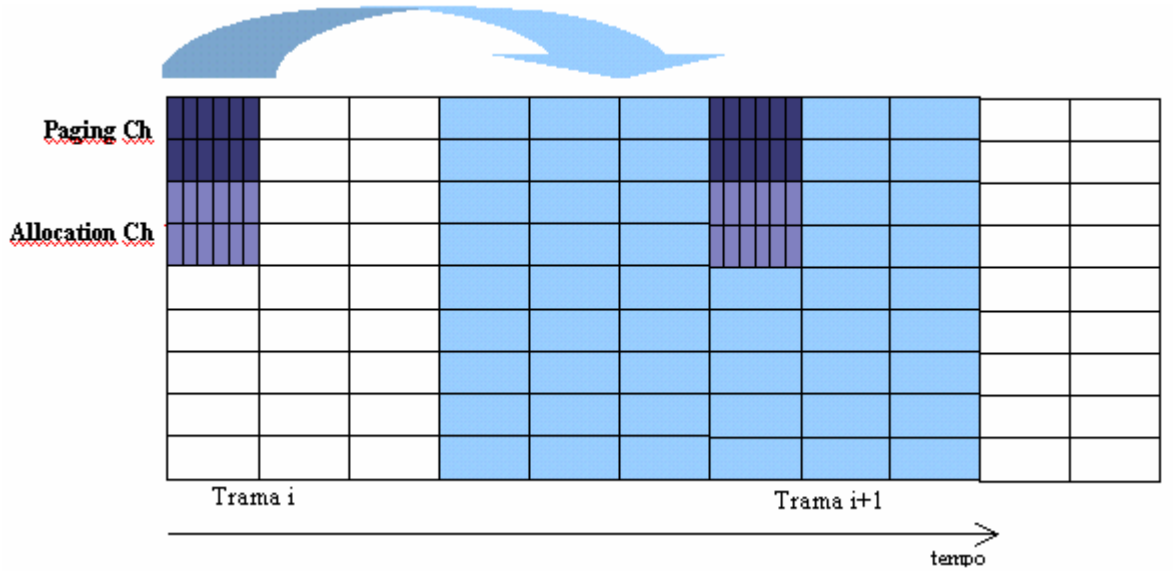


Fig. 3.7 - Struttura a Trama e Matrice dei Codici sulla Tratta in Downlink

L’informazione trasportata dal Paging Channel si riferisce all’area evidenziata sulla destra della figura 3.7.

Questa zona è chiamata Matrice dei Codici e all’interno del simulatore, come spiegato anche nel paragrafo 6.1, risulterà assai importante la distinzione fra le due strutture, Trama e Matrice dei Codici.

Il Canale delle Allocazioni è previsto nella Trama di DL perché necessario al RN per comunicare ai diversi RT in quali risorse della trama di UL possono trasmettere le proprie MAC PDU. In realtà, la struttura dell’Uplink è in questo momento oggetto di studio dell’Unità di ricerca di Roma e si stanno valutando diverse tecniche di accesso al mezzo. Per questo motivo, pur restando validi i ragionamenti svolti per il Paging Channel è più difficile quantificare l’informazione che dovrà essere trasportata dall’ AlCh (che, come il Paging Channel può essere assunto avere dimensione fissa, almeno in una prima fase). Tuttavia, anche esso sarà organizzato in minislot come il Paging Channel, ed ognuno di essi corrisponderà ad un diverso RT. Anche in questo caso, è facile prevedere l’introduzione di un opportuno ritardo fra la fine dell’AlCh e l’inizio della prima risorsa della trama di UL dove i RT possono trasmettere le proprie MAC PDU in attesa, per dar modo ai RT di riuscire a capire dove prelevare le informazioni, in modo analogo a quanto succede nel caso del Paging Channel.

Per quanto riguarda la struttura dell’uplink, ancora oggetto di studio, esso dovrà prevedere un canale, il ReqCh, con cui i terminali richiedono di poter trasmettere al RN che, come si è detto, risponderà attraverso l’ AlCh.

Il canale di richiesta ReqCh consente ai vari RT di inviare le richieste di allocazione di opportunità di trasmissione al RN. Questo significa che ogni RT, sulla base dello stato delle proprie code di traffico, tramite l'accesso al ReqCh, chiede l'assegnazione di una quota parte della capacità disponibile all'interfaccia radio.

Esistono diverse alternative per la struttura del ReqCh:

- accesso *dinamico* al canale delle richieste (cioè con collisione) per tutte le classi di servizio supportate dal MAC;
- accesso *statico* al canale delle richieste (cioè senza collisione) per tutte le

classi di servizio supportate dal MAC;

- accesso *ibrido* al canale delle richieste che corrisponde ad un accesso *statico* per la classe GB ed ad un accesso *con collisione* per la classe BE

In generale il vantaggio di un accesso dinamico sta nel risparmio di banda, poiché il ReqCh deve essere dimensionato in modo da garantire un certo "grado di accesso" e può essere variato dinamicamente dal RN e comunicato su un canale diffusivo a tutti i RT in accordo al numero di collisioni riscontrate.

La controparte di tale meccanismo di accesso dinamico, sta nel fatto di degradare, per alti carichi offerti, le prestazioni del sistema di accesso in termini di *throughput* e di ritardo di trasferimento, e può causare problemi di stabilità a lungo termine.

Al contrario, un accesso statico implica che una certa quantità di banda, nota a priori e costante nel tempo sia completamente dedicata al ReqCh anche se i RT componenti il sistema non sono attivi in un certo istante.

Per queste ragioni, almeno in questa fase dello studio e per quanto concerne questo lavoro è possibile immaginare un ReqCh ad accesso senza collisione e di dimensione fissata.

Data una panoramica di entrambe le tratte è a questo punto opportuno chiarire quanto vale il ritardo di accesso alla tratta in DL (di cui si è parlato nel Paragrafo 3.4 a proposito delle regole di allocazione delle risorse) subito da una MAC PDU di classe GB. Tale valore è composto da due termini (si ricorda che per la definizione data nel paragrafo 3.4 questo ritardo non comprende il tempo di attesa nei buffer):

- il tempo di attesa per la prossima occorrenza di un Paging Channel (nel caso peggiore questo tempo è pari ad una trama: $N \cdot TS = 6 \cdot 1.5ms = 9ms$)
- l'opportunità di trasmissione che nel caso peggiore può essere assegnata ad una MAC-PDU che riesca a partire nella prima matrice dei codici disponibile e, che è rappresentata dall'ultimo TS della matrice stessa, ovvero in termini analitici da $N \cdot TS = 6 \cdot 1.5ms = 9ms$

Ne deriva che il ritardo totale di accesso sarà dato dalla somma dei due termini appena descritti:

$t_{SISTEMA\_MAX} = 2N \cdot TS = 18ms$

3.2.3 Strato fisico

Attraverso la descrizione fatta nel capitolo 2, delle diverse tecniche di modulazione combinate con i differenti schemi di accesso, si è giunti alla conclusione che la migliore scelta per ottenere un buon compromesso tra flessibilità, complessità e prestazioni, anche se con una contenuta complicazione a livello di strato fisico dovuta allo spreading di dati, è quella della tecnica di modulazione OFDM-CDMA. La modulazione OFDM è stata scelta per limitare la selettività in frequenza del canale radio e la tecnica CDMA è stata considerata come schema d'accesso ideale per supportare servizi multimediali, grazie alla sua capacità di adattamento alla natura asincrona del traffico multimediale, che consente di fornire un trasferimento informativo maggiore di quello offerto dalle altre tecniche di accesso.
Ulteriore vantaggio conseguente alla scelta di OFDM-CDMA è relativo al fatto che i segnali usati possono essere facilmente trasmessi e ricevuti usando la FFT (Fast Fourier Transform) senza aumentare la complessità del ricevitore e del trasmettitore ed ottenendo contemporaneamente un'alta efficienza spettrale dovuta alla minima spaziatura tra le sottoportanti.
Il sistema OFDM-CDMA si presta bene quindi ad un accesso multiplo poiché utenti diversi usano lo stesso insieme di frequenze ma un codice diverso che sarà ortogonale a tutti gli altri. Si deve notare, inoltre che in tale tecnica di modulazione, usata nell'ambito del nostro sistema, ci sono 2 livelli di ortogonalità:

- le sottoportanti sono ortogonali: questo consente non solo la loro sovrapposizione, ma anche la suddivisione del flusso di dati su un numero n di sottoportanti guadagnando sul delay spread di un fattore N ed ottenendo una durata di simbolo N volte più grande.
- i codici di spreading sono ortogonali: se una molteplicità di utenti trasmette un segnale a spettro allargato, il RX sarà comunque in grado di distinguerli grazie al codice assegnato ad ognuno di questi. Attraverso la correlazione del segnale ricevuto con la sequenza di codice dell'utente selezionato, avviene il restringimento (despreading) del segnale di questo utente mentre quelli degli altri rimarranno allargati sull'intera banda.

Da quanto appena detto risulta ben visibile la caratteristica importantissima dell'OFDM-CDMA che consiste in un più facile recupero del segnale in condizione di canale non buone. Nella figura sotto vengono messi a confronto gli spettri di densità di potenza di un segnale proveniente da un DS-CDMA e quello di un MC-CDMA..

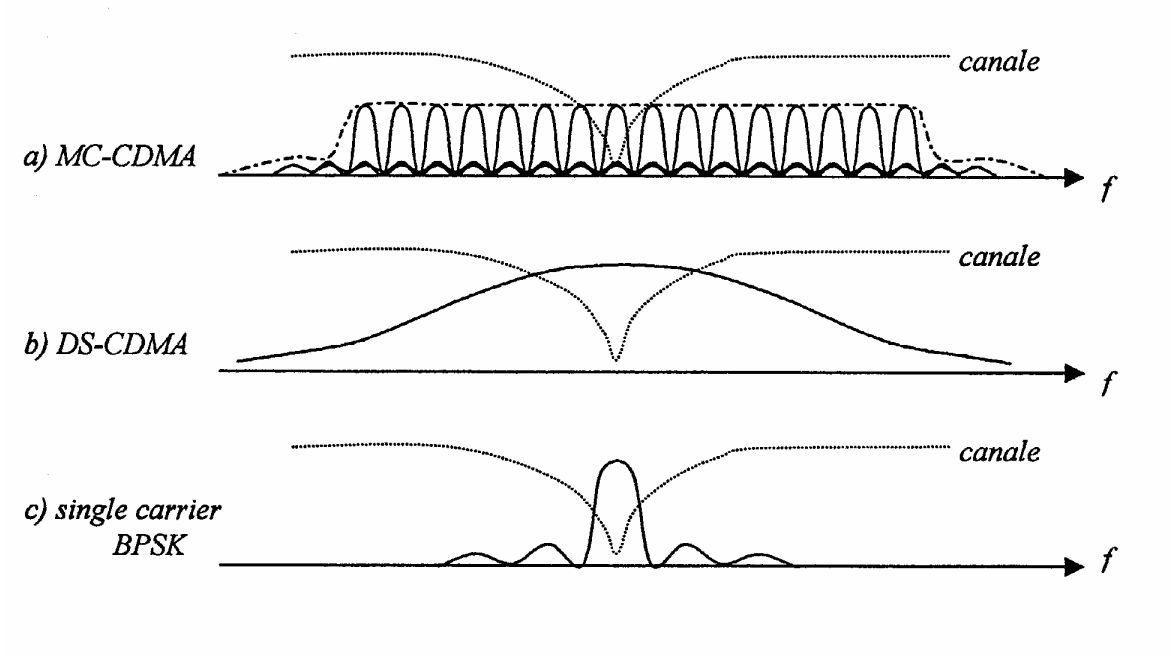


Fig.3.8 – Spettro di a) Segnale proveniente da MC-CDMA (multiportante a banda larga); b) segnale proveniente da un DS-CDMA (singola portante a banda larga); c) Segnale a singola portante a banda stretta

Come si nota facilmente, mentre nel caso DS-CDMA, l'aver usato un codice in tempo che modula il simbolo porta ad avere un unico lobo principale molto largo intorno all'unica portante del sistema, nel MC-CDMA lo stesso simbolo è ripetuto su tutte le sottoportanti del sistema. Un eventuale *notch* come mostrato in figura, cancellerebbe il segnale nel caso DS-CDMA ma non nel caso MC-CDMA, in cui può essere recuperato entro un certo limite di rapporto segnale a rumore, dalle altre sottoportanti. Per i dettagli sull'OFDM e sull'OFDM-CDMA si rimanda al capitolo 2. L'ultimo aspetto da considerare è la scelta del numero di codici da utilizzare nella

struttura che abbiamo definito come “matrice dei codici”. Questa struttura è una vera e propria matrice ed è individuata dal numero di TS e dal numero di codici utilizzati. Basandoci sul numero di codici utilizzati negli standard IEEE 802.11 e HIPERLAN2, anch’essi basati su OFDM-CDMA e sue modifiche, il valore scelto anche nell’ambito del nostro sistema è stato di 52 codici. In realtà entrambe gli standard utilizzano per il trasferimento dei dati solo 48 codici, lasciando disponibili 4 codici per segnalazione e controllo. Il nostro sistema prevede, invece, l’utilizzo per il trasferimento dei dati di tutti i 52 codici, delegando alla segnalazione solo due MAC-PDU per trama organizzate in una struttura a mini-slot.

CAPITOLO 4

INTERFACCIAMENTO DEL MAC CON GLI STRATI SUPERIORI: MODELLI DI TRAFFICO E ALLOCAZIONE DELLA RISORSA TRASMISSIVA

4.1 CARATTERIZZAZIONE E MODELLIZZAZIONE DELLE SORGENTI DI TRAFFICO

Le sorgenti di traffico rappresentano il modo con cui l'utente esprime la richiesta di servizio.

Una sorgente, in prima istanza, può essere caratterizzata secondo due classi di parametri:

- *Caratteristiche in base alla chiamata*: durata e tempo di presentazione della richiesta di chiamata
- *Caratteristiche dell'emissione nell'ambito della chiamata*: capacità di emissione e caratteristiche di attività

Caratterizzare una sorgente vuol dire definire statisticamente il comportamento della sorgente di traffico stessa.

Le sorgenti di traffico si possono raggruppare in quattro grandi famiglie:

1) Sorgenti Audio
2) Sorgenti Video
3) Sorgenti Dati
4) Sorgenti Multimediali che in un certo senso raggruppano le sorgenti precedentemente elencate.

Una prima classificazione delle sorgenti di traffico può, dunque, essere effettuata in base alla modalità con cui i dati sono emessi dal codificatore.

Sulla base di questa premessa è importante definire i tre tipi di sorgenti sotto elencati:

a) Sorgenti di tipo CBR (Constant Bit Rate).

   Questo tipo di sorgente è caratterizzato dalla peculiarità di emettere ad un rate costante come si può vedere, in fig.4.1, dall'andamento "piatto", nel tempo, del rate di picco.

   Un classico esempio è dato dalla voce codificata PCM (Pulse Code Modulation)

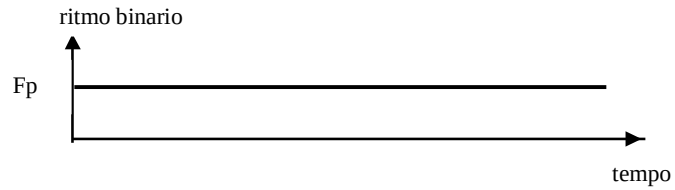


Fig.4.1-Comportamento di una sorgente CBR

b) Sorgenti di tipo VBR (Variable Bit Rate).

   Questo tipo di sorgente ha un rate di emissione variabile nel tempo, come si deduce anche dalla figura 4.2.

   Un esempio di questo tipo di sorgente è dato dalla codifica di immagini in movimento MPEG (Moving Picture Export Group).

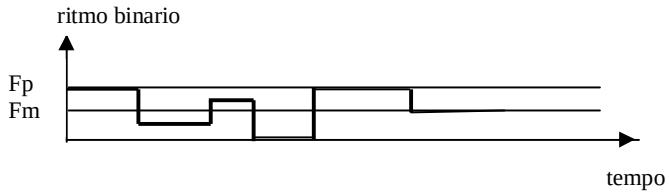


Fig.4.2 - Comportamento di una sorgente VBR

c) Sorgenti di tipo VBR a due stati, anche dette sorgenti ON-OFF ("Tutto o niente").

   Queste sorgenti possono essere considerate come caso particolare delle sorgenti

precedentemente definite. La caratteristica che le differenzia da queste ultime è che l'emissione o avviene al picco o non avviene per niente.

Esempi di sorgente ON-OFF sono le sorgenti vocali con rivelazione di tratti vocali

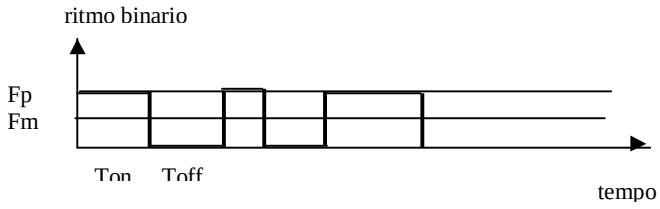


Fig.4.3 - Comportamento di una sorgente VBR a due stati

4.1.1 MODELLI DI TRAFFICO

Traffico vocale e traffico video

Il traffico vocale è ottenuto campionando ad intervalli regolari, e poi codificando, il segnale proveniente da una sorgente vocale. I metodi utilizzati per la codifica sono tali che al ricevitore il messaggio originale viene ricostruito senza degradarne la qualità; inoltre, secondo il tipo di codifica utilizzata, le perdite di celle possono essere compensate o meno.

Come accennato precedentemente, la codifica delle sorgenti vocali gioca un ruolo fondamentale nella modellizzazione del traffico: il traffico generato può essere modellato o come CBR o come VBR a due stati.

Se viene utilizzato un codificatore di tipo PCM a 64 Kbit/s, il traffico è di tipo CBR.

Se invece la codifica si basa sull'uso di tecniche *Speech Activity Detector* (SAD) e *Digital Speech Interpolation* (DSI), il traffico è di tipo VBR a due stati.

Queste tecniche sfruttano la ridondanza intrinseca in un segnale vocale ed eliminano la non necessaria trasmissione degli intervalli di silenzio durante una chiamata.

In letteratura è stato proposto un modello che considera una sorgente vocale come sorgente ON-OFF (ovvero come sorgente VBR a due stati): uno stato di attività (talkspurt o ON) ed uno stato di inattività (silenzio o OFF).

Le sorgenti vocali che emettono questo tipo di traffico sono dette *bursty* ed il periodo in cui la sorgente emette è detto *burst*. Generalmente il burst ha una durata limitata nel tempo, ma è caratterizzato da un bit rate molto elevato. A questo proposito, viene definita una quantità detta *burstiness* che misura l'indice di attività della sorgente, ed espressa dal rapporto tra la banda di picco e la banda media della sorgente. Essa indica se il comportamento della sorgente è più vicino al tipo CBR o al tipo VBR. Per le sorgenti CBR la burstiness è unitaria.

Il numero di celle in trasmissione nella rete nel caso di modellizzazione di una sorgente vocale come sorgente ON-OFF si riduce del 35-40% rispetto al caso in cui la sorgente venisse codificata non considerando i periodi di silenzio. La durata temporale dei periodi di ON e di OFF è rappresentata da una variabile aleatoria. Dalla letteratura si deduce che una distribuzione esponenziale si adatta bene alla descrizione dei periodi di attività e, meno bene a quella dei periodi di inattività, che solitamente possono essere meglio approssimati da una distribuzione geometrica. Tipici valori medi dei due periodi sono $\overline{T}_{ON} \approx 352\text{ms}$ e $\overline{T}_{OFF} \approx 650\text{ms}$ [Figueira et altri,1999].

Per quanto concerne il traffico video, la codifica del segnale può richiedere parecchi megabit al secondo ed è quindi di gran lunga superiore ai 64Kbit/s richiesti dal segnale vocale; inoltre la variabilità del bit-rate di un segnale video è molto più alta di quella del segnale vocale, e pertanto un semplice modello, come quello ON-OFF non è adatto per questo tipo di segnali.

In base al tipo di video considerato, si ha una notevole ridondanza di dati, che deriva da un'alta correlazione intraframe e interframe. Sono dunque necessarie delle codifiche particolari.

La codifica JPEG (Joint Picture Export Group) è molto diffusa per le immagini fisse poiché riesce a caratterizzare molto bene la correlazione intraframe e perché, seppure

con una perdita di qualità, presenta una contenuta dimensione in termini di bytes grazie all'utilizzo delle proprietà della DCT (Discrete Cosine Transform) unitamente ad alcune proprietà della "visione umana". Viceversa lo standard MPEG, è usato per codificare le sorgenti video poiché tiene conto anche della correlazione interframe (tra frames diversi).

Un tipico encoder MPEG genera tre tipi di frame: *Intracoded* (I), *Predictive (P)*, e *Bidirectional* (B). Con l' MPEG si suddivide il filmato in *Group of Pictures* (GOP), caratterizzato da un insieme di frame consecutive. La struttura di un GOP è del tipo:

IBBPBBPBBPBB

dove I, P, B rappresentano le frame dette in precedenza, che avendo caratteristiche diverse vengono codificate in modo diverso.

La frame I è codificata con tecnica JPEG. Le frame P vengono codificate secondo una codifica differenziale rispetto alla frame I ed alla frame B nello stesso GOP. Le frame B vengono, infine, codificate esaminando la frame I e la frame P precedente e quella successiva appartenenti allo stesso GOP.

Le frame I sono quelle che hanno maggior contenuto informativo, mentre le frame B sono quelle che ne sono più povere. La caratterizzazione delle sorgenti MPEG non è per nulla semplice, in quanto è necessario caratterizzare statisticamente tutte e tre i tipi di frame costituenti il GOP in questione.

E' ragionevole, comunque, assumere che uno stream video sia dato dalle repliche di un certo pattern GOP e quindi per una eventuale caratterizzazione della sorgente sarebbe importante tenerne conto. Per quanto riguarda questa tesi, le sorgenti MPEG non sono , però, state considerate. Questo perché al fine di una rappresentazione realistica andrebbero valutate le tracce video e modellizzato il traffico come traffico VBR. Una eventuale caratterizzazione delle sorgenti MPEG come sorgenti ON-OFF non si potrebbe considerare del tutto lecita.

Traffico dati (WWW browsing session)

Sono state considerate tra le sorgenti di traffico dati i seguenti sottogruppi:

1. WWW surfing (o scaricamento di una pagina da internet)
2. FTP (File Transfer Protocol)
3. Linea di comando TELNET

Una tipica sessione *WWW browsing* consiste in una sequenza di chiamate a pacchetto (*packet calls*). L'utente inizia una chiamata quando richiede una informazione. Durante una singola chiamata possono essere generati più pacchetti, ciò significa che una "packet call" è caratterizzata da una sequenza a burst di pacchetti. Questo fenomeno deve essere tenuto in conto nella modellizzazione del traffico [Nieminem,1998].

Una sessione di servizio a pacchetto contiene una o più "packet calls" dipendente dall'applicazione sotto esame. Per esempio in una sessione di WWW browsing una "packet call" corrisponde al download di un documento. Una volta che il documento è interamente arrivato al terminale, l'utente utilizzerà un certo periodo per studiare l'informazione. Questo intervallo di tempo viene detto "*reading time*".E' anche possibile che la sessione contenga una sola "packet call": questo è il caso di un trasferimento di file (o FTP).

Normalmente,quindi, una sorgente di traffico dati sarà denotata con diversi parametri:

- *Processo di arrivo della sessione*

  Ci si riferisce al modo in cui una sessione si presenta al sistema. L'arrivo della sessione di set-up alla rete è modellato come un processo di Poisson. Per ogni servizio, si ha un diverso processo. Il processo considerato genera solo l'instante di tempo di arrivo delle chiamate e non ha nulla a che vedere con la terminazione delle stesse.

- *Numero di "packet calls" per sessione*

  Tale numero è rappresentabile attraverso una variabile aleatoria distribuita in maniera geometrica con valore atteso pari al numero medio di chiamate per

sessione.

- *Tempo necessario per il "reading"*

E' il tempo necessario per leggere l'informazione tra due chiamate a pacchetto consecutive.

- *Numero di datagrammi all'interno di una "packet call"*

Il numero di datagrammi all'interno di una "packet call" possono essere generati mediante l'uso di diverse distribuzioni statistiche. Tale variabile non è stata presa in considerazione, nella caratterizzazione della prima versione del simulatore, in quanto si è supposto di non dimensionare il traffico generato. Nell'evoluzione implementata nell'ambito di questa tesi, si è deciso, per poter rappresentare un comportamento realistico della sorgente, di poter considerare anche il traffico generato. Anziché rappresentare il numero di datagrammi all'interno di una packet call, si è preferito rappresentare il traffico medio generato all'interno della stessa. E' stato, quindi, ipotizzato che tale valore fosse calcolabile attraverso una variabile casuale distribuita geometricamente con valor medio pari alla dimensione media di traffico generato da una sorgente.

- *Tempo di interarrivo tra i datagrammi*:

Questo può essere modellato o mediante distribuzione Geometrica o mediante distribuzione di Poisson. La scelta fatta nell'ambito della tesi è stata quella di modellare i tempi di interarrivo secondo una distribuzione di Poisson.

- *Dimensione del datagramma*

In riferimento alla dimensione dei datagrammi, tipici valori sono:

1. 40 o 44 bytes (per pacchetti di acknoledgment)
2. 512, 576,1024 e 1500 bytes (per pacchetti di dati)

I valori da noi presi in considerazione, come già accennato nel capitolo 3, sono 512 e 1500 bytes.

4.1.2 Prima modellizzazione: Traffico sempre attivo

Un primo studio del sistema di servizio ha portato ad una ipotesi molto semplificata nella gestione e nella caratterizzazione delle sorgenti di traffico. Per poter simulare un primo ambiente di lavoro, infatti, abbiamo ipotizzato di immettere un certo numero di sorgenti all'interno del sistema e di lasciarle attive per tutta la durata della simulazione.
Supporre che la simulazione si basi su un processo di sola nascita sarebbe solo parzialmente corretto dal momento che, come sappiamo, una descrizione di questo tipo sarebbe in evidente contrasto con la realtà di un ambiente di lavoro come quello preso in esame.
Le sorgenti introdotte nella prima versione del simulatore, come già specificato, nascono tutte al momento dell'inizializzazione e si suppone che generino traffico in maniera continuativa. Si evince da queste breve descrizione che il traffico rappresentato tramite un modello statico, in cui è fisso il numero di nascite, non avvengono morti ed il traffico è generato in maniera continuativa, situazione quest'ultima poco vicina alla realtà.
L'unica dinamicità concessa ad un traffico di questo tipo è che le sorgenti, grazie all'utilizzo di una variabile aleatoria che genera i tempi in corrispondenza dei quali queste effettivamente cominciano a trasmettere, nascono in tempi differenti.
Appare chiaro, a questo punto, come il primo tipo di modellizzazione del traffico sia carente sotto due punti di vista: il primo relativo al fatto che non viene preso in considerazione il fatto che sorgenti di diverso tipo dovrebbero essere caratterizzate da differenti probabilità di nascita; il secondo relativo al fatto che non c'è alcun riferimento alla dimensione media di traffico generato da una sorgente in una chiamata ovvero alla durata media di una sessione di servizio.
Per poter tenere conto di questi due aspetti, sarà allora necessario caratterizzare le sorgenti sia con parametri descrittivi del traffico offerto alla rete durante la loro attività, che con parametri in grado di modellarne l'evoluzione dinamica (dal punto di vista del numero ma anche della tipologia).

4.1.3 Modellizzazione del traffico tramite i Processi di nascita e morte

Nel nostro studio siamo interessati a valutare le prestazioni del sistema in un ambiente di lavoro verosimile; si è potuto comprendere, pertanto, che una descrizione statica delle sorgenti di traffico è da escludere.

Per definire l'evoluzione del traffico è necessario specificare in termini statistici gli eventi di nascita e morte delle sorgenti. Ci interessa definire un vero e proprio diagramma a stati per modellare la dinamicità degli utenti che offrono traffico al sistema.

Il primo passo che dobbiamo eseguire riguarda la caratterizzazione statistica della nascita di nuove sorgenti. Si tratta di modellare l'arrivo di queste al sistema di servizio.

Esistono due soluzioni alternative; la prima, di tipo formale, si basa sull'assegnare la distribuzione della variabile aleatoria τ = tempo di interarrivo (mostrata in fig.4.4), il cui valore medio è rappresentato dall'inverso della frequenza media di interarrivo $E[\tau]=1/\lambda$

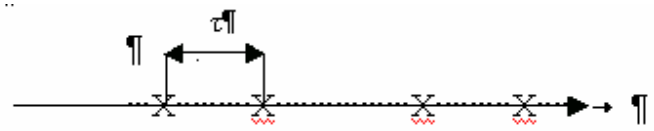


Fig.4.4 – Nascita di una sorgente in un intervallo τ

La seconda, implementata per ottenere una simulazione più fluida e semplice, si basa sulla definizione del numero medio degli arrivi in un intervallo di durata stabilita t (come mostrato in fig.4.5). Si tratta, quindi, di semplificare la distribuzione dei tempi di interarrivo definendo esclusivamente il numero medio di nascite nell'intervallo di riferimento. Attraverso questo parametro, da cui è facilmente derivabile la frequenza media di interarrivo, è quindi possibile approssimare il processo di nascita con un unico parametro: la probabilità che sia nata una nuova sorgente nell'unità di tempo.

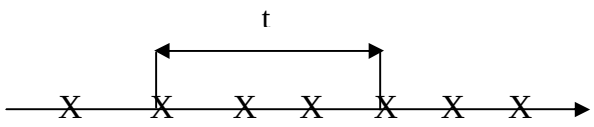


Fig.4.5 – Nascita di più sorgenti nell'intervallo τ

Nell' ambiente di simulazione, dal punto di vista dell'evoluzione dinamica, l'asse dei tempi è stato discretizzato in trame. In questo modo la trama diventa a tutti gli effetti la nostra nuova unità di tempo ed è per questa che ci interessa definire la probabilità di nascita di una nuova sorgente. Chiaramente si sta facendo l'ipotesi che non più di una sorgente dello stesso tipo accede al sistema nel periodo di una trama (ipotesi verosimile in quanto abbiamo visto che la durata di questa è estremamente ridotta e la probabilità di nascita rimane contenuta entro valori considerevolmente bassi). Altra ipotesi assunta in tale modello è quella dell'indipendenza statistica degli eventi di nascita, chiaramente rimovibile in una futura trattazione del problema.

Avendo considerato diversi tipi di sorgente, il procedimento viene ripetuto per tutti i tipi considerati (sia a banda garantita che Best Effort).

Passo successivo è la caratterizzazione della morte di una sorgente. Anche in questo caso la soluzione formale sarebbe quella di introdurre una nuova variabile aleatoria , H, che rappresenti la durata della chiamata o del servizio, di valor medio $E[H]=1/\mu$.

Le proprietà statistiche di tale variabile non dipendono solamente dalle caratteristiche di attività della sorgente, ma vengono influenzate largamente da aspetti legati allo stato del sistema di trasmissione, come la capacità trasmissiva e la velocità di trasferimento dei dati, tra l'altro variabili col tempo.

Per caratterizzare l'evoluzione dinamica di una sorgente è fondamentale conoscere allora due parametri:

- la probabilità di nascita della sorgente nell'intervallo di riferimento (la trama)
- la quantità di traffico generata dalla sorgente.

Per quanto concerne la probabilità di nascita della sorgente, questa può essere definita

con accuratezza attraverso la conoscenza del parametro λ che rappresenta il numero di chiamate nell'unità di tempo e che può variare da sorgente a sorgente.

Definito per ogni tipo di sorgente, il numero di chiamate in una trama e supponendo che le chiamate avvengano in modo aleatorio, si ricava, assumendo una distribuzione esponenziale negativa dei tempi di interarrivo di queste, che la probabilità, che, in $\tau = T_{TRAMA,}$ nasca una chiamata di un certo tipo di sorgente, è espressa da:

$$Ft(\tau) = 1 - e^{-\lambda * \tau} \qquad (1)$$

dove $\lambda * \tau$ rappresenta il numero medio di chiamate nell'intervallo di tempo τ.

Una gestione di questo tipo delle nascite consente al variare del parametro λ, di aumentare la probabilità di nascita di un certo tipo di sorgente. In questo modo si riescono a simulare sia situazioni in cui si hanno molte nascite di sorgenti a qualità garantita e poche di tipo best effort, sia il caso opposto ed ovviamente tutte le situazioni intermedie.

Per quanto riguarda la morte della sorgente, la scelta migliore che si può fare è quella di introdurre un nuovo modello statistico che individui il termine della chiamata in funzione del numero di bit che la sorgente deve trasmettere. L'aleatorietà della morte di una sorgente è nascosta nell'aleatorietà della quantità di traffico, originata nell'arco di tempo di una chiamata, che quella sorgente offre al sistema.

Si considera conosciuto il valore medio del traffico generato da un certo tipo di sorgente. Il generico valore di traffico, sviluppato durante una chiamata, si suppone ricavato da una distribuzione geometrica. Questo che segue, è il procedimento utilizzato nella simulazione, per poter creare ogni volta che nasca una sorgente di un certo tipo, una dimensione di traffico diversa.

La distribuzione geometrica è rappresentata dalla seguente formula:

$$\Pr\{B = k\} = p * (1 - p)^{k-1} \qquad (2)$$

che esprime la probabilità di aver ottenuto per $k-1$ volte un insuccesso e la

$k-esima$ volta un successo, nel senso che fissato il valore di bit pari a k, costituenti il traffico per una certa chiamata, e conoscendo p, la probabilità che il traffico sia costituito da k bit è $\Pr\{B=k\}$.

Chiarito il significato della formula relativa alla distribuzione geometrica, quello che è necessario fare per generare ogni volta valori di traffico differenti, è individuare una variabile aleatoria che rispetti la distribuzione detta, ricordando l'ipotesi che il valore atteso è un parametro conosciuto.

Quando si ha a che fare con distribuzioni discrete come quella geometrica, diventa difficile estrarre la variabile aleatoria rappresentante la distribuzione stessa, mediante una semplice inversione della funzione di distribuzione. E' pertanto necessario ricorrere a degli accorgimenti, che ci consentano di trattare la distribuzione discreta come caso particolare di una distribuzione continua.

Dato il valore del traffico medio e, ricavato il valore atteso $\overline{B}$ [1] della distribuzione , si può calcolare il valore di p, necessario per generare la variabile aleatoria traffico.

Il valore atteso di B è pari a $\overline{B}=\frac{1}{p}$.

Per ipotesi, tale valore coincide con la dimensione media del traffico generato nell'arco di una chiamata indicato di seguito come $\overline{traffico}$.

Si ricava immediatamente che se $\overline{B}=\frac{1}{p}=\overline{traffico} \quad \Rightarrow \quad p=\frac{1}{\overline{traffico}}$.

Ricavato il valore di p, dobbiamo ora generare una v.a. geometrica X con valore medio $\overline{traffico}$. Per poter generare una v.a. di tale tipo, visto che la distribuzione

---

[1]

$$\overline{B}=\sum_{k=1}^{\infty} kp(1-p)^{k-1}=p\sum_{k=0}^{\infty} k(1-p)^{k-1}=-p\sum_{k=0}^{\infty}\frac{\partial}{\partial p}(1-p)^{k}=-p\frac{\partial}{\partial p}\sum_{k=0}^{\infty}(1-p)^{k}=$$

$$=-p\frac{\partial}{\partial p}\sum_{k=0}^{\infty}(1-p)^{k}=-p\frac{\partial}{\partial p}\frac{1}{1-(1-p)}=-p\frac{\partial}{\partial p}\frac{1}{p}=p\frac{1}{p^{2}}=\frac{1}{p}$$

geometrica è una distribuzione discreta, bisogna ricorrere ad un espediente: l'uso della distribuzione esponenziale negativa.

Consideriamo una v.a. esponenziale[2] Z per la quale si consideri la probabilità che $n \le Z \le n+1$:

$$P\{n \le Z \le n+1\} = \left(1 - e^{-(n+1/\lambda)}\right) - \left(1 - e^{-n/\lambda}\right) = e^{-n/\lambda} * \left(1 - e^{-1/\lambda}\right) \qquad (3)$$

Se si impone

$$1 - p = e^{-1/\lambda} \quad \textit{dalla quale poi si ricava} \quad \lambda = -\frac{1}{\ln(1-p)} \qquad (4)$$

la $P\{n \le Z \le n+1\}$ è esattamente la $\Pr\{B = n+1\}$

A questo punto sapendo che per generare una variabile aleatoria esponenziale negativa, è sufficiente invertirne la sua funzione di distribuzione (mostrata in nota 2) ottenendo:

$$X = -\alpha * \ln(1-U) \;\; o \;\; X = -\alpha * \ln U \qquad (5)$$

dove con U si è caratterizzata la distribuzione. Si può dire che, per il calcolo della nostra v.a geometrica, facendo riferimento alle formule (4) e (5), si ottiene per la v.a Z introdotta una espressione data da:

$$Z = \frac{\ln U}{\ln(1-p)} \qquad (6)$$

e quindi per la v.a. B caratterizzante il valore di traffico di ogni chiamata un valore dato da:

$$B = \lfloor 1 + \underline{Z} \rfloor$$

approssimabile, nel caso in cui p sia molto piccolo, con Z stesso.

La durata della chiamata è chiaramente legata al valore di B trovato.

Ne deriva che il modello dell'evoluzione del traffico è un modello a stati, in cui la nascita di una sorgente è identificata da λ, mentre la sua morte è identificata da μ che rappresenta l'inverso del valor medio della durata di una chiamata, legata alla quantità

---

[2] La distribuzione di una variabile aleatoria negativa è data da $F_X(x) = 1 - e^{-\frac{x}{\alpha}} = U$

di traffico generata in media da una sorgente e al ritmo con cui la stessa sorgente trasmette i dati.

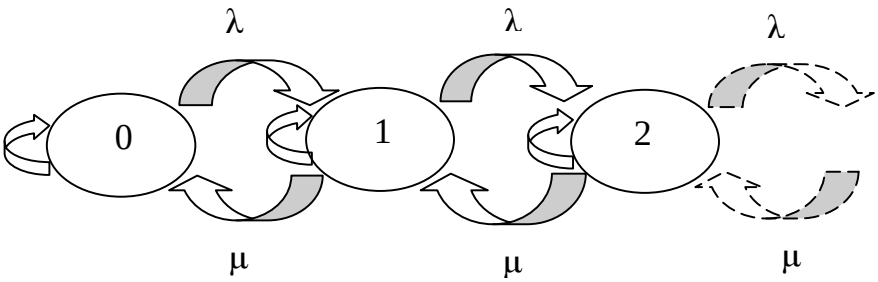


Fig.4.6 – Diagramma di transizione di stato

4.1.4 Parametri delle sorgenti

Come nei paragrafi a seguire verrà chiarito, è necessario definire dei parametri che rappresentino un flusso di dati, sia questo voce, video o dati, e che verranno usati per poter valutare la conformità del traffico generato dalle varie sorgenti alle specifiche di un contratto, che si suppone venga fatto tra utente del servizio e rete. Senza entrare nei dettagli del protocollo usato per garantire il rispetto del contratto di traffico, che viene trattato in maniera approfondita nel paragrafo 4.3.2 insieme ai parametri di Dual Leaky Bucket, definiamo ora i parametri di flusso necessari per la descrizione delle sorgenti che consideriamo all'interno del nostro sistema.
I parametri necessari per poter caratterizzare un flusso sono:

- il ritmo medio *r*
- il ritmo di picco *P*
- la frequenza con cui i pacchetti si presentano al sistema, nel caso di traffico BE
- il massimo ritardo consentito *t*
- la massima dimensione del pacchetto *M*,

- la quantità di traffico sviluppata in media da quel tipo di flusso

La scelta dei parametri si è basata sostanzialmente su valori reali, presi dalla letteratura e da tabelle.

Nel pieno rispetto delle varie tipologie di traffico fatta all'inizio del capitolo, è stato supposto che il sistema fosse in grado di trattare sorgenti ON-OFF, sorgenti CBR e sorgenti BE ed è stata considerata una ulteriore suddivisione degli stessi in sottoclassi che, si è pensato, siano rappresentative di valori tipici di traffico supportati all'interno di una rete wireless.

SORGENTE ON-OFF	SORGENTE CBR	SORGENTE BE
Sottoclassi considerate		
FTP con qualità garantita	Audio	HTTP
Video conferenza	Voce	FTP
Voce codificata ADPCM	Video	SMTP (e-mail)

Tabella 4.1 - Classi e sottoclassi di traffico

Tra le sorgenti ON-OFF è stata individuato anche un tipo di sorgente, il trasferimento di file (FTP) che normalmente rientra tra le BE, come classe di servizio, ma che può essere comunque considerata tra le sorgenti a qualità garantita. Valori caratteristici di un tale tipo di sorgente variano nel range tra 0.1 e 10 Mbit/s.

Per quanto concerne la video conferenza, il range di variazione dei rate va dai 28.8 Kbit/s per la video conferenza a basso rate a valori di ≈ 8 Mbit/s.

Peculiarità del nostro sistema è l'aver considerato la sorgente voce rappresentata sia attraverso una modellizzazione come sorgente ON-OFF che come CBR. Come sorgente ON-OFF abbiamo scelto la voce codificata ADPCM avente rate di picco di ≈ 32 Kbit/s ed rate medio di 11.2 Kbit/s; come sorgente CBR la voce codificata PCM avente un rate tra i 48 Kbit/s ed i 64 Kbit/s. Range relativi a sorgenti audio, invece, variano tra i 256 Kbit/s e 1,5 Mbit/s [Figueira et altri,1999].

Le sorgenti BE sono state identificate in base alle frequenze di arrivo dei pacchetti e rientrano tra le sorgenti di traffico poissoniane: l'arrivo dei pacchetti è modellato tramite processo di Poisson.

I valori presi in considerazione sono di ≈ 16 pacchetti/s per SMTP, ≈ 80 per FTP e ≈ 15 per HTTP. Tali valori sono direttamente ricavabili assumendo di conoscere il ritmo medio di emissione della sorgente r e la lunghezza media dei pacchetti L. Sulla base di questi due valori, è possibile calcolare il tempo di interarrivo pari a L/r, e da questo la frequenza.

I ritardi ammissibili per le sorgenti real time variano tra i 100 ms e qualche secondo. La tabella riportatata di seguito mostra tipici ritardi di differenti applicazioni sia real time che non real time.

Applicazione	Rate medio (kbps)	Rate di picco (kbps)	Max. ritardo (sec)
e-mail, paging	0.01-0.1	1-10	<10-100
computer data	0.1-1	10-100	<1-10
Telephony	10-100	10-100	<0.1-1
Digital audio	100-1000	100-1000	<0.01-0.1
Video conference	100-1000	1000-10000	0.001-0.01

Tabella 4.2 – Rate e ritardi di varie applicazioni di rete

Ultimi due parametri rappresentativi di un flusso sono la lunghezza del pacchetto IP e la dimensione media del traffico generata da una sorgente.

La lunghezza dei pacchetti può essere scelta in modo che ognuno di essi possa essere inviato allo strato sottostante in un unico frame, senza subire frammentazioni. Come già descritto nel capitolo 3, poiché l'IP rappresenta l'offset dei dati in multipli di otto bytes, la dimensione del pacchetto va scelta come multiplo di 8. Naturalmente, scegliere il multiplo di otto bytes più vicino alla MTU (Maximum transfer unit) della rete usualmente consente di dividere il datagramma in frammenti uguali; l'ultimo

pacchetto è spesso più corto degli altri. La sorgente può scegliere qualsiasi dimensione di pacchetto lei pensi sia appropriata. In relazione a queste asserzioni, sono stati scelti 3 valori possibili per la dimensione dei pacchetti IP: 72 bytes, 512 bytes e in ultimo 1500 bytes. Il valore di 72 bytes [cap.3] è stato scelto allo scopo di poter trasmettere il più piccolo pacchetto voce, per l'appunto 72 bytes all'interno di una MAC-PDU senza che questo subisse spezzettamenti. E' stato scelto il valore 1500 in quanto rappresenta la MTU di una rete Ethernet e, come compromesso tra i due il valore di 512 bytes, valore molto utilizzato in letteratura per pacchetti dati. Infine, la durata media di una connessione, mi ha permesso di definire le dimensioni medie del traffico. La tabella riportata di seguito è stata utilizzata con questo obiettivo

Group & Class	Bandwidth Requirement	Connection Duration	Example
1 Class I	30 Kbps (CBR)	1 - 10 m	Voice Service & Audio-phone
2 Class I	256 Kbps (CBR)	1 - 30 m	Video-phone & Video-conference
3 Class I	1-6 Mbps	5 m - 5 h	Interact. Multimedia & Video on Demand
4 Class II	5-20 Kbps	10 - 120 s	E-mail, Paging & Fax
5 Class II	64-512 Kbps	30 s - 10 h	Remote Login & Data on Demand
6 Class II	1-10 Mbps	30 s - 20 m	File Transfer & Retrieval Service

Tabella 4.3 - Servizi di rete (traffico multimediale)

4.2 CARATTERIZZAZIONE DELLE CLASSI DI TRAFFICO

Una volta caratterizzate e modellizzate le sorgenti che si intende trattare nel presente lavoro, è necessario inserirle in un contesto più ampio, quello delle classi di traffico. Il protocollo MAC da realizzare dovrà far fronte come già più volte chiarito a dei requisiti di qualità del servizio, se presenti, che rispettino le caratteristiche di ogni

singola sorgente di traffico. Per il supporto di servizi a qualità garantita in una rete IP, il modello “Integrated Service” (IS), da considerarsi come estensione del modello IP classico, ha definito le seguenti classi di traffico:

- ***Elastic Traffic***: traffico tradizionale delle reti TCP/IP, indifferente alla variazione di ritardo di transito dei pacchetti e quindi, con una qualità di servizio percepita dall’utente, pari al tempo necessario al trasferimento di un intero elemento informativo tipico dell’applicazione
- ***Inelastic Traffic*** o “***real time traffic***”: traffico con requisiti di tempo reale in cui la qualità è fortemente dipendente dal ritardo di transito in rete ed i cui parametri rappresentativi sono il *throughput* e il *delay*, per i quali è richiesto un valore limite garantito dalla rete, il *jitter* di ritardo e la *perdita di pacchetti*, per i quali è richiesto un limite massimo. Il trattamento del traffico real time è preferenziale rispetto al traffico di tipo “elastic” e necessita di procedure di riservazione delle risorse.
-

sensitive/tolerant to...	**Real-time**	**Elastic**
delay	sensitive	tolerant
jitter	sensitive	tolerant
bandwidth	sensitive	sensitive
loss of data	tolerant	sensitive

Tabella 4.4 – Traffico “real time” vs traffico “elastic”

La tabella appena mostrata riassume le caratteristiche generali del traffico “real time” e di quello“elastic”.

Nell'ambito delle classi di traffico appena definite, il modello Integrated Service comprende tre classi di servizio:

• ***Best effort service (BE)***

• ***Guaranteed service (GS)***

• ***Controlled load service (CLS)***

Sostanzialmente il primo tipo di servizio, quello "BEST EFFORT" rientra nella classe di traffico di tipo "elastic", mentre le altre due classi di servizio, la "GUARANTEED SERVICE" e la "CONTROLLED LOAD" possono trovare collocazione nella classe di traffico di tipo "real time", differendo tra loro sostanzialmente per la priorità con cui il servizio viene portato a termine: maggiore priorità è concessa alla GS.

La tabella di seguito mostra le principali differenze tra la "Guaranteed Service"e la "Controlled Load".

	Guaranteed Service	**Controlled Load Service**
Applicazioni	• applicazioni in tempo reale	• applicazioni sensibili alla congestione di rete • applicazioni non in tempo reale con vincoli meno stringenti di ritardo
Comportamento end-to-end	• ritardo massimo garantito • throughput garantito • nessuna perdita causata da overflow	• di tipo Best Effort in una rete poco carica

Tabella 4.5 - Classi di servizio a banda garantita nel modello Integrated Service

Come si può vedere, per la classe "GUARANTEED SERVICE" a priorità più alta, è imposto un requisito stringente sul ritardo e sull'assenza di overflow nei buffer: tali vincoli sono rispettati riservando una certa quantità di banda R.

Per la classe “CONTROLLED LOAD” a priorità più bassa, non può essere garantita l’assenza di trabocco nei buffer pur riservando una certa quantità r di banda, e viene quindi solamente imposto un limite superiore al ritardo massimo tollerabile [Blefari-Melazzi et altri].

Questo lavoro mira alla definizione di un MAC che supporti sia la classe di servizio a qualità garantita che quella best effort, non considerando l’ulteriore suddivisione appena menzionata tra la classe GS e quella CL.

4.2.1 La classe di servizio Guaranteed Bandwidt (GB)

La classe di servizio GB comprende, come già accennato prima, quelle applicazioni che hanno requisiti stringenti sia sul ritardo di trasferimento complessivo da estremo a estremo, sia sulla variazione (jitter) dei ritardi subiti dalle unità informative appartenenti a tali servizi.
Oltre a questo, il grado di integrità informativa è un altro parametro che deve essere garantito. Voce e video sono normalmente applicazioni che rientrano in questa classe di servizio in quanto i relativi codificatori producono un flusso continuo di dati che deve essere riprodotto a destinazione in modo consecutivo ed entro un limite massimo temporale. Per fare un esempio, se i dati video non sono trasferiti nell’intervallo di tempo definito, ciò ostacola “l’utilità” del frame video successivo, nel senso che la visione del filmato non è lineare ma procede a scatti.
Analogamente, per quanto riguarda la voce, un trasferimento che non rispetti gli usuali limiti che stanno intorno ai 200 ms di ritardo non consentono una percezione piacevole della voce.

Di seguito sono riportati esempi di applicazioni appartenenti alla classe GB.

REAL TIME STREAMING APPLICATIONS	BANDWIDTH REQUIREMENTS (AFTER COMPRESSION)
Audio (downstream)	
Qualità CD stereo: 10Hz-20KHz	256Kb/s
Qualità Broadcast: 50Hz-7KHz	64/56/48 Kb/s
POTS(PCM, G 711) 0.2-3.4KHz	64 Kb/s
Low Bit Rate POTS (G 723.1)	6.4/5.3 Kb/s
Low Bit Rate POTS (G 729.A)	8 Kb/s
Video (downstream)	
HDTV (downstream)	Circa 20Mb/s
Video on Demand (MPEG2)	Circa 4-6Mb/s
Video on Demand (MPEG1)	1-2 Mb/s
Low rate videoconferencing (H 263)	Ottimizzato per <28.8 Kb/s

Tabella 4.6 - Esempi di applicazioni GB

Appare ovvia conseguenza, da quanto appena detto, che questa classe di traffico deve essere trattata al fine di garantire la QoS pattuita in termini di banda , ritardo e perdita.
Per ogni flusso sarà necessario allocare un determinato quantitativo di risorse. Tali risorse, nella fattispecie, sono banda e buffer.
La maniera con cui si è provveduto a ciò è stato quella di caratterizzare ogni flusso mediante dei parametri detti "*Traffic Desciptor*", tramite i quali non è solo possibile allocare in maniera opportuna la risorsa, ma è anche possibile gestire meccanismi di controllo di accettazione di flussi. Il protocollo utilizzato per conseguire la qualità di servizio considerando come strato superiore lo strato IP è il protocollo RSVP

(Resource ReSerVation Protocol). Questo protocollo è un protocollo di riservazione della risorsa, che garantisce grazie ad uno scambio di messaggi tra sorgente e destinatario, la riservazione del quantità garantita di banda in ciascun ramo in modo da ridurre al minimo la perdita per code.
Questo meccanismo verrà illustrato in un paragrafo seguente.

4.2.2 La classe di servizio "Best Effort" (BE)

La classe di servizio BE è una classe che si adatta molto bene alla descrizione di servizi che non hanno alcun requisito di tempo reale e richiedono, solamente, il soddisfacimento dei vincoli sull'integrità informativa, questo a causa del fatto che l'esigenza fondamentale di tali servizi è relativa al trasporto dei dati, che deve essere il più possibile esente da errori. Di conseguenza si può dire che l'informazione fa "quanto di meglio" per arrivare a destinazione e lo fa nel minor tempo possibile. Un tipico esempio è il traffico Internet (Web Browsing, trasferimento di file e posta elettronica).

All'interno della classe di servizio BE è opportuno, tuttavia, fare una distinzione tra applicazioni che presentano si requisiti meno stringenti rispetto a quelli della classe GB sul ritardo, ma che comunque, sono molto legate al tempo con cui il servizio viene espletato e servizi per i quali il tempo necessario al trasporto dell'informazione a destinazione non ha alcuna rilevanza.

Appartengono alla prima delle due sottoclassi le funzioni di rete tradizionali, come il trasferimento di file, che pur non essendo sensibili al ritardo sono comunque legate al tempo che l'utente è disposto ad aspettare prima che il servizio sia completato.

Nella seconda sottoclasse possono invece essere considerati gli usuali servizi di posta elettronica.

Dal momento che per tali applicazioni deve essere assicurato un ottimo livello di integrità informativa e le perdite di informazioni degradano notevolmente il servizio offerto, l'obiettivo primario della classe BE è quello di assicurare un servizio che sia esente da errori.

Questo scopo può essere raggiunto considerando meccanismi di FEC e ARQ o un misto tra i due in modo tale da raggiungere un compromesso tra utilizzazione della risorsa radio e vincolo sull'integrità informativa stessa.

A tutt'oggi si sta cercando un modo di trattare le unità dati appartenenti ai vari servizi e che prevedano la trasmissioni di blocchi più o meno lunghi e con certi requisiti di trasferimento. Tuttavia, mentre può essere valutata seriamente l'idea di utilizzare un protocollo di riservazione di risorse per quelle applicazioni come FTP, vincolate al ritardo di trasferimento lato utenza (tempo di espletamento del servizio), per altre applicazioni come telnet, invece, non ha molto senso usare procedure di segnalazione al trasmettitore e al ricevitore poiché queste richiederebbero alla rete un impegno, anche se per breve tempo, per trasferire blocchi di dati di poche centinaia di bytes. Il sistema di accesso non può garantire alcuna qualità di servizio se non con l'instaurazione della sessione.

La classe BE, oltre a consentire un corretto trattamento delle informazioni provenienti da Internet che non impiegano RSVP, consente di ottenere un'efficiente utilizzazione della risorsa. Il fatto di dover garantire dei flussi alla classe GB non permette di tenere in conto per questa del guadagno statistico che si otterrebbe mediante la multiplazione dei flussi, ma una gestione che accanto ai flussi a qualità garantita consideri anche flussi di tipo BE, consente di riempire quelle "carenze" di risorsa lasciata dai flussi GB. Gli algoritmi di scheduling proposti nei paragrafi seguenti consentono esattamente di ridurre per quanto possibile gli sprechi di risorsa.
Di seguito è riportata una tabella che riassume tipi di servizio e requisiti delle classi di servizio a banda garantita e Best Effort.

	Tipo di Servizio	Requisiti
Guaranteed Bandwidth GB	applicazioni in tempo reale audio, video, telefonia via internet, conferenza multimediale, giochi interattivi etc.	Trasferimento in tempo reale (vincolo sul massimo ritardo e sul jitter di ritardo). Probabilità di perdita non nulla. Probabilità di errore sul bit (BER) di 10^{-3}, 10^{-5}.
Best Effort BE	Web browsing, trasferimento di file FTP, chatting, E-mail	Traffico senza esigenze di tempo reale o con vincoli molto larghi sul ritardo. Probabilità di perdita molto bassa (10^{-8}). Probabilità di errore sul bit (BER) di 10^{-7}.

Tabella 4.7 – Classi di servizio del MAC

4.3 SUPPORTO DELLA QoS E ALLOCAZIONE DELLA RISORSA

Nel concetto di qualità del servizio è già presente la distinzione chiave dei servizi i servizi con prestazioni garantite e non. Nel primo caso l'operatore fornisce un servizio garantendo determinati livelli prestazionali, nel secondo, si limita fondamentalmente a fare quanto di meglio sia possibile, non escludendo la possibilità che lo scambio delle informazioni avvenga con le prestazioni volute. La qualità del servizio percepita dal generico utente può essere classificata in due diverse categorie:

– Qualità del Servizio a livello di chiamata ;

– Qualità del Servizio durante la fase di trasferimento dell'informazione.

La prima va definita all'interno di reti, con connessione o senza, nelle quali il singolo utente prima di poter trasferire informazioni, deve instaurare una connessione e quindi chiedere un'autorizzazione. In altre parole prima di iniziare il trasferimento dati vero e proprio l'utente è sottoposto ad un controllo di accesso. In presenza di un controllo di questo tipo, c'è interesse nel valutare con quale probabilità una richiesta di chiamata sia accettata o rifiutata. La Qualità del Servizio di chiamata misura tale aspetto. Un predeterminato livello di questo tipo di QoS può essere dato solo in termini probabilistici, dal momento che non si può prevedere in generale quanti utenti esprimeranno una richiesta di servizio. Il caso in cui non sia presente controllo di accesso, come accade per Internet, può essere visto come caso limite in cui la QoS di chiamata è sempre garantita, nel senso che la rete non oppone mai rifiuto ad una richiesto di servizio.

La qualità del servizio di trasferimento fa invece riferimento alle prestazioni percepite dagli utenti durante la fase di trasferimento, tipicamente espressa in termini di parametri prestazionali come ritardo, variazione del ritardo, tasso di perdita delle unità informative, throughput. La QoS di trasferimento può essere caratterizzata da prestazioni garantite solo in presenza di controllo di accesso. Il modo in cui le risorse vengono allocate determina le caratteristiche prestazionali di una infrastruttura di telecomunicazioni.

4.3.1 Allocazione delle risorse secondo il modello Dual Leaky Bucket

Analizzate le sorgenti e caratterizzate le classi di servizio, bisogna individuare quali sono i descrittori di traffico grazie ai quali si caratterizzano le sorgenti dal punto di vista della rete.

E' importante a questo punto definire le strategie che sono alla base del MAC e le

scelte che sono state fatte per garantire la qualità del servizio richiesta alla rete.

Il modello Integrated Services è stato scelto al fine di fornire un servizio garantito a certi flussi di traffico.

Il protocollo RSVP, già accennato precedentemente, è il protocollo utilizzato per comunicare le richieste e per riservare le risorse. Questa comunicazione preventiva ha lo scopo di assicurare ad un certo flusso informativo una quota parte di banda e di memoria nei buffer, in modo da assicurare il ritardo di trasferimento da estremo ad estremo richiesto ed evitare perdite di dati. La riservazione delle risorse avviene tramite dichiarazione da parte delle applicazioni coinvolte di un certo numero di parametri di traffico. Per specificare il traffico offerto, il protocollo RSVP assume che quest'ultimo venga regolato in modo deterministico tramite il modello Dual Leaky Bucket. Tale modello è usato per caratterizzare le specifiche del contratto di traffico (Tspec), fornisce una descrizione concisa del carico offerto da ciascun flusso nonché i parametri di policing.

Il suo ruolo fondamentale è sostanzialmente quello di regolare la generazione di traffico di una sorgente fungendo da "TRAFFIC SHAPER" (sagomatore di traffico).

I parametri da dichiarar nel contratto di traffico sono quelli che seguono

- il ritmo medio sostenibile o *token rate r,*
- la tolleranza di burst o *token buffer dimension b,*
- il ritmo di picco *P,*
- la massima dimensione del pacchetto *M,*
- il valore minimo di dati soggetti a controllo *m.*

ma per poter caratterizzare in maniera completa un flusso è sufficiente far riferimento ai primi tre: ritmo medio sostenibile, tolleranza di burst e ritmo di picco.

Per essere in grado a partire da una conoscenza più o meno dettagliata della sorgente, di dimensionare la banda R da assegnare a ciascun flusso e lo spazio di buffer, è necessario fare riferimento al vincolo imposto sul ritardo massimo di trasferimento di un pacchetto da estremo ad estremo che deve essere minore di un certo valore massimo

e al fatto che si vuole ottenere una probabilità di overflow nei buffer nulla.

E' necessario allo scopo definire una maschera che rappresenti l'andamento nel tempo dei bit emessi.

Questa viene definita con il numero totale di bit emessi da una sorgente in un intervallo di tempo [0,t] . Con la funzione $A^*(\tau)$ si individua il numero massimo di bit emessi da una sorgente in una finestra temporale estesa tra [t,t+τ].

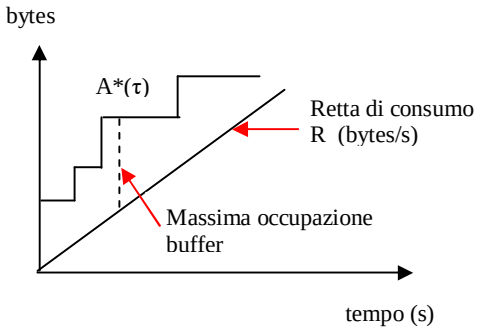


Fig.4.7 – Inviluppo $A^*(\tau)$ insieme alla retta retta di consumo

Il massimo numero di bit trasmessi da un canale di capacità R in un intervallo temporale τ è rappresentato dalla retta R*τ. Di conseguenza, dalle considerazioni fatte precedentemente, si deduce che il numero di bit non trasmessi nel caso in cui si voglia trasmettere il flusso attraverso il canale di capacità R è dato proprio da:

$A^*(\tau) - R^*\tau$.

Volendo evitare un trabocco dei buffer con conseguente perdita di dati, occorre dimensionare il buffer in modo che :

$B \geq MAX_{\tau}\left(A^*(\tau) - R'^*\tau\right)$

Se B è il massimo riempimento del buffer che si verifica nelle condizioni indicate, essendo il buffer di coda FIFO, risulta che affinché l'ultimo bit presente in coda sia trasmesso da un canale di capacità R il tempo di coda sarà espresso dalla seguente equazione:

$t_{CODA} = \left(MAX_{\tau} \left(A^*(\tau) - R^*\tau \right) \right) / R$

Se il sistema introduce ulteriori ritardi la cui somma sia minore di $t_{MAXSISTEMA}$, allora affinché il ritardo massimo totale sia minore del massimo ritardo da estremo a estremo, deve essere verificata l'equazione che segue:

$t_{CODA} + t_{MAXSISTEMA} \leq D_{MAX}$

In particolare sarà possibile determinare R dalla condizione di eguaglianza, visto che l'unica incognita è R stessa. Pertanto la soluzione dell'equazione è banale.Questa trattazione matematica per il calcolo della banda da attribuire ad un flusso è corretta, ma il vero problema, che si riscontra nella realtà, è quello di non avere una conoscenza esatta della maschera di emissione della sorgente in un preciso intervallo temporale.

E' per questo che è necessario determinare una maschera di flusso che definisca un numero di bit maggiore o uguale al numero di bit trasmessi nella finestra temporale τ. Per far questo dalla prima maschera considerata $A^*(\tau)$, si passa ad una seconda maschera, $M(\tau) \geq A^*(\tau)$, che provvede attraverso una descrizione approssimata del flusso con i parametri di Dual Leaky Bucket al calcolo di un valore di banda da dedicare a quel flusso, $R' > R$, che sia il più possibile prossimo al valore di R che si sarebbe ottenuto considerando la vera curva di inviluppo. Questo, al fine di evitare sovrastima di banda e dello spazio da allocare nel buffer.

La condizione che dovrà essere verificata per rispettare il vincolo sul ritardo può essere riscritta come segue:

$\left(MAX_{\tau} \left(A^*(\tau) - R'^*\tau \right) \right) / R' + t_{MAXSISTEMA} \leq D_{MAX}$

I parametri di Dual Leaky Bucket, fino a questo momento non interessati nei calcoli, forniscono ora un'approssimazione lineare a tratti della maschera di flusso $M(\tau)$.

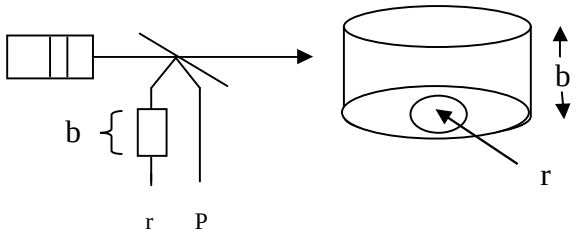


Fig.4.8 – Modello Dual Leaky Bucket

Considerando la coppia (b,r), r rappresenta la portata del foro del contenitore cilindrico e b la sua capacità. Il contenitore è in grado di accettare non più di b bit insieme e qualora sia pieno può essere riempito ad un rate massimo pari a r.
Tornando alla rappresentazione grafica, quanto appena detto è supportato dal grafico di fig.4.9.

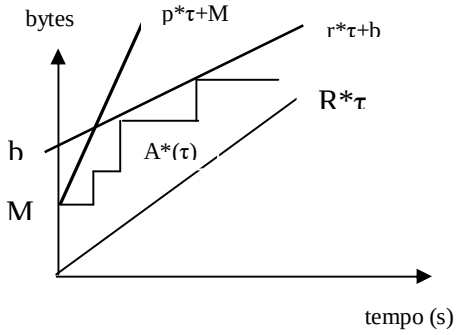


Fig.4.9 – Determinazione dei parametri Dual Leaky Bucket a partire dall'inviluppo

Le due rette $p*\tau+M$ e $r*\tau+b$ ci dicono che finchè è possibile, la sorgente trasmette al picco raggiunto il punto di gomito, $\tau=\frac{(b-M)}{(p-r)}$, in cui si raggiunge la massima distanza tra la maschera DLB e la retta passante per l'origine, il ritmo di emissione diventa r.
La funzione minimo delle due rette

$M(\tau) = \min(M + p*\tau, b + r*\tau)$

rappresenta esattamente la maschera individuata dai parametri DLB con la condizione che $r < p$ e $M < b$.

Sostituendo alla maschera denominata $M(\tau)$ la funzione minimo tra le due rette e sostituito il tutto all'interno dell'equazione sul ritardo di trasferimento si ottiene:

$$\left(MAX_{\tau}\left(\min\left(p*\tau + M, r*\tau + b\right) - R'*\tau\right)\right)/R' + t_{MAXSISTEMA} \leq D_{MAX}$$

La funzione MAX a primo membro restituisce un valore finito se solo se R'>r.

Sostituendo a questo punto il ritardo massimo di trasferimento da estremo a estremo con il ritardo massimo tollerabile e il ritardo di sistema con quello che si suppone introdotto dal sistema in esame si ricava che:

$$\left(\frac{r*(b-M)}{(p-r)} + b - R'*\frac{(b-M)}{(p-r)}\right)*\frac{1}{R'} + D_{SISTEMA} \leq D_{MAXTOL}$$

e risolvendo rispetto ad R' si ottiene

$$R' \geq \frac{p*b - r*M}{\left(D_{MAXTOL} - D_{SISTEMA}\right)*(p-r) + b - M}$$

che posto a sistema con R'>r, porta a:

$$R' \geq MAX\left(\frac{p*b - r*M}{\left(D_{MAXTOL} - D_{SISTEMA}\right)*(p-r) + b - M}, r\right)$$

La massima distanza tra la maschera DLB e la retta di consumo di pendenza R' rappresenta la massima occupazione del buffer. Da questo segue che, affinché non ci siano perdite nel buffer, deve essere allocato uno spazio di memoria B, che rispetti la $B \geq MAX_{\tau}\left(A*(\tau) - R'*\tau\right)$ e sia :

$$B \geq \frac{p*b - r*M}{p-r} - R'*\frac{b-M}{p-r}.$$

Come detto precedentemente, la scelta dei parametri Dual Leaky Bucket deve rispettare i vincoli imposti sulla maschera. La scelta deve mirare quindi a trovare coppie di valori b, r e p ed M , che siano le migliori.

Se si considerano due maschere, l'una inclusa nell'altra e si dimensiona il sistema a partire dalla maschera più interna, è chiaro che le risorse calcolate non possono che essere minori od uguali a quelle della maschera più esterna [White P.P.,1997].

Nella caratterizzazione del protocollo RSVP, sia per quanto riguarda il calcolo della banda che il dimensionamento dei buffer, vengono introdotti due valori aggiuntivi, *C* e *D* che sono termini correttivi necessari per considerare deviazioni dal modello fluidico.

Il ritardo massimo tollerabile sarebbe in tal caso pari a :

$$D_{MAXTOL} = \frac{(b-M)(p-R')}{R'(p-r)} + \frac{M+C}{R'} + D \qquad p \geq R' \geq r \qquad \text{(a)}$$

$$D_{MAXTOL} = \frac{M+C}{R'} + D \qquad R' \geq p \geq r \qquad \text{(b)}$$

Dal punto di vista del ricevitore, volendo un ritardo massimo tollerabile pari a D_{MAXTOL}, il ritmo R' (in bytes/s) che deve essere riservato da chi trasmette si ricava direttamente da (a) o (b):

$$R' = \begin{cases} \dfrac{p*(b-M)+(M+C)*(p-r)}{(D_{MAXTOL}-D_{SISTEMA})*(p-r)+b-M} & p \geq R' \geq r \\ \dfrac{M+C}{(D_{MAXTOL}-D_{SISTEMA})} & R' \geq p \geq r \end{cases}$$

Per quanto riguarda la dimensione dei buffer, tenendo conto dei parametri correttivi la formula che esprime il valore di B è:

$$B = \frac{(b-M)(p-R')}{p-r} + M + C + R'*D$$

C’è da aggiungere che, dal momento che vengono trattati aggregati di flussi, in luogo di $r, p \; e \; b$ si considereranno i valori dati da :

$$\overline{r} = \sum_i r_i \qquad bytes$$

$$\overline{p} = \sum_i p_i \qquad bytes / s$$

$$\overline{b} = \sum_i b_i \qquad bytes / s$$

[Schmitt,1998].

4.3.2 ReSource reserVation Protocol

RSVP è un protocollo di prenotazione usato in router di tipo Integrated Services per garantire una predeterminata QoS nelle reti IP.

Tale protocollo è stato progettato per consentire a mittenti, destinatari e routers di comunicare tra loro tramite messaggi di controllo, per stabilire opportuni stati e modalità di trattamento dei datagrammi all’interno dei router, in modo da poter supportare le classi di servizio definite precedentemente. I messaggi di controllo stabiliscono lo “stato” dei nodi della rete e lo “stato” di un nodo determina la politica di allocazione di risorsa da perseguire per i diversi flussi di comunicazione che attraversano il nodo stesso. Affinché questi messaggi di controllo si propaghino in modo corretto tra sorgente e destinatario, i router partecipanti alla sessione devono in qualche modo conoscere i percorsi correnti nella rete. Le tabelle di routing che definiscono tutti percorsi da sorgente a destinazione, vengono aggiornate ad intervalli regolari [Blefari-Melazzi et altri]. Assumendo che esista un solo nodo sorgente, i percorsi che si vengono ad originare da questo verso tutte le destinazioni, formano una sorta di “albero” detto Distribution Tree (RFC 1190), in cui la sorgente è il nodo radice, la destinazioni sono le foglie ed i routers sono i nodi intermedi.

Il RSVP definisce sei tipi di messaggi:

Tipo di Messaggio	Mittente del messaggio	Significato del Messaggio
PATH REQUEST	trasmettitore	Attivazione di uno stato di "waiting for session setup". Fornitura di informazioni al destinatario circa le caratteristiche di traffico di sorgente e del percorso da estremo a estremo in modo che il destinatario possa effettuare richiesta opportuna ai router lungo tutto il percorso da sorgente a destinazione
PATH ERROR	router lungo il cammino	Generazione d'errore scaturita da una richiesta di PATH
PATH TEARDOWN	trasmettitore	Accelera la liberazione delle risorse
RESV (reservation) request	ricevitore	Originato dall'utente destinatario del flusso. Il suo scopo è quello di portare la richiesta di riservazione della risorsa ai router lungo il percorso da sorgente a destinazione
RESV ERROR	router lungo il percorso	Rifiuto di riservazione della risorsa (o perché non disponibile o perché è fallita l'autorizzazione)
RESV TEARDOWN	ricevitore	Accelera la liberazione delle risorse

Tabella 4.8 - Messaggi RSVP

Dal momento che RSVP è un protocollo iniziato dal destinatario, e dal momento che i destinatari che desiderano una certa QoS necessitano di riservare risorse lungo uno specifico percorso, essi devono ricevere messaggi di PATH su quel percorso dalla sorgente.

Il messaggio di PATH presenta le seguenti informazioni:

- *Indirizzo di destinazione*
- *Phop*, indirizzo dell'ultimo nodo in grado di supportare RSVP che lo ha inoltrato

- Un campo chiamato *Sender Template* che identifica la sorgente
- Il *descrittore del traffico* di sorgente
- *Adspec* (opzionale) che identifica le caratteristiche del cammino da estremo a estremo, può essere usato dai destinatari per conoscere in modo più preciso il livello di prestazioni disponibili e le risorse di cui chiederanno l'allocazione

Ogni router che riceve un messaggio di PATH ne controlla la validità, in caso negativo risponde con un messaggio di PATHERROR a colui dal quale lo ha ricevuto.
Altrimenti crea una variabile di stato in cui memorizza le specifiche di traffico, il prossimo HOP ed eventualmente l'Adspec. Associato ad ogni stato di PATH c'è un timer di reset , allo scadere del quale lo stato di path è cancellato esplicitamente, mediante messaggio di tear-down. La cancellazione è evitata se e solo se è ricevuto un nuovo messaggio di PATH prima dello scadere del timer (*refresh timeout*). Alla ricezione del messaggio di path il timer è riavviato. Questo meccanismo dello stato di connessione avviato dal messaggio di path è detto "*soft*" RSVP e serve di supporto a situazioni di crash della rete.

Le richieste di PATH e le richieste di riservazione saranno ripetute regolarmente, ogni intervallo di "refresh". Il messaggio RESV è inviato dall'utente destinatario del flusso ed il suo scopo principale è quello di portare le richieste di riservazione ai router lungo il percorso da sorgente a destinatario. Il messaggio di RESV contiene un descrittore del traffico di sorgente (*Tspec*) che è generalmente posto uguale al Tspec della sorgente, e la caratterizzazione della "riservazione" (*Rspec*), che comprende banda calcolata R da "riservare" ad ogni router e un termine di correzione detto Slack Term che serve per tener conto di eventuali margini del sistema per il ritardo da estremo a estremo rispetto a quello tollerato dall'applicazione. Quando un router RSVP riceve un messaggio di RESV, fornisce le informazioni Tspec e Rspec al proprio controllo di ammissione, questo controlla se è disponibile sufficiente banda e spazio nei buffers per soddisfare la richiesta. L'accettazione della richiesta comporta anche un eventuale modifica della

coppia di parametri ricevuti Rspec da (Rin, Sin) in (Rout, Sout) secondo le regole del protocollo. Se la riservazione è accettata allora è inviato un nuovo messaggio di RESV al router che lo precede nel percorso da sorgente a destinatario. Come lo stato di PATH anche quello di riservazione è memorizzato nei router in maniera temporaneo. Se non c'è un aggiornamento periodico, la cancellazione dello stato tramite il messaggio RESV TEARDOWN avviene allo scadere del timer.

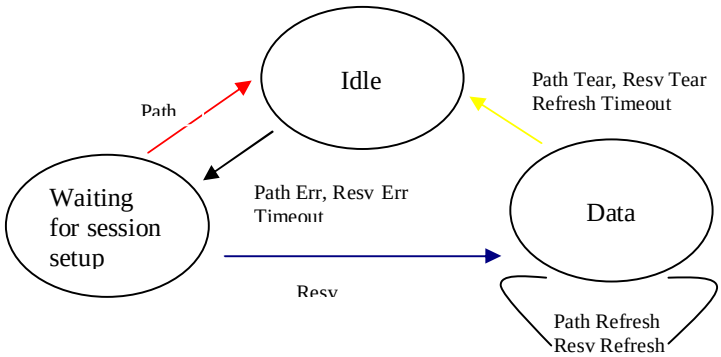


Fig.4.10 – Diagramma di stato per l'instaurazione di una sessione RSVP

Nel caso del sistema sotto esame e facendo riferimento alla tratta di Downlink, quando un messaggio di PATH viene inviato da una sorgente remota, arriva tramite il router di accesso al RN. Questo consulta la tabella di routing nella quale l'indirizzo di ogni RT è associato a tutti gli indirizzi di tutti gli utenti attestati allo stesso RT.

Se l'indirizzo IP non viene trovato, il RN invia un messaggio di PATH ERROR, altrimenti il messaggio di PATH viene segmentato e mandato in Downlink come un set di MACPDU destinate al RT a cui l'utente è connesso. A questo punto l'RT se ha ricevuto correttamente il messaggio lo riassembla e lo invia all'utente destinatario. Ricevuto il messaggio di PATH, se l'utente finale decide di accettare l'instaurazione della sessione manda un messaggio di RESV al RT, il quale riserva risorse tra lui e l'utente come specificato dal protocollo RSVP, richiedendo al RN l'assegnazione di banda per inviare le MACPDU in cui è stato segmentato il messaggio. Se la trasmissione del messaggio di RESV avviene prima dello scadere del timeout dello

stato di Path attivato dal RN, l'RN riserva banda e buffer all'interfaccia radio e inoltra il messaggio di RESV verso la sorgente remota.

4.3.3 ADMISSION CONTROL

Lo scopo dell'Admission Control proposto è accelerare la decisione se accettare un nuovo flusso nella rete o non. Comunemente, gli organi di controllo della rete mirano a riservare capacità per un nuovo flusso nel momento in cui le richieste di un utente siano accettate. Tale capacità non è sotto-utilizzata se non usata dai flussi con riservazione, dal momento che il surplus può essere sfruttato dai flussi BE.
L'accettazione di un flusso deve rispettare poche semplici regole.
Indicata con R_{req} la capacità richiesta da un nuovo flusso, con R_{ex} la capacità già allocata e con R_{res} la capacità riservata preallocata, l'organo di controllo di rete, all'atto dell'inserimento della nuova sorgente nel sistema, deve verificare che sia rispettata la seguente disuguaglianza:

$$R_{req} + R_{ex} \leq R_{res}$$

Se la disuguaglianza è rispettata allora il nuovo flusso è accettato e la quantità di capacità allocata è aggiornata al valore $R_{es} = R_{es} + R_{req}$, altrimenti il nuovo flusso non viene accettato [Bolognesi,1999].
Come visto nel paragrafo precedente, relativo al calcolo della banda necessaria ad un flusso, avevamo indicato con R' la banda allocabile ad esso, misurata a partire dai parametri descrittori del traffico. Sulla base delle premesse fatte, per l'insieme dei flussi attestati al RN e diretti verso ciascun RT (Downlink), le condizioni che devono essere verificate sulla banda allocata per attuare un controllo di ammissione sono le seguenti:

Classe GB $\begin{cases} \sum_{i=1}^{N_{GB}\in RN} R_i \le C_{DL,netta} - C_{\min,DL} \\ \sum_{i=1}^{N_{GB}\in RN} r_i \le C_{DL,netta} - C_{\min,DL} \end{cases}$

Le somme mostrate in formula sono l'esplicitazione di $R_{req} + R_{ex}$, mentre con $C_{DL,netta}$ si considera quella variabile che ho indicato con R_{res}.

Le precedenti disuguaglianze sono da intendersi alternative. Il calcolo della banda da allocare ricordiamo era espresso dalla relazione:

$$R' \ge MAX\left(\frac{p*b - r*M}{\left(D_{MAXTOL} - D_{SISTEMA}\right)*\left(p-r\right)+b-M}, r\right)$$

pertanto il suo valore per ogni flusso sarà o R_i o r_i. Le due sommatorie indicate nella formula intendono semplicemente mostrare che il confronto viene fatto tenendo conto di volta in volta dell'effettiva banda allocata, qualunque essa sia.

Una volta stabilita la capacità allocata alla classe di servizio GB, la restante capacità è allocabile alla classe BE.

Ne segue che la banda complessivamente disponibile per i flussi di tipo Best Effort è pari a :

$$C_{BE} = C_{DL,netta} - \sum_{i=1}^{N_{GB}} F_{m,i} + C_{\min,DL}$$

Le equazioni descritte sono da intendersi nella seguente maniera, con $C_{DL,netta}$ si indica la capacità in DL, dell'interfaccia radio vista dallo strato 3, a cui è stata sottratta l'informazione di segnalazione (per questo è netta), con $C_{\min,DL}$ si è indicata la banda che eventualmente è riservata in DL per la classe BE. Infine con $F_{m,i}$ si è indicata la banda media che il singolo flusso accettato e diretto in DL ha effettivamente utilizzato.

Il termine indicato con $C_{\min,DL}$ viene utilizzato nel sistema, solo in determinate occasioni. Vediamo di chiarire brevemente questo concetto. La disciplina di

Scheduling adattiva, illustrata nel cap.5, mostra come sia possibile implementare un'assegnazione delle risorse variabile in funzione del carico del sistema. L'allocazione differente si basa su assegnazione di banda via via differenti. E' previsto infatti un meccanismo di accettazione che a seconda del traffico presente in rete, garantisca a volte una quantità di banda certa al traffico Best Effort, mentre altre volte fa del suo meglio per allocare il disponibile. Questa procedura di accettazione anche per la Best Effort è implementata all'interno dell'algoritmo di scheduling stesso e si prefigge lo scopo di non degradare troppo la classe BE in condizioni di carico GB, molto pesante, che non lasci spazio ad allocazione delle risorse per la BE. Questo è il motivo per cui nella formula precedente si è considerato il termine $C_{\min,DL}$.

CAPITOLO 5

STRATEGIE AVANZATE DI SCHEDULING

5.1 PRINCIPI DI SCHEDULING

Molte applicazioni multimediali (quali, ad esempio, audio e video conferenze) richiedono alla rete di provvedere ad una grande quantità di requisiti sulla qualità di servizio (in relazione a banda, ritardo di pacchetto, delay, jitter e perdita).
Per abilitare la rete a provvedere a tutto ciò, le sorgenti specificano le proprie caratteristiche di traffico. La rete, dal canto suo, provvede a rispettare i requisiti imposti mediante l'utilizzo dei protocolli descritti nel capitolo precedente e riservando risorse e provvedendo allo scheduling di queste nella maniera migliore possibile.
Un buon protocollo MAC deve essere in grado di gestire ritardi dei pacchetti non elevati, di garantire soddisfacenti valori di throughput, di concedere accesso al mezzo ai diversi nodi, qualora questi siano caratterizzati da requisiti di traffico con differenti priorità. Inoltre il MAC deve essere robusto nei confronti del fading di canale ma soprattutto deve poter supportare traffico multimediale.
Con la convergenza nella rete di voce, video e dati, è al momento necessario per i protocolli MAC gestire un traffico di tipo multimediale.
I protocolli richiedono meccanismi di trattamento dei pacchetti delle varie applicazioni diversi a seconda dei vincoli sul ritardo. Due meccanismi comuni per il trattamento dei pacchetti sono le priorità di accesso e lo scheduling.
Le priorità di accesso provvedono un servizio differenziato consentendo a certi nodi di ottenere accesso ai servizi di rete con più alta probabilità di altri.

Lo Scheduling può fornire invece garanzie di banda, ritardo e jitter e consiste sostanzialmente in una allocazione intelligente delle risorse, che permetta di incontrare le garanzie di QoS richieste dall'utenza.
Il sistema che noi consideriamo, come già ampiamente descritto nel Cap.3, è una rete wireless indoor di tipo centralizzato. Protocolli MAC centralizzati sono protocolli dove l'arbitraggio e la complessità si trovano in quello che noi abbiamo classificato come RN (Radio Node).
Il RN ha il completo controllo sull'utente e sull'istante nel quale questo accede al mezzo e sostanzialmente è proprio il RN a decidere a chi fornire risorse e con quale priorità assegnarle.
Una disciplina di scheduling serve a risolvere i problemi di contesa, definendo l'ordine secondo il quale le richieste debbano essere servite e fornisce lo spunto per allocare le risorse in modo equo rispettando le garanzie richieste.
Un algoritmo di scheduling, in generale, è caratterizzato dai seguenti aspetti:

1. Scelta tra procedura *work-conserving* e procedura *non work-conserving*
2. Definizione dei livelli di priorità
3. Scelta dell'ordine del serivizio all'interno del livello

Per chiarire il significato e la differenza tra algoritmo work-conserving e algoritmo non work conserving, devono essere introdotti due aspetti della politica di scheduling[Keshav,1997].
Il primo si riferisce alla scelta del tempo di processamento di una richiesta: la sua "esecuzione" può avvenire immediatamente o essere rinviata (*Sleep policy*). Se ci troviamo in una condizione di rinvio è perché

- il carico nel sistema era già alto;
- la richiesta fatta era di bassa priorità ed erano presenti richieste concorrenti ad alta priorità.

Il secondo aspetto fa riferimento al momento in cui riammettere una richiesta rinviata nel sistema (*Wakeup policy*).

Una richiesta rinviata viene riammessa solo nel momento in cui un'altra richiesta è stata soddisfatta. Adempiuta la richiesta, lo scheduling deve decidere quale tra le richieste rinviate deve essere selezionata al suo posto tra quelle presenti.
Se la politica consente alle richieste a più bassa priorità di essere soddisfatte, in caso di richieste a alta priorità non disponibili, la disciplina di scheduling è detta *work conserving.* Altrimenti è detta *non work conserving.*
Una politica work conserving evita di bloccare richieste che devono essere soddisfatte.
Nelle implementazioni naturali degli algoritmi di scheduling, le politiche di Sleep e Wakeup sono state implementate in modo tale che alle richieste che possono essere soddisfatte subito sono assegnate immediatamente un certo numero di risorse, le altre vengono messe in attesa in opportuni buffer per essere soddisfatte non appena possibile.

Non-Work Conserving Policy

Supponiamo che esistano richieste di tipo A e richieste di tipo B, che rappresentino le richieste delle classi di servizio GB e BE. La politica non-work-conserving permette alle richieste A o B di occupare solamente le risorse che sono disponibili per quel tipo, consentendo quindi l'esecuzione di una richiesta solo se il numero totale di richieste di A o B è sotto la soglia stabilita durante la fase di contratto utente-rete; le altre richieste, almeno in un primo momento, vengono lasciate a riposo (Sleep policy).

Per fare un esempio, considerando i due livelli di priorità A e B e assumendo che il massimo numero di richieste che possono essere soddisfatte di A è 6 e quello di B è 3, se arrivano tre richieste di tipo A e quattro di tipo B, le tre richieste di A vengono soddisfatte occupando le risorse assegnate ed una richiesta di tipo B è bloccata. Considerando poi la politica di "Wakeup", possiamo accogliere tutte le richieste che non hanno ecceduto i loro limiti. Queste vengono soddisfatte in base alla priorità che le denota ed alla loro età.

Work Conserving Policy

La politica work conserving non permette di lasciare risorse inutilizzate, e consente a B di occupare, ad esempio, le risorse lasciate inutilizzate da A (Sleep policy). Nell'esempio fatto precedentemente, tutte le richieste verranno eseguite: 3 risorse saranno occupate da A ed una risorsa di A sarà occupata dalla richiesta in più fatta da B. Se successivamente arrivano tre richieste di A, in tale scenario, due di loro saranno eseguite, la terza sarà eseguita se la politica di servizio è con priorità, altrimenti bloccata. Per quanto concerne la politica di Wakeup le richieste vengono soddisfatte, come nel caso non work conserving, in base alla loro priorità e alla loro età.

Vengono definiti come algoritmi work conserving gli algoritmi VC (Virtual Clock), GPS (Generalized Processor System) e alcune loro generalizzazioni (come, ad esempio, il Packet-by-Packet GPS e il Self-Clocked Fair Queueing) in grado di allocare bande variabili ai pacchetti di un flusso e mostrare come ciò porti alla banda garantita; tra quelli non conserving invece si può considerare l'algoritmo Stop & Go Queueing.

Livelli di priorità e ordine del servizio

Il supporto di differenti classi di traffico è caratteristica comune a tutti gli algoritmi di scheduling studiati, la distinzione dei flussi è generalmente tra due tipi, real time e non real time, benché ci siano state proposte, tipicamente sviluppate per reti ATM, in cui il numero di flussi supportati dallo scheduling è pari a 5: CBR, real-time VBR, non-real-time VBR, ABR (Available Bit Rate), UBR (Unspecified Bit Rate) [Moorman,1999].

I flussi di traffico appartengono a classi di servizio diverse. Tutti i flussi di traffico che richiedono capacità riservata (GB) vengono trattati separatamente rispetto a quelli senza capacità riservata (BE).

Negli algoritmi di scheduling questa distinzione viene considerata in due modi diversi:

- o si suppone di avere due buffer distinti che mantengano memoria delle richieste da soddisfare per la classe GB e per la classe BE e che vengono svuotati con priorità diverse
- o si suppone l'esistenza di un unico buffer in cui mettere tutte le richieste, distinte perchè marcate con valori di priorità differenti.

Ai pacchetti di classe GB, spesso, vengono attribuite due classi di priorità, nel rispetto di protocolli ma soprattutto di servizi come Integrated Services che definiscono più classi a qualità garantita [rif.cap.4].

Le risorse vengono riservate solo per i pacchetti appartenenti alla classe GB. La porzione di risorsa riservata lasciata libera dalla classe GB a priorità più alta è sfruttata dalla classe GB a priorità più bassa, mentre i pacchetti BE, qualora ci sia ancora risorsa disponibile, occupano solo quella. In tal modo la capacità riservata non viene sottoutilizzata.[Mowbray]

Solitamente il traffico BE non ha alcun peso nello schema di riservazione.

Per quanto visto, un generico algoritmo di scheduling propone un certo numero di livelli di priorità: un pacchetto di un dato livello di priorità è servito se non esistono pacchetti di livello più alto, con rispetto del fatto che pacchetti appartenenti a livelli più alti godono di un ritardo più basso [Zhang,1995]. Usualmente i livelli di priorità sono mappati secondo i ritardi delle diverse classi di traffico:

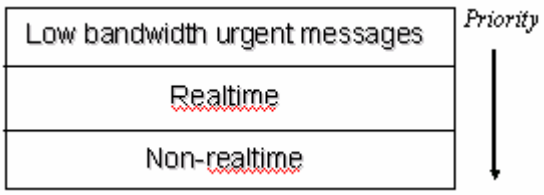


Fig.5.1- Relazione tra priorità e classi di traffico

Lo schema più comune supporta uno scenario in cui l'entità centrale fa una "riservazione" che è sufficiente per ciò che deve essere assolutamente trasferito, tutto il resto viene inviato a priorità bassa nello stesso flusso e trattato come traffico non real time.

E’ compito dell’utente decidere le discipline di priorità all’interno del livello (scelta 4). Solitamente le più utilizzate sono:

- in ordine di arrivo (FCFS)
- in ordine di etichetta di servizio
- in modalità Round Robin (RR)

Algoritmi noti in letteratura

Prima di procedere nella spiegazione dell’ algoritmo di scheduling supportato dal sistema sotto esame, è opportuno introdurre alcuni degli algoritmi di scheduling utilizzati maggiormente nelle reti wireless, considerando i loro aspetti comuni e mostrando come gli elementi caratterizzanti che abbiamo citato in precedenza siano particolarmente importanti nella realizzazione di un buona gestione delle risorse.

La tecnica di scheduling **Virtual Clock** (VC) [Zhang,1995]**,** come dice il nome stesso**,** si basa sul fatto che ogni flusso ha un proprio orologio virtuale. Questo orologio è incrementato ogni volta che un pacchetto appartenente ad un flusso arriva al buffer. Il pacchetto viene marcato con il nuovo tempo del clock e messo in coda. L’ordine con cui i pacchetti vengono tolti dal buffer è secondo un tempo crescente**,** di conseguenza è previsto un riordino della coda all’atto dell’emissione dei pacchetti. La disciplina di servizio è FIFO (o FCFS) per flusso, ma non per flussi aggregati.
L’incremento del clock è pari all’inverso del rate riservato al flusso diviso la lunghezza del pacchetto (per esempio a 3 Mb/s un pacchetto di 375 bytes, deve aumentare il tempo di 1ms). Per quanto concerne il rispetto dei vincoli di contratto, si dice che il rate di un flusso eccede la riservazione quando il suo virtual clock supera il clock di rete.
E’ consentito comunque dall’algoritmo un caso per cui il flusso generico possa eccedere la sua riservazione. Questo accade quando gli altri flussi risultano inattivi.

In questa condizione, l'orologio virtuale del flusso che ha violato la "riservazione" è regolato ad un valore più alto, finchè gli altri flussi non reclamano la loro quota riservata. Questo potrebbe accadere dopo un lungo periodo di tempo. L'eventuale violazione, se accettata, non comporta un trattamento dei pacchetti differente rispetto al caso di non violazione del contratto.

L'algoritmo **Weighted Round Robin** (WRR) è stato pensato allo scopo di allocare diversa quantità di servizio a code diverse. In WRR ad ogni coda è assegnato un peso proporzionale alla quantità di servizio che deve ricevere. Il WRR visita ogni coda in ordine ciclico. Durante ogni visita alla coda, il numero di pacchetti serviti, assumendo la coda non vuota, è proporzionale al peso della coda stessa. Consideriamo tre code di pesi $w_1 = 2, w_2 = 3, w_3 = 1$, se tutte le code hanno fino a 3 pacchetti, lo scheduler WRR servirà 2 pacchetti dalla coda 1, 3 pacchetti dalla coda 2 e 1 pacchetto dalla coda 3, avendo assunto una lunghezza di pacchetto fissa. Basandosi su uno scheduling di tipo round robin, si ottiene una suddivisione equa della banda in relazione all'importanza (peso) della coda. Il WRR si complica notevolmente nel caso in cui oltre al peso assunto dai pacchetti, anche la dimensione degli stessi risulti variabile. Diventa fondamentale in questo caso conoscere la lunghezza dei pacchetti almeno in media.

L'algoritmo **Generalized Processor Sharing** (GPS) è una disciplina ideale di scheduling di tipo work conserving che provvede un servizio equo ad un insieme di code in accordo ai loro pesi. GPS può essere visto come WRR in cui l'unità di servizio, ad ogni visita alla coda, diventa infinitamente piccola (di qui deriva il nome di algoritmo fluidico). Se consideriamo tre code di pesi, $w_1 = 2, w_2 = 3, w_3 = 1$, la più buona approssimazione allo schedulatore GPS è dato proprio dal WRR che serve, come già detto, 2 bits dalla coda 1, 3 dalla coda 2 e 1 dalla coda 3. Ora se immaginiamo che l'unità di servizio diventi piccolissima, potremmo pensare che la coda anziché contenere pacchetti contenga fluido. Il GPS svuota le code simultaneamente a rate proporzionali ai loro pesi, se le code sono, come dichiarato

in letteratura, “backlogged” (non vuote). La quantità di servizio offerta alla coda i-esima sarà chiaramente proporzionale all’importanza della coda stessa. Ciascun flusso viene etichettato al momento del suo ingresso nel sistema. La sua uscita dipenderà non solo dalla sua dimensione ma anche dal peso che gli è stato attribuito.Una definizione più formale dell’algoritmo GPS, si potrà trovare nella caratterizzazione del Weighted Fair Queueing, illustrato di seguito e che ne fornisce una generalizzazione.

L’algoritmo **Weighted Fair Queueing** (WFQ), anche chiamato **Packet by Packet-GPS** [Pareck,1993], è una disciplina di scheduling che approssima la disciplina di servizio GPS. L’idea è servire i pacchetti in ordine dei loro tempi di fine normalizzati in base al peso della coda a cui i pacchetti appartengono

In questo algoritmo proprio come avveniva per il GPS, viene simulato un servizio fluido esente da errori, per qualsiasi tipo di flusso. Ogni pacchetto che entra nel sistema è marcato con due etichette: una di ingresso (di nascita) ed una di fine.

L’etichetta di nascita di un flusso è rappresentata dal massimo che si ottiene tra il tempo di fine servizio per quel flusso e l’istante di fine del flusso precedente ottenuto a partire dall’istante di nascita e dalla lunghezza del pacchetto. Questo calcolo è necessario al fine di evitare sovrapposizione di flussi, in quanto benché un pacchetto possa essere etichettato con un certo valore di fine, il servizio che gli viene fornito richiede un tempo maggiore. In sostanza ogni pacchetto che arriva è marcato con il tempo in cui dovrebbe essere completato il servizio, che possiamo indicare con R(t). Il pacchetto che deve terminare il servizio è sottoposto a scheduling per primo per poi essere trasmesso. Ogni flusso (dove per flusso si intende pacchetto) ha un peso W_i, come nel caso GPS, che gli garantisce una porzione della capacità totale pari a $W_i \Big/ \sum_{j=1}^{n} W_j$, con $\sum_{j=1}^{n} W_j$ pari alla somma di tutti i pesi dei diversi flussi.

Quando un flusso non usa tutta la capacità che gli è stata allocata, questa può essere distribuita secondo una disciplina di tipo round robin tra tutti gli altri flussi che ne necessitano sempre in accordo ai loro pesi. Anche flussi di tipo non real time sono contrassegnati da un peso ed il loro trattamento è paritetico a quello dei flussi GB. Sostanzialmente tutto è regolato tramite i pesi assegnati ai singoli flussi.
La complessità cresce con il numero di flussi da sottoporre a scheduling. All'atto dell'accettazione di un nuovo flusso, tutti i pesi devono essere ricalcolati. La disciplina di servizio anche in questo caso si basa sull'ordine dei tempi di fine del pacchetto.Lo scheduler WFQ sceglie tra i pacchetti nel sistema quello etichettato con il più piccolo tempo di fine. Il WFQ non è molto semplice da implementare in pratica a causa della difficoltà del calcolo del servizio normalizzato R(t), per il quale è necessario mantenere traccia dell'insieme delle code attive nel sistema che possono cambiare molto velocemente. Considerando poi qualche variante al WFQ, si possono citare il **Self Clocked Fair Queueing** (SCFQ) e lo **Start time Fair Queueing** (SFQ) [Bhargavan,1997]. Il primo del tutto simile all'algoritmo appena illustrato evita la necessità di calcolare il servizio usando al suo posto l'etichetta del pacchetto al momento servito. La difficoltà incontrata dal calcolo di R(t) è completamente superata poichè non è necessario mantenere traccia delle code attive, e l'unica informazione necessaria è il tempo di fine del corrente pacchetto di servizio. Il SCFQ è più semplice da implementare del WFQ, e rispetta sempre la caratteristica di equità del WFQ se considerato per scale di tempo non brevi. Il secondo, servito come spunto per la strategia del sistema sotto esame, ha una complessità di implementazione dello stesso tipo di SCFQ, ma si comporta meglio dal punto di vista del ritardo nel caso peggiore. L'idea di base è definire l'istante di inizio del pacchetto ed in base alla lunghezza di questo ed al rate relativo alla sorgente considerata, calcolare l'istante di fine.

L'ultimo algoritmo citato, lo **STOP & GO QUEUING** [Golestani,1996] si differenzia dagli altri in quanto non conserving. In questo algoritmo viene imposto un limite massimo alla dimensione del flusso da schedulare.

Un flusso può mandare al più B bytes di dati in un periodo di lunghezza T (un frame). Un pacchetto che arrivi al frame i-esimo, non verrà servito fino all'inizio del frame successivo.
La lunghezza del frame determina il ritardo subito dal flusso nella rete: un T basso comporta ritardi minori, ma limita la granularità dell'allocazione di capacità, che è un pacchetto per frame. Ciò può essere migliorato considerando dimensioni di frame multiple ma aumentando la complessità del sistema. La caratteristica peculiare di questo algoritmo è il modo in cui vengono trattati i flussi non real time. I pacchetti appartenenti a questo tipo di traffico, sfruttano i momenti di attesa dei pacchetti real time, che per essere serviti debbono aspettare l'inizio del frame successivo.
Il traffico real time viene preparato per essere sottoposto a scheduling nel frame successivo, alla fine di ogni frame. Questo peggiora le performance.
L'algoritmo è semplice da implementare se i pacchetti sono serviti con disciplina FIFO ed è usata una sola lunghezza di frame.

5.2 SCHEDULING STATICO

Nel sistema preso in considerazione, per il progetto della nostra unità di ricerca, le sorgenti a cui si fa riferimento, vengono distinte, inizialmente, in due classi: GB (Guaranteed Bandwidth) o alla classe BE (Best Effort). Di seguito presenteremo l'algoritmo di scheduling in una prima fase di lavoro, considerando un ambiente di lavoro statico.
La "riservazione" delle risorse (BANDA) è fatta esclusivamente per i due tipi di sorgente a qualità garantita : sorgenti ON-OFF e sorgenti CBR.
Le risorse sono rappresentabili, all'interno dell'algoritmo di scheduling, in MAC-PDU/TRAMA. Questo valore è ricavabile a partire dal valore della banda RSVP calcolata al capitolo precedente, sulla base delle caratteristiche della sorgente presa in considerazione.

Come già mostrato nella definizione del sistema, si è ipotizzata una architettura a strati in cui sopra il MAC si trovano lo strato di adattamento e quello IP. Per poter mappare la banda da strato IP a strato MAC è necessario tenere presenti le lunghezze dei pacchetti nei diversi strati, comprensive di overhead di strato e non. Questo porta ad un valore di banda a strato MAC ricavabile dalla seguente formula iterata due volte, per tenere in considerazione lo strato di adattamento:

$$BW_n = BW_{n+1} * \left\{ \left\lceil L_{n+1}PDU / L_nSDU \right\rceil * \left(L_nPDU / L_{n+1}PDU \right) \right\} \qquad bytes/s \text{ [1]}$$

Tale banda, in bytes/s, deve essere convertita nell'unità di misura usata nello scheduling, MAC-PDU /TRAMA. Ciò è ottenuto dalla formula precedente con il seguente calcolo:

$$BW_{MAC} = \frac{BW * NumeroTS * DurataTS}{L_{MACPDU}} \qquad Mac - Pdu / Trama$$

Una volta stabilita la quantità di risorse da assegnare ad ogni flusso, bisogna accertare che queste risorse siano effettivamente consegnate a questo. La quantità di banda assegnabile al traffico a qualità garantita è imposta indipendentemente dall'algoritmo di scheduling vero e proprio, infatti all'atto dell'ingresso nel sistema delle differenti sorgenti, il fornitore del servizio può decidere fino a quando inserire sorgenti di traffico a qualità garantita essendo vincolato ad un massimo consentito, imposto dalla capacità disponibile. Questa operazione si riflette direttamente nell'implementazione della prima versione del simulatore, dove l'utente ha la possibilità di fissare un valore per la variabile "CAPACITA'_DISPONIBILE", che regola, appunto, il limite massimo di MAC-PDU per trama assegnabili alle sorgenti GB.

E' possibile valutare sempre nel simulatore varie statistiche, come ad esempio quelle relative alla portata GB e BE nelle varie situazioni (molto carico GB, molto carico BE e tutte le situazioni intermedie). Lo scheduling preso in considerazione è

[1] $n+1$ ed n , a pedice della lunghezza dei pacchetti, indicano l'appartenza di un pacchetto allo strato superiore o a quello sottostante.

di tipo centralizzato e l'entità preposta all'allocazione delle opportunità a trasmettere in DL ,il Radio-Node (RN), servirà in primo luogo la classe GB, che rappresenta la classe con priorità più alta e, solo successivamente quella BE, che rappresenta la classe a priorità più bassa.

Questo modo di procedere si concretizza nella presenza di 3 steps fondamentali nella definizione dell'algoritmo:

- STEP 1: viene calcolata l'assegnazione delle MAC-PDU "dovute" alle connessioni GB dirette ai vari RT. Ciò viene fatto stabilendo il minimo tra quanto presente nella coda di quel RT e quanta banda gli è stata concessa, sulla base delle sue caratteristiche di traffico.
- STEP 2: Una volta calcolato il numero di MAC-PDU assegnate, viene decrementato un contatore che tiene conto delle numero delle MAC-PDU disponibili in quella trama, nonché lo stato delle code per quel RT. Questo secondo step rappresenta quindi una scansione delle GB, una volta assegnate quelle dovute, che fornisce una sorta di doppia garanzia per il traffico a qualità garantita. A questo punto si procede iterativamente con l'assegnazione di una MAC-PDU ad ogni RT finché non ci siano più risorse da assegnare o non ci siano più richieste GB in attesa. L'assegnazione di queste risorse alla classe GB avviene in modalità Round Robin, assegnando una MAC-PDU alla volta ad ogni RT, qualora questo presenti ancora uno stato delle code diverso da zero, e decrementando, per ogni assegnazione, il valore delle MAC-PDU ancora disponibili nonché lo stato code relativo al RT considerato. Si prosegue in questa maniera finché o il numero delle PDU assegnabili diventa 0 oppure non ci sono più richieste GB da soddisfare. In questo secondo caso, si entra nel terzo step dello scheduling in cui cominciano ad essere assegnate MAC-PDU per la classe BE, aventi priorità più bassa di quelle della classe GB.
- STEP 3: Il meccanismo è identico a quello descritto nella seconda fase: si ha l'assegnazione di una MAC-PDU ad ogni RT finché non ci siano più MAC-PDU da assegnare o non ci siano più richieste BE pendenti. Anche in questo

caso, man mano che avviene l'assegnazione delle MAC_PDU, viene decrementato il numero delle PDU dallo stato code per la classe BE e decrementato il numero delle risorse ancora assegnabili.

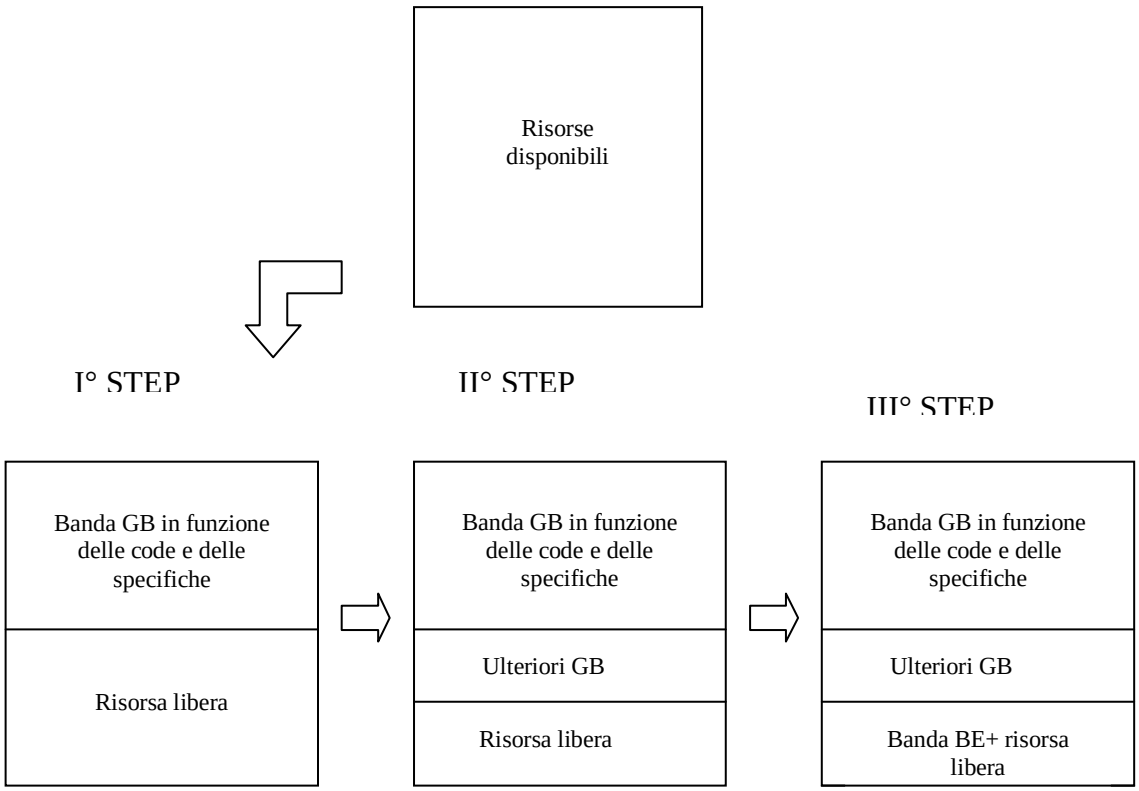


Fig.5.2 - Assegnazione delle risorse

La procedura di scheduling appena descritta mostra una scelta ben precisa per il trattamento delle MAC-PDU di classe GB. Queste sono caratterizzate da una doppia priorità; ciò si riscontra nella doppia scansione delle code GB effettuata nel secondo step dello scheduling. Questo chiaramente va a scapito di eventuali richieste BE, che aspettano in coda , e vengono allocate solo in assenza di richieste GB pendenti.

5.3 SCHEDULING DINAMICO E STRATEGIE ADATTIVE

Le scelte fatte per realizzare uno scheduling adattativo sono servite per supportare quanto segue:

- provvedere in modo opportuno ai cambiamenti del traffico ;
- realizzare, all'interno dello scheduling, un'assegnazione di risorsa variabile in relazione ai parametri che descrivono nel tempo le allocazioni delle PDU di classe GB e BE.

Si è ottenuto un ibrido tra un servizio work conserving e non work conserving sia per i flussi multiplati con riservazione di capacità che per quelli BE.
Il simulatore implementato per verificare le prestazioni dell'algoritmo di scheduling statico ha permesso il monitoraggio della quantità di banda che viene allocata alle sorgenti GB, in particolare la banda in eccesso rispetto a quella assicurata per contratto, legata alla doppia priorità descritta in precedenza. Come risultato di tale verifica, è stata evidenziata una necessità di limitare questo privilegio introducendo, eventualmente, una strategia di allocazione della risorsa in grado di variare nel tempo. A tale propositosono stati presi in considerazione tre differenti algoritmi di scheduling proposti da Acampora per le reti wireless [Acampora]. L'algoritmo di scheduling elaborato a partire da tale riferimento, non solo si adatta alle condizioni di carico della rete, ma provvede anche ad una sorta di equità tra il trattamento dei flussi appartenenti alla classe a qualità garantita e quello dei flussi appartenenti alla classe best effort. Per quanto riguarda questa classe di servizio è stata inserita una ulteriore evoluzione di questo scheduling attraverso la considerazione di due classi di servizio BE, una ad alta priorità, l'altra a più bassa priorità. Chiaramente questo ha portato a dover gestire per la classe BE due code con priorità diverse.

Il meccanismo di allocazione della risorsa reagisce alle condizioni di variabilità del traffico dovuto ai processi di nascita e morte e aggiorna in maniera continua il

traffico ammesso, sia quello GB che quello BE, fornendo una ripartizione della banda radio disponibile in tre modi diversi.

L'algoritmo di scheduling lavora in modo centralizzato ed assicura che la capacità al RN sia efficientemente suddivisa tra le diverse classi di connessioni.

Gli algoritmi di scheduling proposti da Acampora vengono classificati in base all'utilizzo della risorsa che viene fornito e si distinguono in:

- Complete Partitioning (CP)
- Class I Complete Access (CA)
- Class I Restricted Access (RA)

Nel primo schema, CP, la capacità disponibile è completamente partizionata. Le connessioni real time utilizzano una porzione di banda che in figura è indicata con C_I mentre le connessioni non real time utilizzano la porzione di banda rimanente. L'algoritmo è di tipo non work conserving, infatti le risorse lasciate inutilizzate da una o dall'altra classe non possono essere eventualmente assegnate a chi abbia ancora delle PDU da trasmettere in coda. Questo algoritmo, così come mostrato, sembrerebbe non dare garanzie di sfruttamento della risorsa in modo intelligente, ma in realtà ha un ruolo ben preciso, che verrà descritto in seguito.

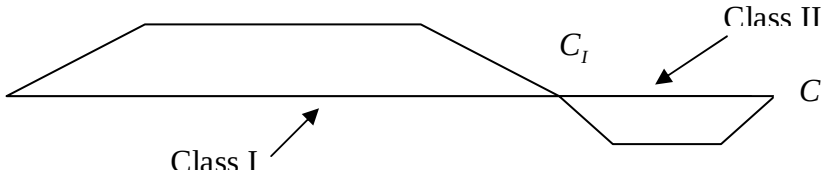


Fig.5.3 - Complete Partitioning

Nel secondo schema, CA, le connessioni real time possono utilizzare l'intera capacità con maggior priorità sulle connessioni non real time, ma in ogni momento la capacità non utilizzata dalle connessioni real time è ugualmente condivisa tra tutte le connessioni non real time, sia quelle a priorità più alta che quelle a priorità più bassa. Questo algoritmo adotta la classica politica work conserving e le risorse lasciate inutilizzate dalla classe GB sono totalmente a disposizione della classe BE.

Dato il meccanismo di priorità, potrebbe comunque accadere che la maggior parte delle volte il traffico senza requisiti di priorità non venga mai soddisfatto; per questo viene in soccorso un terzo algoritmo di scheduling, il Restricted Access.

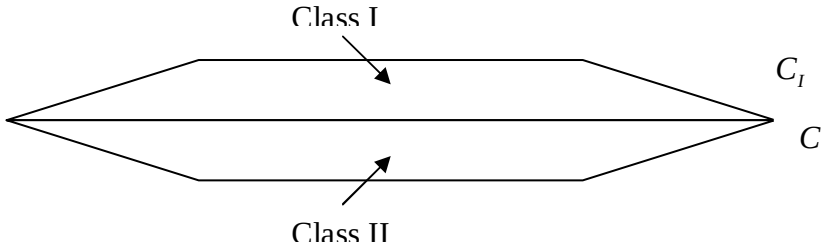


Fig.5.3 – Complete Access

Nel meccanismo di accesso ristretto, le connessioni real time possono usare la banda fino ad una percentuale massima, fissata ad esempio all'80%. La banda rimanente è allocabile esclusivamente alle sorgenti BE, che possono accedere comunque alle risorse riservate alle sorgenti GB che non state però utilizzate da queste. In figura 5.5 si può vedere quanto descritto: il tratto di banda fino a C_I è dedicato alle connessioni real time con priorità maggiore, e il tratto di banda da C_I a C è dedicata alle connessioni non real time (sia quelle a priorità più alta che quelle a priorità più bassa). In modo simile al meccanismo CA, la capacità non utilizzata dalle connessioni real time, come già accennato, è disponibile per le connessioni non real time. Quello che si può fare pertanto, è prevedere un passaggio dal meccanismo CA ad RA nel momento in cui si verifichi un elevato carico GB nel sistema, con eccessiva penalizzazione del traffico BE. Questo consente alle connessioni di tipo non real time di ricevere la parte della risorsa di cui non potrebbe beneficiare con l'utilizzo esclusivo dell'algoritmo precedente.

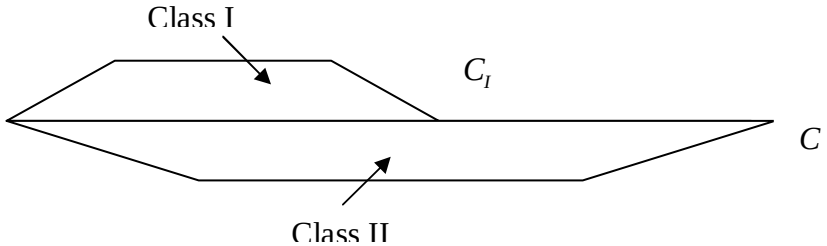


Fig.5.5 - Restricted Access

Gli ultimi due schemi di allocazione (RA e CP) sfruttano una politica di tipo work conserving e consentono una allocazione delle risorse lasciate inutilizzate come è mostrato in figura 5.6, l'eccesso è egualmente condiviso:

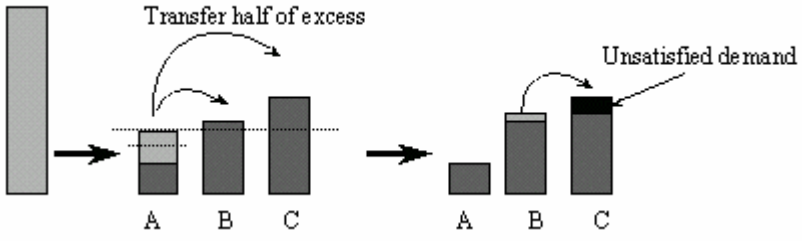


Fig. 5.6 - Trattamento del surplus di risorsa da assegnare

Si potrebbe pensare di definire, a questo punto, un algoritmo ibrido in grado di variare la strategia di "riservazione" della risorsa in funzione dello stato di congestione della rete. Tale algoritmo potrebbe partire dall'utilizzo della strategia CA, che sostanzialmente mette delle sorgenti GB tutta la banda del sistema. Le sorgenti BE hanno accesso, in ogni trama, alle MAC-PDU eventualmente lasciate libere dalle sorgenti a qualità garantita. E' stato già evidenziato che un approccio di questo tipo può risultare troppo penalizzante per le sorgenti BE nel momento in cui l'intera banda risulti effettivamente occupata dalle unità informative prodotte dalle sorgenti a priorità maggiore, soprattutto considerando il fatto che queste potrebbero anche essere state emesse in violazione dei parametri individuati in fase di contratto. Nel momento in cui si dovesse valutare l'effettiva impossibilità, da parte delle sorgenti BE, a trasmettere le proprie MAC-PDU, sembra allora opportuno immaginare di passare ad una strategia di tipo RA, che assicuri almeno un minimo della capacità per le sorgenti a priorità più bassa. Deve essere chiaro, in ogni caso, che questa possibilità deve essere comunque vincolata al rispetto dei parametri di qualità del servizio delle sorgenti GB. Il passaggio dallo schema CA a quello RA può essere inteso come una restrizione degli eccessi di MAC-PDU trasmesse per le sorgenti a qualità garantita. Proprio questa considerazione ci induce, ad immaginare un passaggio intermedio attraverso lo schema CP. Questo schema, evidentemente meno efficiente dello schema RA in termini di allocazione

della capacità trasmissiva, presenta comunque una semplicità di implementazione che può portare a preferirlo nel momento in cui siano alte le garanzie di utilizzare quasi completamente le due partizioni da questo individuate (che equivale ad avere una perdita ridotta in efficienza di occupazione della banda rispetto allo schema RA). L'algoritmo ibrido, pertanto, parte dallo stato standard associato allo schema CA. Nel momento in cui si verifichi l'impossibilità di accedere alle risorse del sistema, da parte delle sorgenti BE, si opera una transizione allo schema CP, fissando ad esempio le due partizioni pari all'80% della banda totale per le sorgenti GB e al restante 20% per le sorgenti BE (sempre se risultano rispettati i vincoli QoS per le sorgenti a qualità garantita). Poiché ci si trova in una situazione nella quale le sorgenti GB tendono ad occupare tutta la capacità del sistema, sarà ragionevole assumere che l'80% riservato loro sia completamente occupato. Allo stesso modo, poiché la transizione è stata motivata dall'impossibilità per le sorgenti BE di accedere al sistema, è possibile immaginare una buona occupazione del 20% della banda messa adesso a disposizione. Si permane in questo stato, fino a quando non si incorre in una delle due seguenti situazioni:

- si verifica una scarsa occupazione della partizione assegnata alle GB
- si verifica una scarsa occupazione della partizione assegnata alle BE.

Nel primo caso, è opportuno abbandonare la politica non work conserving dello schema CP e ammettere che le sorgenti BE occupino lo spazio lasciato libero dalle sorgenti a qualità garantita. In questo caso, infatti, la complicazione del protocollo è ampiamente giustificata dal potenziale incremento nell'efficienza di occupazione della capacità a disposizione. Nel caso contrario, quando è la partizione assegnata alle sorgenti BE a mostrarsi scarsamente occupata, è invece opportuno tornare allo schema standard CA, permettendo in sostanza l'ingresso di nuove sorgenti a qualità garantita o, perlomeno la trasmissione delle relative MAC-PDU in eccesso. Un ritorno allo schema standard di "riservazione" e allocazione della risorsa diventa opportuno per le stesse motivazioni, anche operando con la politica RA, sempre quando risulti poco utilizzato quel 20% della capacità trasmissiva interdetto per le sorgenti GB a favore delle BE.

La disciplina di scheduling supportata è una disciplina di tipo FIFO, nella quale il primo pacchetto che entra nella coda è il primo ad uscire. Dal momento che ci sono tante code per quanti sono i RT, il modo in cui si procede allo scheduling di tutti i pacchetti di tutte le code è quello già considerato per la prima modellizzazione, quello round robin, che fornisce una uguale suddivisione delle risorse tra tutte le code. Dopo aver servito una coda lo scheduler procede a quella successiva non vuota, in ordine ciclico.

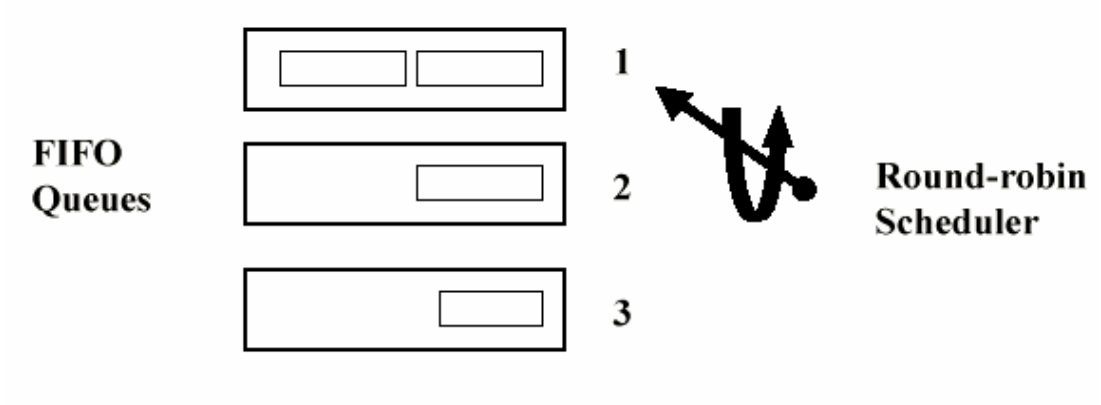


Fig.5.7 - Gestione delle code

Gli algoritmi di scheduling esaminati nel primo paragrafo sono serviti anche come riferimento per la gestione delle etichette dei tempi dei pacchetti e per la scelta della modalità di servizio di ogni pacchetto all'interno delle code. La suddivisione della risorsa, come visto, si è basata esclusivamente sull'algoritmo proposto da Acampora. Ancora qualche parola sulla gestione della doppia priorità per la classe di servizio BE. Nell'algorimo di scheduling statico, era stata ipotizzata la presenza di una unica coda per la classe di servizio BE. Allo scopo di considerare una condizione di traffico di servizio differente e per vedere come il sistema si comportava nel caso della presenza nel sistema di due classi BE, si è pensato di aggiungere nel sistema una coda rappresentativa di richieste BE a più bassa priorità rispetto a quelle già considerate fino a questo momento. Il meccanismo di trattamento di queste risorse è molto semplice. Le risorse vengono servite secondo il seguente ordine di priorità:

1. quota parte riservata delle risorse a qualità garantita (GB)

2. risorse presenti nelle code GB fino al valore previsto dal particolare algoritmo di scheduling
3. risorse BE ad alta priorità
4. risorse BE a bassa priorità

La figura 5.1 relativa all'assegnazione delle risorse deve essere quindi modificata tenendo presente che oltre i tre step già definiti è presente un quarto step in cui la quantità di risorsa eventualmente ancora libera è lasciata a disposizioni di eventuali MAC-PDU di classe BE a bassa priorità che attendono in coda.

CAPITOLO 6

DESCRIZIONE DEL SIMULATORE

6.1 Descrizione della Versione 1.0

6.1.1. Introduzione

All'interno del Progetto di uno Strato MAC che risponda alle caratteristiche richieste dalla WLAN in considerazione in questo studio, ha un ruolo importante l'implementazione di un simulatore per poter valutare appieno la convenienza delle scelte progettuali, prima ancora che si passi alla fase della realizzazione vera e propria del progetto.

Il Team MURST dell'Unità di Roma ha completato nel mese di Luglio 2001 la realizzazione della prima versione di questo simulatore (versione 1.0), di cui è importante evidenziare alcune caratteristiche generali prima di passare ad una descrizione più accurata del suo funzionamento.

La versione 1.0 del simulatore implementa solamente la tratta *downlink* del sistema in questione (collegamento dal Radio Node ai Radio Terminal). La scelta di implementare dapprima questa tratta è stata presa sia in considerazione del fatto che è la tratta su cui normalmente c'è un maggiore traffico, sia perché la tecnica di accesso al mezzo da utilizzare sulla tratta di uplink è al momento in fase di studio da parte della stessa Unità di Roma.

Inoltre, è rilevante evidenziare che, in prima approssimazione, si è assunto che il sottostante canale di trasmissione sia ideale (cioè privo di errori) e che le sorgenti di traffico presenti all'inizio della simulazione restino attive per tutto il periodo preso in considerazione.

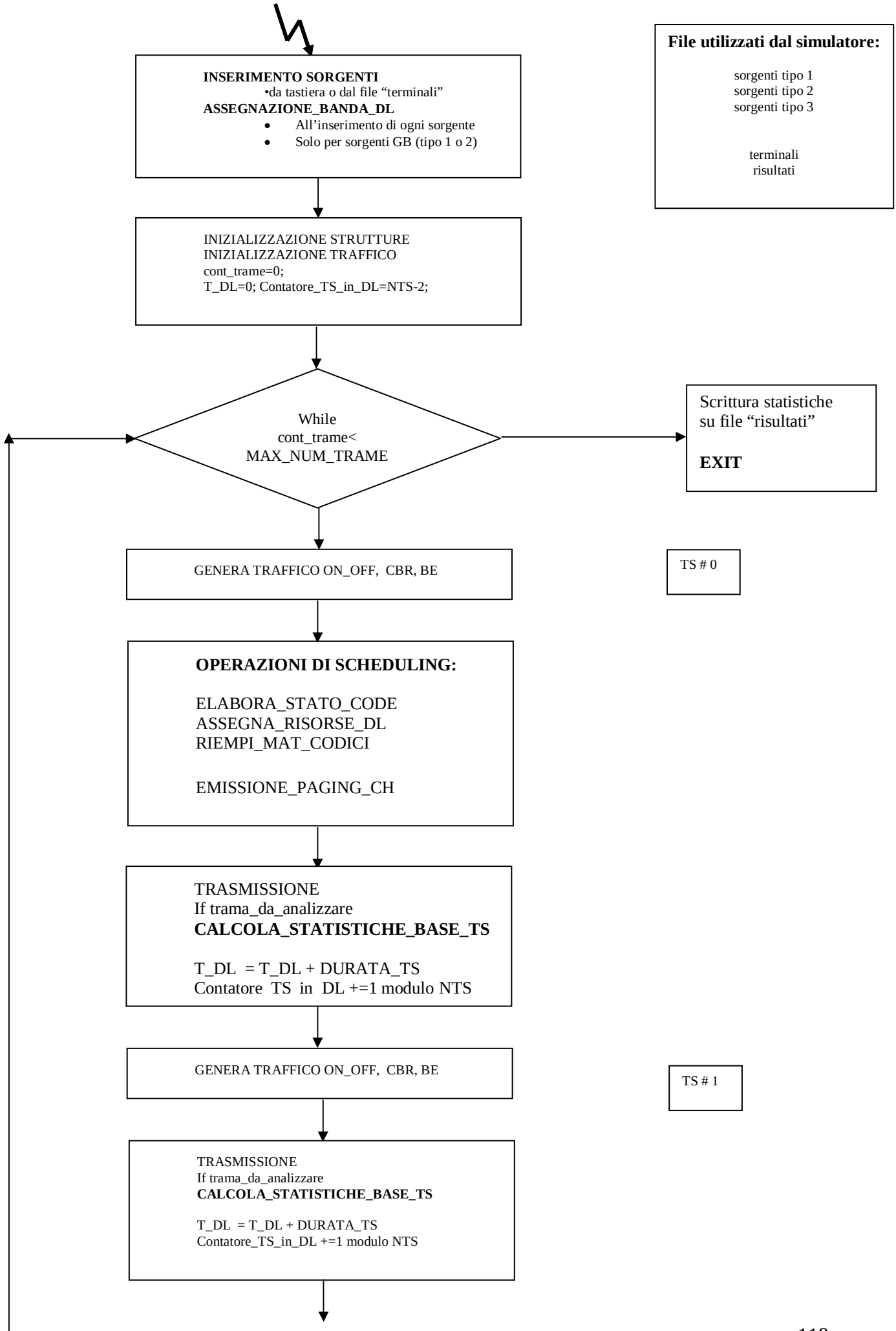
L'Unità di Roma, terminata la versione 1.0 del simulatore, è correntemente impegnata nello sviluppo di una successiva versione del simulatore che rimuova i limiti e le ipotesi appena descritte.

Questa studio, in particolare, è orientato allo sviluppo di un Protocollo di Scheduling adattivo alle condizioni di traffico che si suppone non più statico ma mutante nel tempo grazie all'introduzione dei processi di nascita e morte, che possa essere convenientemente integrato nel simulatore realizzato.

Infine, è di notevole importanza evidenziare che sia la versione 1.0 del simulatore, sia le versioni successive, che integrano e arricchiscono la versione base in determinati aspetti, sono realizzate utilizzando come linguaggio di programmazione il C (C++), con l'ottica di poter utilizzare in un prossimo futuro il simulatore realizzato all'interno del simulatore di rete realizzato dall'Università di Berkley e denominato Network Simulator (*ns*). D'altra parte, poiché l'interesse primario inizialmente è stato sviluppare un simulatore di Strato MAC per poter testare fin da subito le scelte di progetto effettuate, si è proceduto ad implementare un simulatore che, almeno inizialmente, fosse in grado di funzionare come entità a sé stante, senza la necessità di essere inserito nell'architettura prevista da *ns*. Per questo motivo, sono presenti procedure che simulano la presenza degli Strati che logicamente confinano con lo Strato MAC. Fin dalla versione 1.0, infatti, la generazione del traffico IP in entrata al Protocollo di accesso al mezzo è simulata da apposite procedure.

Per quanto riguarda l'interfacciamento con lo Strato di Rete, esso è stato introdotto proprio nella versione aggiornata del simulatore proposta da questo studio, in quanto si è trattato di introdurre delle procedure ausiliarie che non solo rendessero possibile variare le condizioni di traffico della rete durante l'evolversi del tempo, simulando dei veri e propri processi di nascita e morte, ma anche realizzando uno scheduling capace di cambiare il suo funzionamento in relazione alla quantità di risorsa assegnata alla classe di servizio a qualità garantita e best effort. Il canale è invece stato supposto sempre ideale ovvero è un canale che non commette errori.

Nella Sezione 6.1.2 è presentato ad alto livello il funzionamento del simulatore nella sua versione 1.0, mentre nelle Sezioni 6.2 e 6.3 viene descritto come è stato modificato per l'introduzione dei processi di nascita e morte e di uno scheduling adattivo alle mutevoli condizioni di traffico.


File utilizzati dal simulatore:
sorgenti tipo 1
sorgenti tipo 2
sorgenti tipo 3
terminali
risultati
INSERIMENTO SORGENTI
•da tastiera o dal file "terminali"
ASSEGNAZIONE_BANDA_DL
All'inserimento di ogni sorgente
Solo per sorgenti GB (tipo 1 o 2)
INIZIALIZZAZIONE STRUTTURE
INIZIALIZZAZIONE TRAFFICO
cont_trame=0;
T_DL=0; Contatore_TS_in_DL=NTS-2;
While
cont_trame<
MAX_NUM_TRAME
Scrittura statistiche
su file "risultati"
EXIT
GENERA TRAFFICO ON_OFF, CBR, BE
TS # 0
OPERAZIONI DI SCHEDULING:
ELABORA_STATO_CODE
ASSEGNA_RISORSE_DL
RIEMPI_MAT_CODICI
EMISSIONE_PAGING_CH
TRASMISSIONE
If trama_da_analizzare
CALCOLA_STATISTICHE_BASE_TS
T_DL = T_DL + DURATA_TS
Contatore TS in DL +=1 modulo NTS
GENERA TRAFFICO ON_OFF, CBR, BE
TS # 1
TRASMISSIONE
If trama_da_analizzare
CALCOLA_STATISTICHE_BASE_TS
T_DL = T_DL + DURATA_TS
Contatore_TS_in_DL +=1 modulo NTS

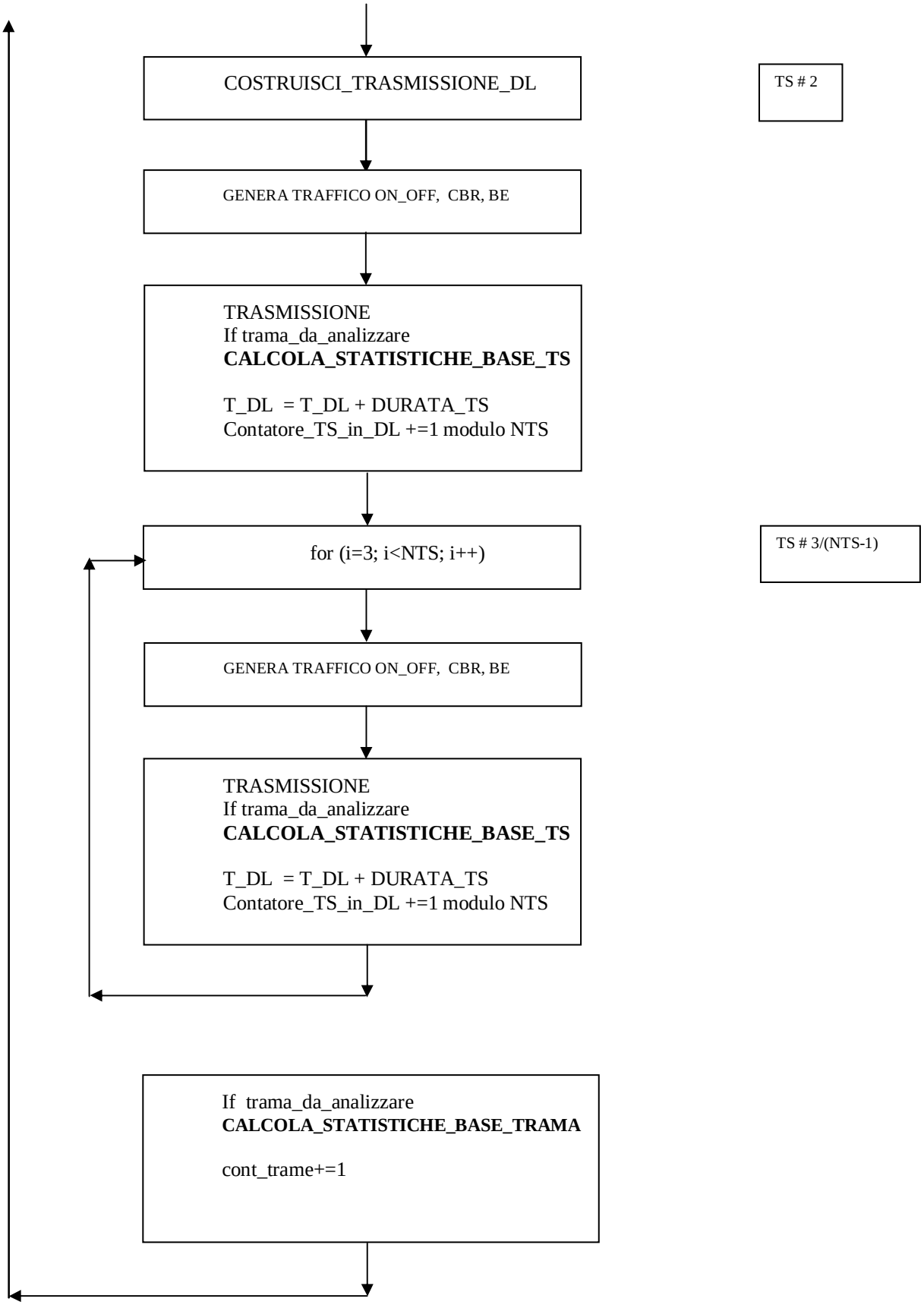


Fig. 6.1 - Diagramma di flusso del Simulatore (Versione 1.0)

6.1.2. Funzionamento del simulatore

La versione 1.0 del simulatore prevede che le sorgenti presenti e attive nella WLAN all'inizio della simulazione restino attive per tutto il periodo preso in considerazione. Per questo motivo, la simulazione inizia con l'inserimento delle informazioni riguardo il numero delle sorgenti attive presenti al Radio Node e del terminale (RT) di destinazione dei pacchetti dati che le sorgenti emetteranno.
L'inserimento delle sorgenti avviene quindi all'inizio della simulazione e può essere effettuato da tastiera o tramite il *file* "terminali.txt". Quest'ultima modalità è stata inserita per permettere all'utente di lanciare delle simulazioni su computer remoti.
Effettuata questa operazione, il simulatore provvede autonomamente a prelevare, per ciascuna sorgente inserita, le caratteristiche di emissione dai file descrittori delle sorgenti ("sorgentitipo1.txt", "sorgentitipo2.txt", "sorgentitipo3.txt"). I file descrittori delle sorgenti sono tre poiché la versione 1.0 definisce tre classi di traffico principali (ON/OFF, CBR, BE) e quindi a seconda del tipo di sorgente che è stata inserita dall'utente (ON/OFF, CBR, BE), il simulatore preleva i parametri descrittori dal relativo file. Inoltre, per ciascuna classe di traffico sono definite tre sottoclassi (vedi figura 6.1) e quindi, sempre in riferimento alla descrizione dell'utente, vengono scelti i parametri della particolare sottoclasse.

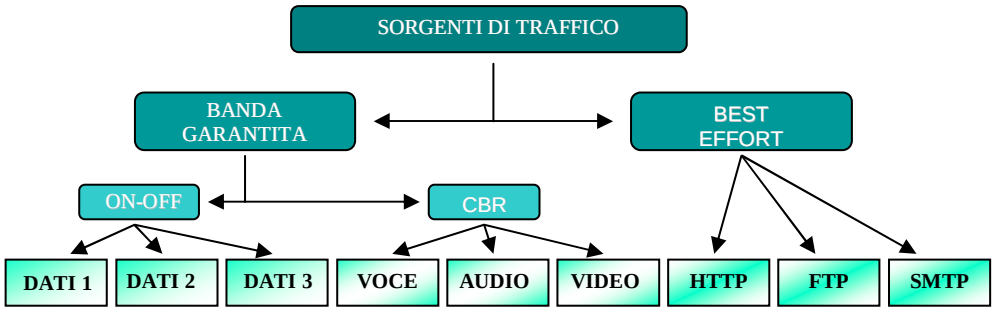


Fig. 6.2 - Sorgenti di traffico implementate nel simulatore (versione 1.0)

Per le sorgenti ON/OFF e CBR i parametri descrittori sono i parametri Dual Leaky Bucket (rate di picco, rate medio, Token Buffer Dimension) oltre che la dimensione del pacchetto IP tipicamente usata da quel particolare tipo di traffico e il massimo

ritardo accettabile per quella connessione.Si noti che, naturalmente, per le sorgenti CBR il rate di picco coincide con il rate medio.

Per le sorgenti BE i parametri descrittori sono la frequenza media di arrivo dei pacchetti (i tempi di inter-arrivo dei pacchetti del traffico BE sono generati secondo il modello di Poisson) e la dimensione tipica del pacchetto IP per quel particolare traffico.

Classe di Traffico	Parametri
ON/OFF	• Rate di picco • Rate medio • Token Buffer Dimension • Lunghezza pacchetto IP • Massimo ritardo tollerabile
CBR	• Rate • Lunghezza pacchetto IP • Massimo ritardo tollerabile
BE	• Frequenza media di emissione dei pacchetti IP • Lunghezza pacchetto IP

Tabella 1: Parametri delle sorgenti presenti nei tre file sorgenti tipo 1, 2, 3

I parametri descrittori delle connessioni attive sono raccolte e ordinate per terminale di destinazione in una struttura dinamica che assicura la possibilità di gestire ogni sorgente come un elemento indipendente (vedi figura 6.3).

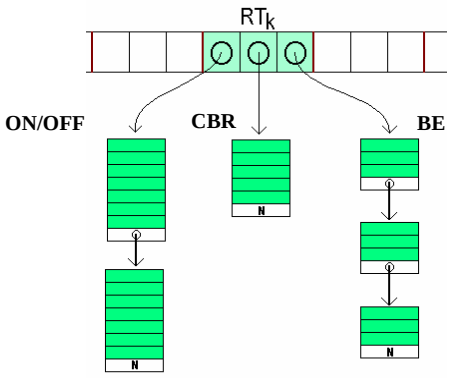


Fig. 6.3: Sorgenti ON/OFF, CBR e BE attive al RT_k

Per quanto riguarda le sorgenti GB (cioè sia le ON/OFF che le CBR), la versione 1.0 assume che l'emissione del traffico sia coerente con la maschera individuata dai parametri DLB. In particolare si sceglie, tra tutte le sorgenti ON/OFF (o CBR) che rispettano la maschera DLB individuata, la sorgente ON/OFF (o CBR) che in media emette lo stesso numero di bit/s che sarebbero emessi dalla sorgente DLB vera e propria (vedi figura 6.4).

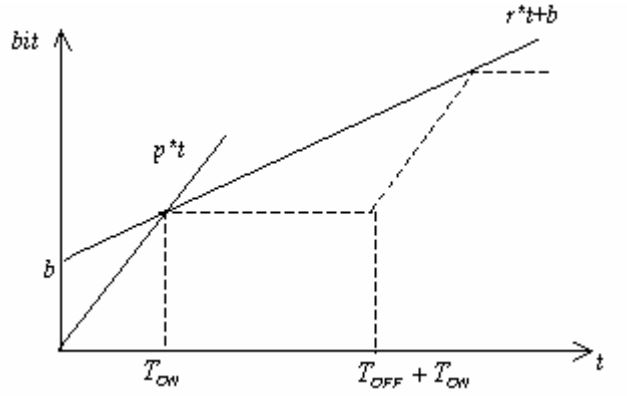


Fig.6.4 - Scelta di T_{ON} e T_{OFF} fra tutti quelli coerenti con la maschera DLB

In questo modo si determinano i periodi di emissione (T_{ON}) e di silenzio (T_{OFF}) della sorgente ON/OFF scelta fra quelle compatibili con la maschera DLB. In altre parole si sceglie quella sorgente ON-OFF che emette per un certo periodo (T_{ON}) al rate di picco, finchè non raggiunge la maschera DLB e successivamente rispetta un periodo di silenzio (T_{OFF}) sufficientemente lungo da permetterle di ricominciare subito dopo ad emettere al rate di picco (successivo T_{ON}). Per quanto riguarda le sorgenti CBR, essendo il rate di picco coincidente per definizione con il rate medio, tra tutte le sorgenti CBR che hanno una maschera di emissione compatibile con quella descritta dalla maschera DLB (cioè la cui curva di emissione è costantemente "sotto" la maschera DLB) si sceglie la sorgente CBR che emette al rate di picco previsto dalla maschera DLB.

Per le sorgenti GB, quindi, l'emissione è perfettamente nota e descritta dalla maschera DLB, quindi è possibile calcolare le risorse che devono essere assegnate

al particolare flusso affinché siano rispettati i predefiniti requisiti di qualità del servizio. Per questo motivo, il rate minimo R_{MIN} ricavato nel paragrafo 4.3 è proprio il rate che deve essere garantito alla particolare sorgente ON/OFF (o CBR). Quindi, la caratteristica della versione 1.0 è una conoscenza perfetta dell'emissione nel tempo delle sorgenti GB; questo è il motivo per cui all'inizio della simulazione è possibile associare a ciascuna connessione GB la banda minima che deve essergli garantita (in termini di MAC PDU / Trama).

A questo punto, oltre alla inizializzazione delle strutture usate dal simulatore, viene inizializzato il traffico per preparare le successive chiamate alle procedure di generazione del traffico.

Il traffico GB è generato su base Time Slot ed assume che i pacchetti IP vengano passati agli strati sottostanti solo quando sono completi. Quindi un pacchetto IP proveniente da una sorgente di classe ON/OFF o CBR si assume generato nel Time Slot in cui cade l'istante di fine generazione del pacchetto. Questo permette di operare la segmentazione di interi pacchetti IP in MAC PDU e di operare funzionalità di Strato di Adattamento su tutto il pacchetto, volte a garantire i parametri di QoS richiesti dalle diverse sorgenti da cui provengono i pacchetti IP.

Per le sorgenti ON/OFF, la fase di inizializzazione controlla se la sorgente si trova attualmente in un periodo di attività o di silenzio e calcola l'istante in cui emetterà il primo pacchetto IP. La fase di generazione del traffico ON/OFF prosegue aggiornando l'istante di arrivo del successivo pacchetto IP ogni volta che sia stato generato un pacchetto. Naturalmente, il calcolo dell'istante di generazione del pacchetto successivo è rispettoso dei periodi di inattività (T_{OFF}) della sorgente.

Per le sorgenti CBR, la fase di inizializzazione calcola l'istante in cui verrà generato il prossimo pacchetto IP e la fase di generazione del traffico CBR prosegue aggiornando l'istante di arrivo del prossimo pacchetto IP.

La generazione del traffico BE è anche essa effettuata su base Time Slot e si trascura il tempo di pacchettizzazione. La fase di inizializzazione genera l'istante di

arrivo del primo pacchetto IP basandosi su una estrazione di una variabile aleatoria (V.A.) a distribuzione esponenziale negativa.

La procedura di generazione del traffico BE genera tutti gli istanti di nascita dei pacchetti IP che cadono nel Time Slot corrente calcolando i tempi di interarrivo come realizzazione della stessa V.A. esponenziale negativa.

Inoltre prepara l'istante di generazione del prossimo pacchetto IP che verrà generato nel successivo o nei successivi TS.

Dopo l'inizializzazione del traffico, inizia la simulazione vera e propria. Si tratta di un ciclo che si ripete per quante sono le trame prese in considerazione nella simulazione. Come è stato detto in precedenza, ogni trama si compone di N_{TS} Time Slot (le simulazioni sono state effettuate con N_{TS}=6). Ciascun Time Slot (TS) è caratterizzato da alcune operazioni comuni a tutti i TS e, eventualmente, da altre operazione tipiche del particolare TS. Ad esempio, in tutti i TS viene generato il traffico e vengono trasmesse delle MAC PDU.

Il primo TS di ogni trama, il RN trasmette ai RT , tramite il Canale di Paging, le informazioni su dove ascoltare all'interno della Matrice dei Codici le unità dati a loro dirette. Naturalmente, prima di emettere il Paging Channel, il RN effettua l'algoritmo di scheduling descritto nel paragrafo 5.2. A questo scopo di attivare lo *scheduler*, il RN ha bisogno di conoscere lo stato delle code (la loro lunghezza). Dopo aver assegnato una percentuale della prossima trama di DL a ciascuna connessione attiva, il RN posiziona le MAC PDU di tutte le connessioni nella Matrice dei Codici.

Nella versione 1.0, essendo un problema tipicamente di Strato Fisico la scelta istante per istante dei codici migliori dove effettuare la trasmissione, effettua un riempimento *casuale* della Matrice dei Codici, come mostrato in figura 6.5.

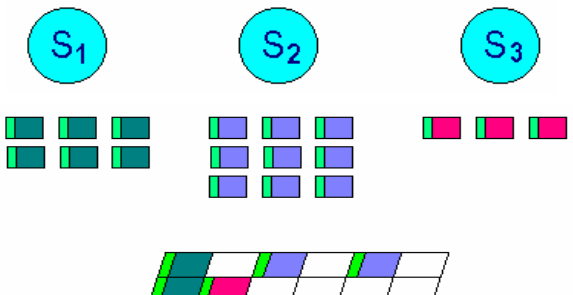


Fig. 6.5 - Riempimento della Matrice dei Codici

Come si è accennato, in ogni TS vengono trasmesse delle unità dati. Nella versione 1.0, poiché il canale è ipotizzato ideale, una volta individuate le MAC PDU che vengono trasmesse nel TS in considerazione, la trasmissione si riduce semplicemente a togliere quelle MAC PDU dal relativo buffer.
In realtà, in occasione della trasmissione, se la trama corrente appartiene all'insieme di trame da considerare nelle statistiche, vengono calcolate e aggiornate le statistiche di rilievo che riguardano la trasmissione delle MAC PDU:

- Ritardo min e max per le sorgenti ON-OFF per ogni RT
- Ritardo min e max per le sorgenti CBR per ogni RT
- Ritardo min e max per le sorgenti BE per ogni RT
- Stime dei valori medi e delle varianze dei ritardi per ogni sorgente attiva
- Istogramma del ritardo per le sorgenti BE
- Istogramma del numero di MAC PDU presenti nel buffer GB in un TS
- Istogramma del numero di MAC PDU presenti nel buffer BE in un TS
- Istogramma dei codici che vengono assegnati per trasmissioni GB
- Istogramma dei codici che vengono assegnati per trasmissioni BE

Il terzo TS è l'altro TS dove il simulatore compie altre operazioni, oltre alla generazione del traffico, alla trasmissione e all'incremento dei contatori del tempo

(T_DL è il assoluto dei tempi mentre Contatore_TS_in_DL individua il TS corrente all'interno della matrice dei codici).

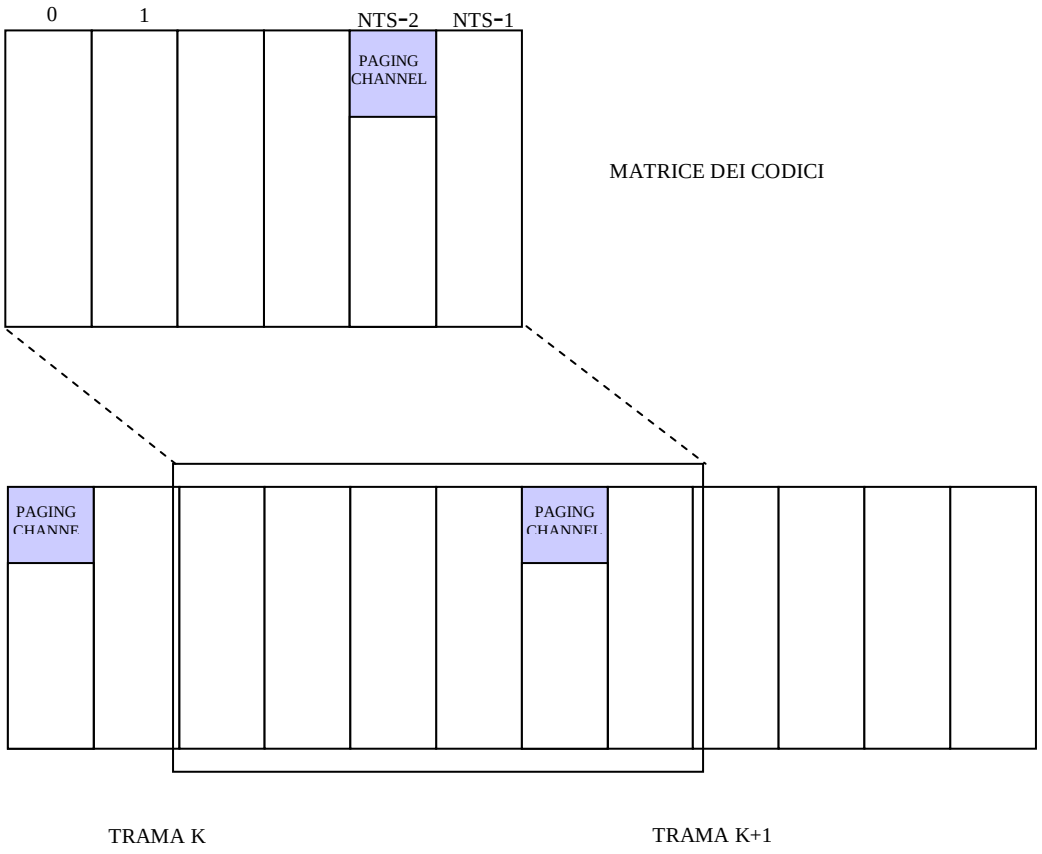


Fig.6.6 - Relazione fra Matrice dei Codici e Struttura a Trama

Come si può vedere in figura 6.6, la definizione della matrice dei Codici non corrisponde alla trama a causa di una traslazione temporale. La trama corrisponde invece esattamente all'evolversi dei Time Slot come indicato dal grafo di flusso di figura 6.1.

Il motivo di ciò è che si deve dare tempo ai RT di ricevere correttamente il Paging Channel e di interpretare le informazioni che trasporta. Questo significa che il Paging Channel porta informazione su dove i terminali devono ascoltare le unità dati loro indirizzate all'interno della zona evidenziata in figura 6.6 (zona che corrisponde alla definizione di matrice dei codici utilizzata nel simulatore).

Questa traslazione fra la definizione di Matrice dei Codici e trama spiega anche perché all'inizio del grafo di flusso di figura 6.1 Contatore_TS_in_DL è stato inizializzato a -2.

Le operazioni eseguite per la prima trama vengono eseguite tante volte quante sono le trame considerate nella simulazione. Inoltre, alla fine di ciascuna trama, se essa appartiene all'insieme di trame da considerare nelle statistiche, vengono calcolate e aggiornate ulteriori statistiche:

- Istogramma del numero di MAC PDU ON/OFF trasmesse per trama
- Istogramma del numero di MAC PDU CBR trasmesse per trama
- Istogramma del numero di MAC PDU GB trasmesse per trama
- Istogramma del numero di MAC PDU BE trasmesse per trama

6.2. Integrazione dei Processi di nascita e morte alla Versione 1.0

Come descritto nei paragrafi precedenti, il modello scelto per l'inserimento delle sorgenti nel sistema, durante la prima fase del lavoro, ha previsto l'introduzione da tastiera del numero e del tipo di sorgenti desiderate per la particolare simulazione. L'inserimento delle sorgenti nella 1° versione avveniva una sola volta prima dell'inizio della simulazione e le sorgenti introdotte rimanevano sempre attive, senza possibilità di morte.
Un approccio di questo tipo è apparso non rispettare il reale comportamento di una qualsiasi sorgente, che si suppone nasca, sviluppi traffico e muoia. E' per questa ragione, che è stato affrontato il problema della nascita e della morte delle sorgenti e della sua realizzazione all'interno del simulatore di un sistema di tipo più realistico. Partendo dalla prima versione del simulatore, in cui il primo passo era dato dall'inserimento delle sorgenti, nella nuova versione la nascita della sorgente viene valutata alla fine di ogni trama. E' stato necessario apportare una variante al simulatore che consentisse di caratterizzare la nascita delle sorgenti non una sola

volta bensì in maniera cadenzata nel tempo. Una approssimazione che è sembrata lecita, è stata quella di andare a valutare l'eventuale nascita di sorgenti all'interno di una finestra temporale. Dal momento che l'intervallo di un TS, pari a 0,0015 secondi, è sembrato molto piccolo come intervallo di possibili nascite, è stata considerata la nascita di una o più sorgenti, all'interno di un intervallo maggiore, la trama. Chiaramente in base alle ipotesi fatte, la valutazione delle nascite avviene a fine trama, la prima trama risulterà "scarica".
Con l'evolversi del tempo si assiste ad una trasformazione delle sorgenti presenti nel sistema: crescente se si realizza la nascita di una nuova sorgente, decrescente se si è verificata una morte.
Tre sono le scelte fondamentali per il supporto dei processi di nascita e morte:

1. il numero di sorgenti che nascono in un'ora per ogni sottotipo
2. il tempo medio di attività della sorgente in questione
3. il numero di bytes totali da trasferire per assistere alla morte della sorgente una volta che questa abbia trasmesso il dovuto: ciò è realizzato a partire dalla conoscenza del tempo di attività della sorgente, estraendo una variabile aleatoria di distribuzione geometrica rappresentante la quota di traffico sviluppata per quella sorgente.

Una volta realizzata la nascita delle sorgenti, queste genereranno un traffico proporzionale al valore estratto (come descritto nel punto 3). Ci si aspetta che il sistema sia delineato da momenti in cui ci sia molto traffico da smaltire e momenti in cui, invece, il sistema risulti vuoto.
Una volta che la sorgente abbia trasmesso tutto quello che doveva, viene realizzata la morte della stessa. La sorgente relativa deve essere staccata dalla struttura che tiene conto di tutte le sorgenti presenti nel sistema, mantenendo memorizzati però i valori di statistiche necessari per poter valutare i risultati. L'eliminazione della sorgente da un punto di vista logico avviene con la "cessazione" della generazione del traffico della stessa.

I valori di probabilità di nascita e la dimensione media del traffico generato da una sorgente, dal momento che tutto è regolato in maniera casuale nel tempo, sono stati inseriti come ulteriori parametri nei file “sorgenti”, nominati all’inizio del capitolo.
Il parametro su cui si basa la nascita della sorgente è rappresentato dal numero medio di nascite, per quel tipo di sorgente, in una ora. Conosciuta la durata della simulazione è possibile tramite il numero medio appena detto risalire al numero di sorgenti che devono nascere durante la simulazione in analisi. Una giusta “calibratura” del simulatore dovrebbe garantire un numero di nascite prossimo a quello calcolato, seppure non in maniera puntuale, in media. Per ogni sottoclasse di ogni sorgente di traffico viene valutato con cadenza di trama se è nata una sorgente di quel tipo o meno. Poiché la nascita di una sorgente è l’ultima operazione fatta all’interno di una trama, è presto ammesso che il sistema nasca scarico e la prima possibilità di allocazione delle risorse si ottenga nel primo TS della trama successiva, non appena venga realizzata la generazione del traffico e venga realizzato lo scheduling. E’ chiaro dal meccanismo descritto che il massimo numero di sorgenti che nascono è pari al numero delle sottoclassi totali considerate. Realizzata la simulazione delle nascite, per riuscire a caratterizzare le morti è stato necessario andare ad agire sulle procedure della generazione del traffico.
La prima versione del simulatore supponeva che una volta che la sorgente avesse fatto l’ingresso nel sistema non ne sarebbe più uscita, la conseguenza ovvia di questo è stata una generazione continua di traffico per la sorgente inserita con una trasmissione dei pacchetti IP via via generati permanente. Rimovendo tale ipotesi, è stato necessario bloccare la generazione del traffico per la generica sorgente che abbia trasmesso la quantità di bytes calcolata come descritto precedentemente.
E’ ovvio che un funzionamento della procedura di questo genere, come è facilmente constatabile dai grafici, abbia permesso di rappresentare molto bene, come la vita di una sorgente può occupare solo una trama come estendersi per un numero più alto di trame. Al termine della trasmissione la sorgente cessa la sua attività e verrà eliminata dalla struttura che memorizza tutte le sorgenti presenti.

Dal punto di vista delle statistiche, tre sono state le valutazioni fondamentali per la descrizione dei processi di nascita e morte: le sorgenti di ogni tipo e di ogni sottoclasse nate durante la simulazione, le sorgenti morte e le sorgenti presenti nel sistema in una particolare finestra temporale.

6.3. Integrazione di un Protocollo di SCHEDULING alla Versione 1.0

L'algoritmo di scheduling previsto nella prima versione del simulatore si è constatato si comportasse bene in condizioni di equità di traffico a qualità garantita e best effort, ma non troppo equamente qualora una grande quantità di sorgenti GB venivano inserite rispetto ad un numero non eccessivo di sorgenti BE. La doppia garanzia fornita alle GB era un po' troppo stringente nei confronti del traffico BE.
L'introduzione di una ulteriore coda per la classe di tipo best effort, nonché il desiderio di realizzare uno scheduling che garantisse comunque una parte della risorsa alle sorgenti con minore priorità, ci ha indotto a realizzare uno scheduling adattativo, che in condizioni di carico pesante GB, è comunque in grado di allocare una buona parte della risorsa alla classe BE, qualora questa non riceva allocazione o soffra del troppo traffico GB generato.
In realtà l'algoritmo di scheduling si divide in tre fasi che realizzano ognuna una allocazione differente della risorsa e che si "rimbalzano" tra di loro per consentire un'assegnazione della risorsa anche alla classe best effort, non troppo penalizzante, pur non avendo questa alcuna garanzia di banda, almeno in linea di principio.
Come mostrato nel paragrafo 5.3, sono stati realizzati "tre scheduling" : uno con accesso completo sia per la GB che per la BE, uno con partizionamento completo della risorsa tra GB e BE, l'ultimo con accesso ristretto alla risorsa da parte della GB e completo alla BE. L'evoluzione del numero delle richieste e l'assegnazione alle GB, comporta passaggi successivi a seconda della situazione del sistema da un algoritmo di scheduling all'altro. Questi passaggi sono facilmente visibili nei grafici successivi rappresentativi dello scheduling utilizzato trama per trama. L'obiettivo che si intende conseguire con questo nuovo algoritmo è il raggiungimento di equità

di trattamento delle sorgenti. Questo al fine di allocare una quota parte di banda maggiore, rispetto all'algoritmo di scheduling statico descritto nel paragrafo 5.2, alla classe BE, che non presenta nessun requisito di garanzie di banda. L'overflow considerato per ogni classe dovrebbe essere minore rispetto alla 1° versione del simulatore, così come il ritardo medio delle sorgenti BE.

Per quanto concerne la doppia coda per la classe best effort, tale algoritmo dovrebbe consentire una buona allocazione delle risorse anche a quelle a più bassa priorità, non presenti nella prima versione del simulatore. Come avevamo detto nella descrizione di quest'ultima, la struttura che raccoglieva le sorgenti era suddivisa in tre parti, sorgenti ON OFF, sorgenti CBR e sorgenti BE. La struttura evoluta è stata arricchita di un nuovo gruppo di sorgenti le sorgenti BE a bassa priorità che hanno stesse caratteristiche delle prime ma priorità più bassa. Questa doppia coda per le sorgenti BE, ci servirà per capire quanto l'algoritmo implementato per favorire un comportamento equo si comporti in maniera corretta anche per le sorgenti BE.

Come al solito la modalità di assegnazione delle risorse è in pieno rispetto di quanto descritto nel capitolo del MAC. Viene assegnato il dovuto alla classe a qualità garantita e poi cominciano ad essere soddisfatte le code di ogni RT per ogni tipo di classe, in misura proporzionale all'algoritmo di scheduling usato.

Si parte da un meccanismo di accesso completo in cui tutta la risorsa è disponibile per la classe a qualità garantita. Questa riceverà dal RN, il dovuto e successivamente riceverà ulteriore risorsa in proporzione al suo stato delle code. Se lo stato delle code di questa classe fosse troppo elevato, l'ovvia conseguenza sarebbe un'assegnazione sconsiderata alla classe GB, che non tiene conto della presenza di eventuali MAC-PDU di classe BE in coda, che subiscono un ritardo eccessivo. Al fine di ottenere un trattamento equo, che non consenta troppe assegnazioni in più alla classe GB e verificata, l'effettiva "sofferenza" della classe BE, quello che si fa è passare ad una situazione di scheduling a partizionamento completo in cui una buona parte della banda viene assegnata alle GB ma una parte è riservata "strettamente" alle BE e solo a loro. Chiaramente un approccio di questo genere penalizza le sorgenti a priorità più bassa in misura minore rispetto al

meccanismo di accesso completo. Pur essendo riservata una quantità di risorsa alla BE, può accadere che questa non sia sufficiente. Quello che, a questo punto, si può verificare, è che se le sorgenti BE hanno bisogno di più risorsa e se le sorgenti GB non occupano tutta la quota parte di risorse loro assegnate, si passa all'ultimo meccanismo di assegnazione della risorsa, l'"accesso ristretto". Al fine di evitare sotto-utilizzo della banda, si lascia sempre una piccola parte riservata alle sorgenti a priorità inferiore (le sorgenti BE), ma viene resa disponibile a queste anche quella parte di banda non sfruttata. Chiaramente di un tale algoritmo di scheduling si dovrebbero avvantaggiare le sorgenti BE a più bassa priorità che in tale sistema sono quelle che soffrono più di tutte. Il ciclo riparte dall'"accesso completo" se si percepisce che le BE non sfruttano tutta la banda a disposizione, rendendo di nuovo disponibili alla classe GB tutte le risorse del sistema. Il rimpallo da un algoritmo all'altro è rappresentativo dell'adattività alle condizioni di carico della rete.

CAPITOLO 7

RISULTATI

La prima valutazione che è stata fatta per la verifica del corretto funzionamento del simulatore si è basata su una calibratura dello stesso. Imposto il numero medio di nascite in un'ora, si è fatta una serie di simulazioni con numero di trame di volta in volta maggiori per essere in grado di stimarne il comportamento. Si è voluto capire se al variare del numero delle trame il comportamento del sistema in relazione alle nascite migliorasse, peggiorasse o approssimasse in maniera corretta il valore reale, qualunque fosse il numero di trame considerato.

Si è constatato che il comportamento medio, a parte poche eccezioni, si avvicina moltissimo all'andamento reale e ciò accade indipendentemente dal numero di trame simulate. Si è verificato, cioè, che qualsiasi sia il numero di trame imposto nella particolare simulazione, il modello realizzato approssima molto bene quello reale. Sulla base di queste premesse viene mostrato il risultato di 5 set di simulazioni compiute con un numero di trame pari a 10000, 20000, 30000, 40000 e 50000, avendo ipotizzato che il numero medio di nascite in un'ora fosse pari a 120 sorgenti per ogni classe di traffico.

I risultati attesi in relazione alle premesse fatte sono di 12 nascite nel caso di 10000 trame, di 24 nascite nel caso di 20000 trame, di 36 nascite nel caso di 30000 trame, di 48 nascite nel caso di 40000 trame e di 60 nascite nel caso di 50000 trame. Tali valori vanno sempre considerati in media. I grafici di seguito mostrano il dettaglio di quanto appena descritto. Il primo grafico è rappresentativo del valore reale di sorgenti che dovrebbero nascere rispettando il valore imposto inizialmente (120 sorgenti per ogni classe in un'ora) e del valore delle nascite ottenuto mediando i risultati dei set delle 5 simulazioni considerate. Il secondo grafico, mostrando il dettaglio delle singole prove, è rappresentativo del numero di nascite effettivo ottenuto nelle varie simulazioni eseguite (5 simulazioni per 10000 trame, 5

simulazione per 20000 trame, etc). Come si può facilmente constatare, valori reali, mediati e simulati sono tutti molto vicini tra loro.

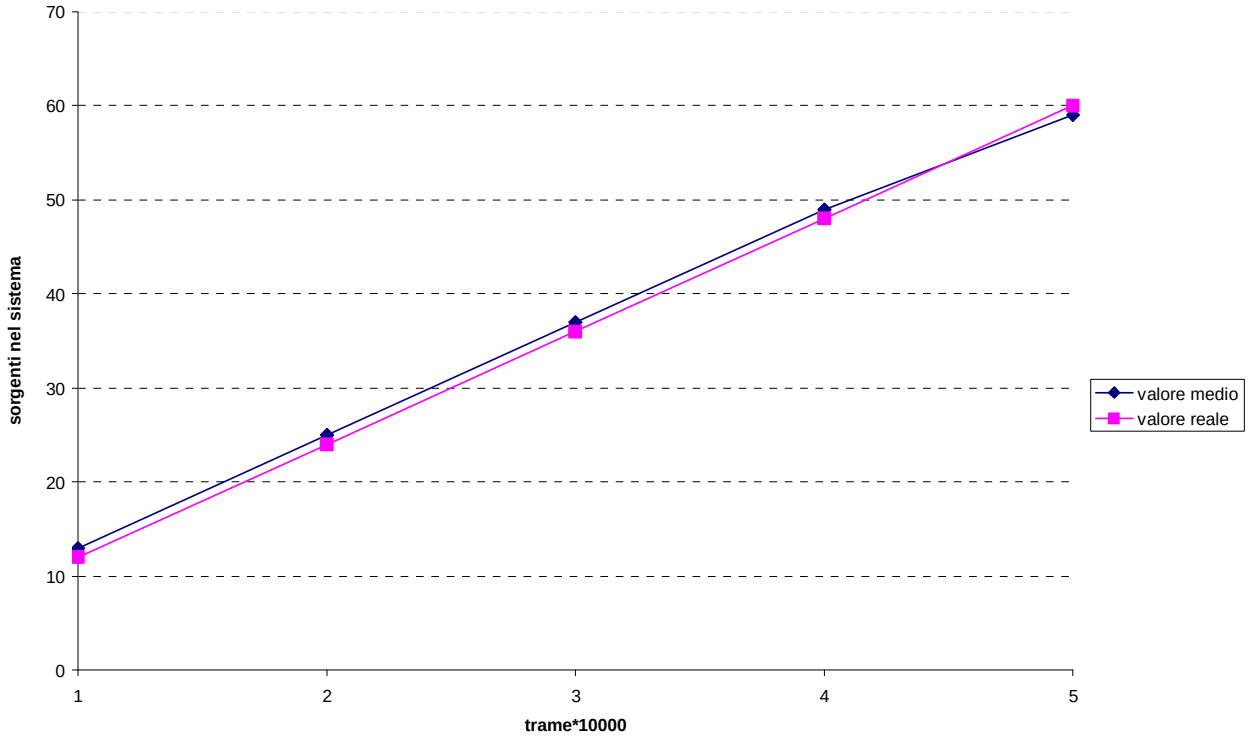


Grafico 1 – Valore reale e mediato sulle simulazioni compiute del numero di nascite al variare del numero di trame simulate.

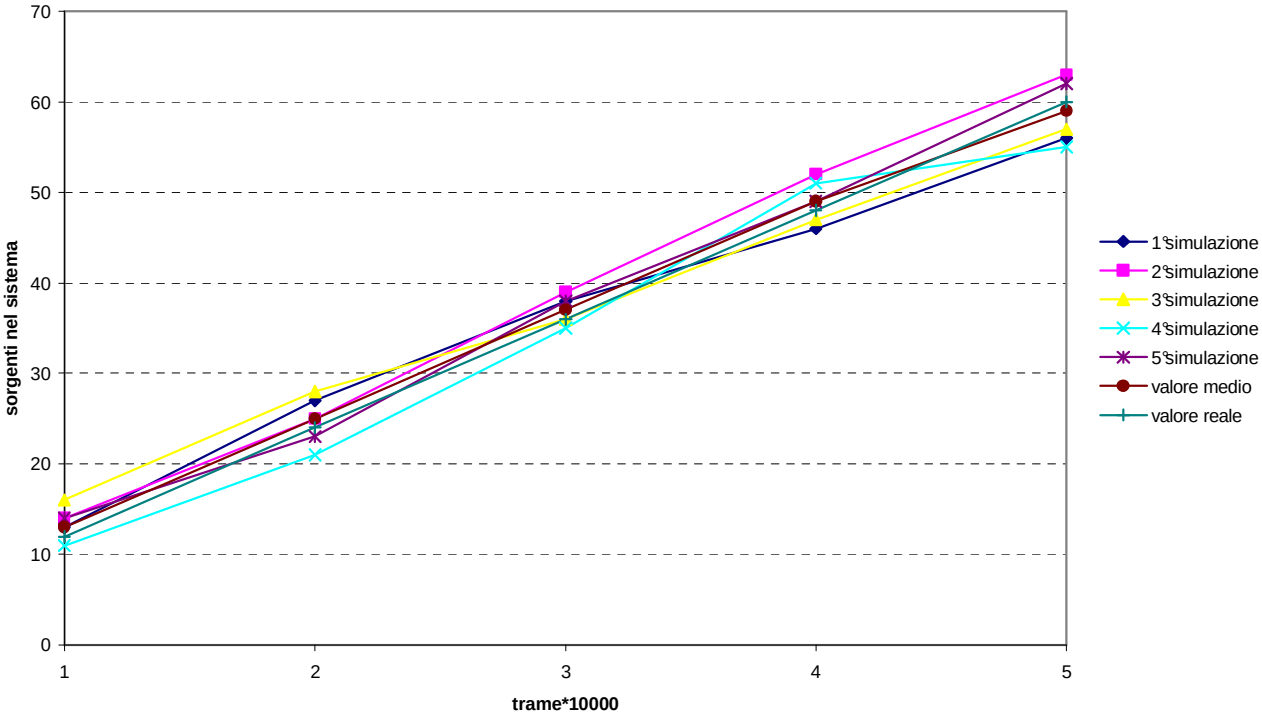


Grafico 2 – Dettaglio delle singole simulazioni

La verifica della corretta “calibratura” del sistema ci ha consentito di scegliere molto liberamente il numero di trame da considerare per i risultati successivi.
Si è così fissato un numero di trame per tutte le simulazioni valutate di seguito pari a 30000 che corrispondono a 4,5 minuti.
Il parallelo, che è stato affrontato nel seguito, è tra l’algoritmo di scheduling “statico”, in cui sono stati inseriti i processi di nascita e morte, lasciando inalterato l’algoritmo di scheduling proposto nella prima versione del simulatore, dove veniva lasciata piena libertà alle risorse a qualità garantita di occupare tutta la risorsa disponibile, anche quella non assicurata, e l’algoritmo di scheduling dinamico proposto e descritto nel paragrafo 5.3, in cui invece vengono limitati i privilegi concessi alle risorse GB a favore delle risorse BE.
Chiaramente per quanto concerne i processi di nascita e morte i due algoritmi si comportano nella medesima maniera; le differenze che ci si aspettano sono, invece, relative all’assegnazione delle risorse.
La verifica di un andamento variabile nel tempo delle nascite e delle morti delle sorgenti è stata fatta trama per trama per il sistema complessivo, sia nel caso di simulazione in ambiente dinamico che in caso di simulazione in ambiente statico, ed è stato compiuto un ulteriore controllo all’interno di ogni singola classe di traffico per verificare la rispondenza tra i risultati totali del sistema e quelli “parziali” delle singole classi. Anche in questo caso, la risposta del sistema è stata molto buona in quanto è facile notare, in ogni simulazione mostrata nel seguito, un effettivo cambiamento dello stato del sistema durante l’evolversi del tempo. Caratteristica su cui è necessario soffermarsi un po’ è relativa alla morte di una sorgente ed in quanto tempo questa avvenga. Valutando i grafici, spesso è possibile notare come compaiano dei picchi. Questi ultimi rappresentano la nascita e la morte di una sorgente avvenuta nell’intervallo di poche trame. La cosa non deve destare sospetto dal momento che è possibile che si verifichino due condizioni: il traffico da trasferire è poco; la sorgente trova il sistema scarico ed emettendo ad un rate elevato riesce a smaltire il traffico in breve tempo. Qui la spiegazione dei picchi che si

notano sia nella rappresentazione delle nascite e morti nello scenario statico sia in quella nello scenario dinamico.

Simulazione 1

Questa simulazione è stata fatta considerando che il traffico generato dalle varie sorgenti non fosse eccessivamente pesante. Chiaramente, come si vedrà anche nelle simulazioni successive, non è possibile ottenere un andamento delle nascite e delle morti nonché un valore di traffico emesso esattamente uguale per il caso statico e quello dinamico. I confronti effettuati vanno sempre valutati tenendo presente che si sono considerati dei risultati che in media riproducessero le stesse condizioni generali. Tutto questo perché non solo le nascite sono dettate dal caso, ma lo stesso traffico emesso dipende dall'estrazione di una v.a. di distribuzione geometrica, che di volta in volta assume valori differenti, pur essendo generata a partire dalla conoscenza della dimensione media di traffico sempre uguale.

Cominciamo con il confronto tra le nascite/morti delle sorgenti nel caso statico e le nascite/morti nel caso dinamico.

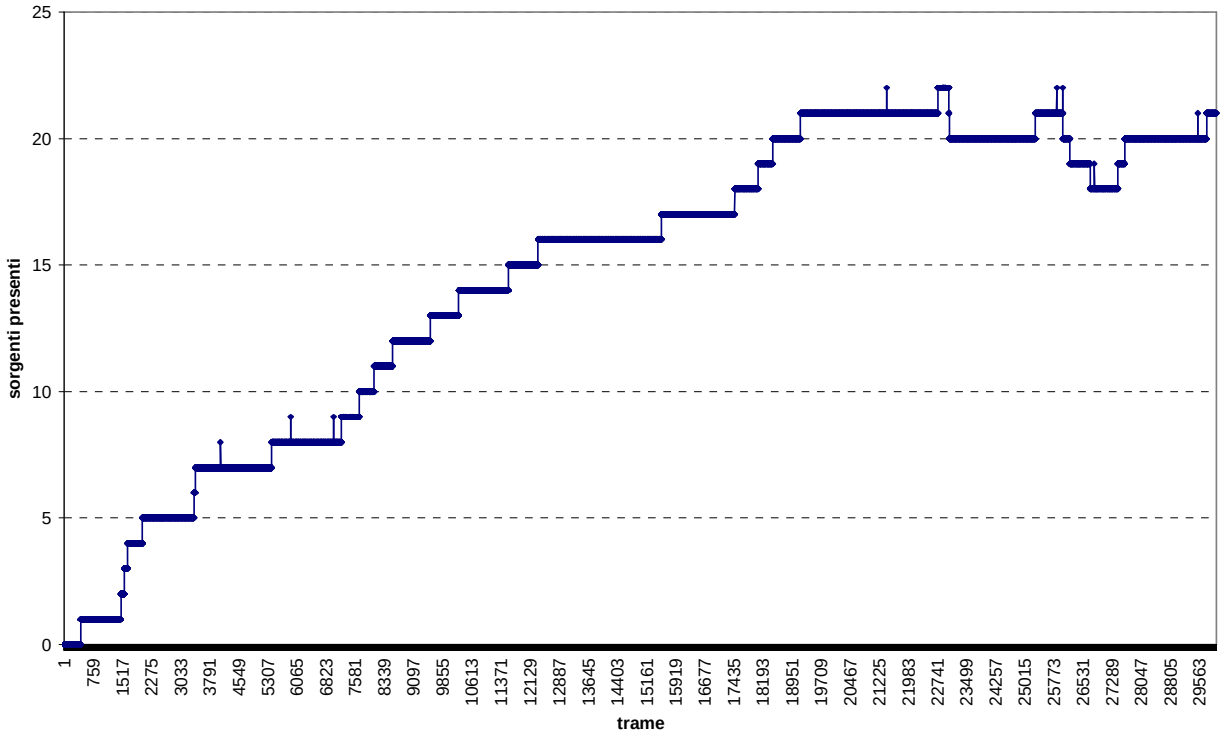


Grafico 3 – Sorgenti presenti nel sistema (caso statico)

Il numero di sorgenti nate in totale, nel caso statico, è pari a 35. Ricordo che il valore calcolato era di 36, di conseguenza questa simulazione dal punto di vista delle nascite rispetta bene il modello. Il dettaglio sulle nascite è dato da 8 sorgenti ON-OFF, 11 sorgenti CBR, 8 sorgenti BE al alta priorità e 7 sorgenti BE a bassa priorità

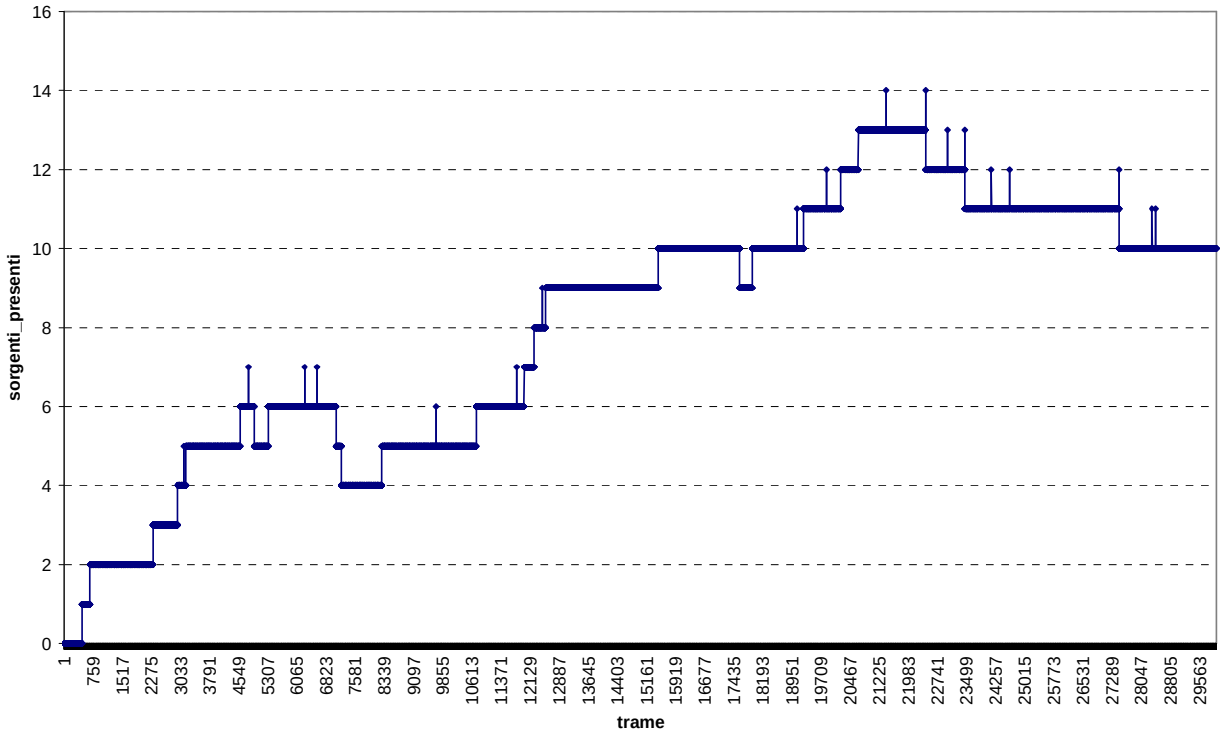


Grafico 4 – Sorgenti presenti nel sistema (caso dinamico)

Anche il numero di sorgenti nate in totale nel caso dinamico è pari a 35. Il dettaglio sulle nascite è dato da 8 sorgenti ON-OFF, 11 sorgenti CBR, 7 sorgenti BE al alta priorità e 9 sorgenti BE a bassa priorità.

Dal punto di vista dei processi di nascita e morte i due sistemi si comportano in maniera molto simile come era naturale attendersi e come si può constatare dal dettaglio sulle nascite in entrambe i sistemi. L'andamento temporale delle nascite/morti è chiaramente differente, ma anche questo era un risultato atteso dal momento che i valori di traffico generati nei due casi è molto improbabile che risultino gli stessi.

Passando ora alla assegnazione della banda per entrambe i casi statico e dinamico, vediamo se si verificano grandi differenze di comportamento tra i due.

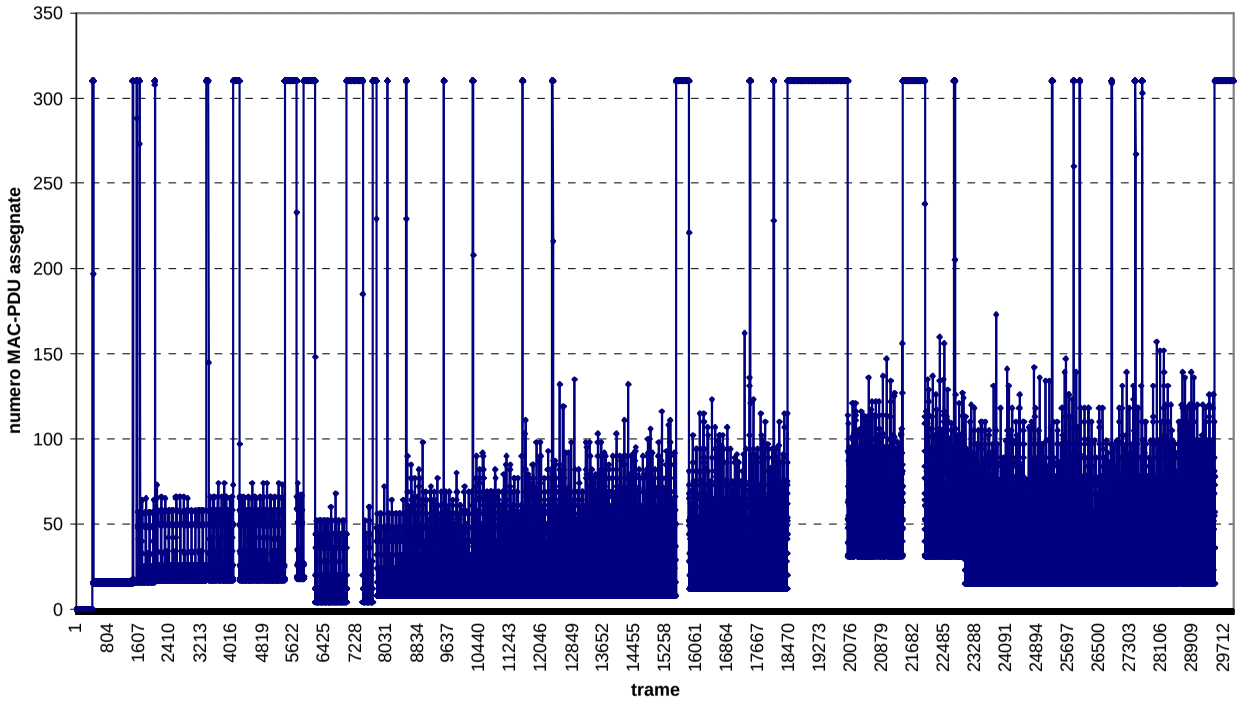


Grafico 5 – Banda assegnata totalmente (caso statico)

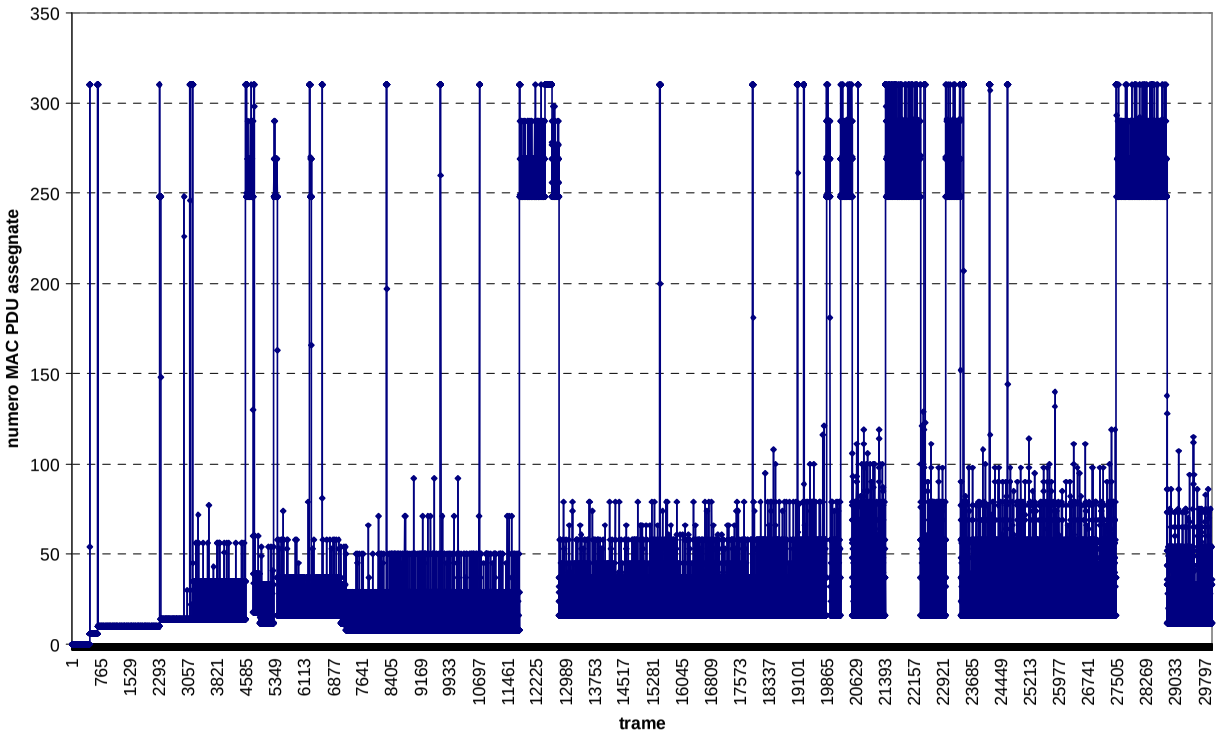


Grafico 6 – Banda assegnata totalmente (caso dinamico)

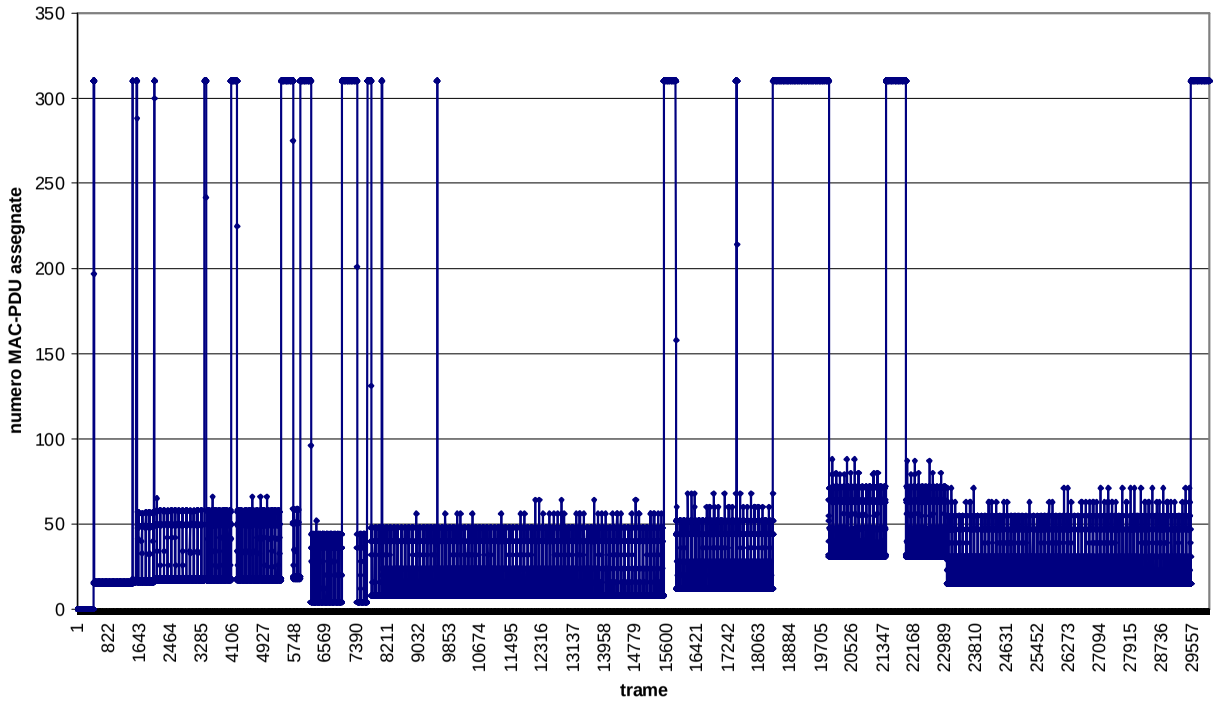


Grafico 7 – Assegnazione banda per le risorse GB (caso statico)

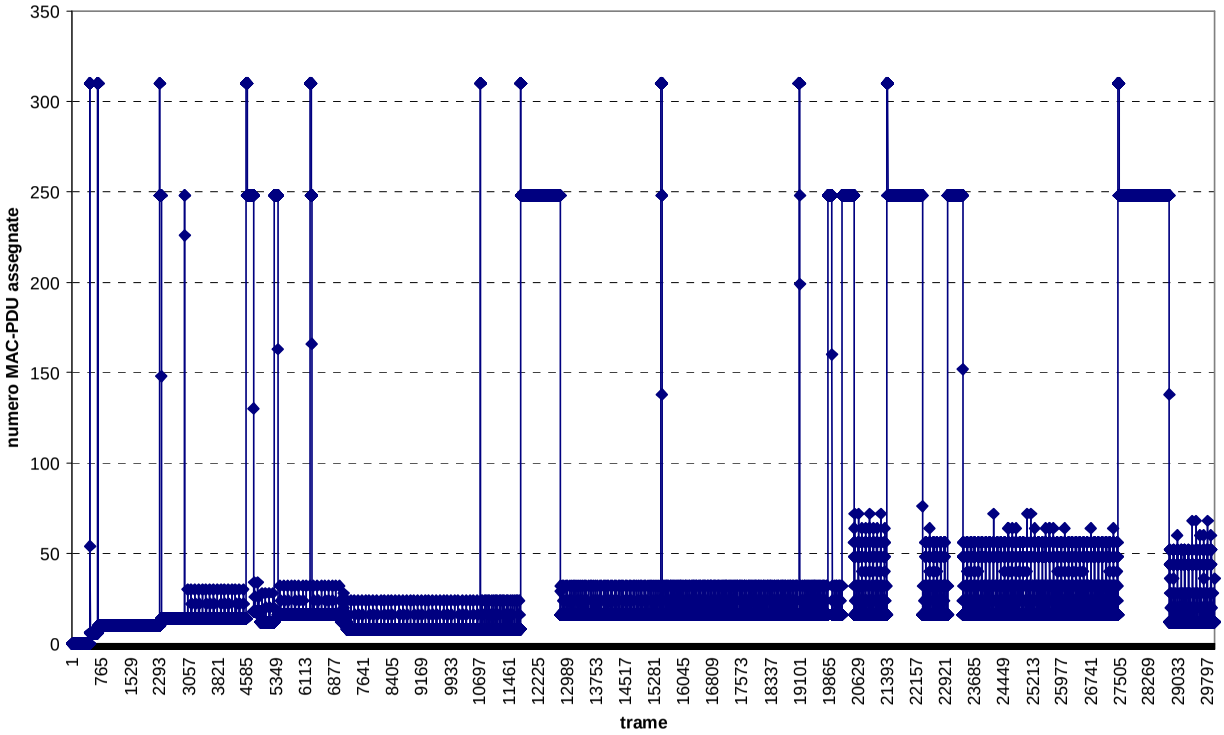


Grafico 8 – Assegnazione banda per le risorse GB (caso dinamico)

I grafici 5 e 6 mostrano l'assegnazione totale di banda nel sistema nel caso statico e nel caso dinamico. Come si può facilmente notare, nel caso statico questa simulazione ha prodotto un'assegnazione di risorsa leggermente maggiore. Questo è dovuto da un lato al fatto che la quantità di PDU generate, nel caso statico, è leggermente più grande e dall'altro al diverso meccanismo di scheduling usato. Si cominciano a percepire le prime differenze nell'allocazione della banda, soprattutto nei momenti in cui il sistema risulta carico.

I grafici 7 ed 8 mettono invece a confronto la sola assegnazione di risorsa per la classe a qualità garantita. Anche qui si nota una maggiore assegnazione delle GB nel caso statico, imputabile alla quantità di traffico in più generato nell'ambito di questa simulazione ed in parte alla minore assegnazione di code alle risorse GB da parte dell'algoritmo di scheduling adattativo implementato. Il confronto compiuto tra le due simulazioni si è basato poi sulla valutazione di parametri globali come ad esempio il rapporto tra le richieste effettuate dalla classe GB e quelle accettate. I valori di richieste riscontrati sia per la simulazione in ambiente statico che per la simulazione in ambiente dinamico sono risultati uguali e pari a 19. Il numero di richieste accolte, identico a quelle fatte, anch'esso uguale nel caso statico e nel caso dinamico, ci ha permesso di concludere che, data la situazione di carico non troppo pesante, il sistema è stato in grado di accogliere quanto richiesto.

Altri due parametri valutabili in relazione alla risorsa a qualità garantita, nel caso statico ed in quello dinamico sono:

- la percentuale di MAC-PDU GB trasmesse in eccesso rispetto a quelle garantite per trama, il cui dettaglio nela simulazione risulta dato da:
  % Eccesso medio per trama (caso statico)=46,4%
  % Eccesso medio per trama (caso dinamico)=42,3%
- la percentuale di MAC-PDU GB presenti in coda eccedenti un limite stabilito a 5000, il cuidettaglio nella simulazione risulta dato da:
  % Overflow GB rispetto al riferimento (caso statico)=8.6%
  % Overflow GB rispetto al riferimento (caso dinamico)=6,9%

La percentuale degli eccessi è analoga. Questo perché il sistema presenta un grado di utilizzazione non tanto alto e, quindi, non si riesce a percepire una grande differenziazione nelle assegnazioni di quanto non assicurato alla classe GB tra il caso statico e quello dinamico. Anche la differenza negli overflow è contenuta e può essere attribuita ad un numero totale maggiore di MAC-PDU prodotte nel caso statico ed in misura minore ad un contenuto miglioramento apportato dall'algoritmo di scheduling adattativo implementato nel caso dinamico.

Sempre in questo ambito, vengono di seguito mostrati due grafici: uno relativo alle successive variazioni di scheduling nel tempo, l'altro rappresentativo delle volte in cui è stato preferito un tipo di scheduling ad un altro.

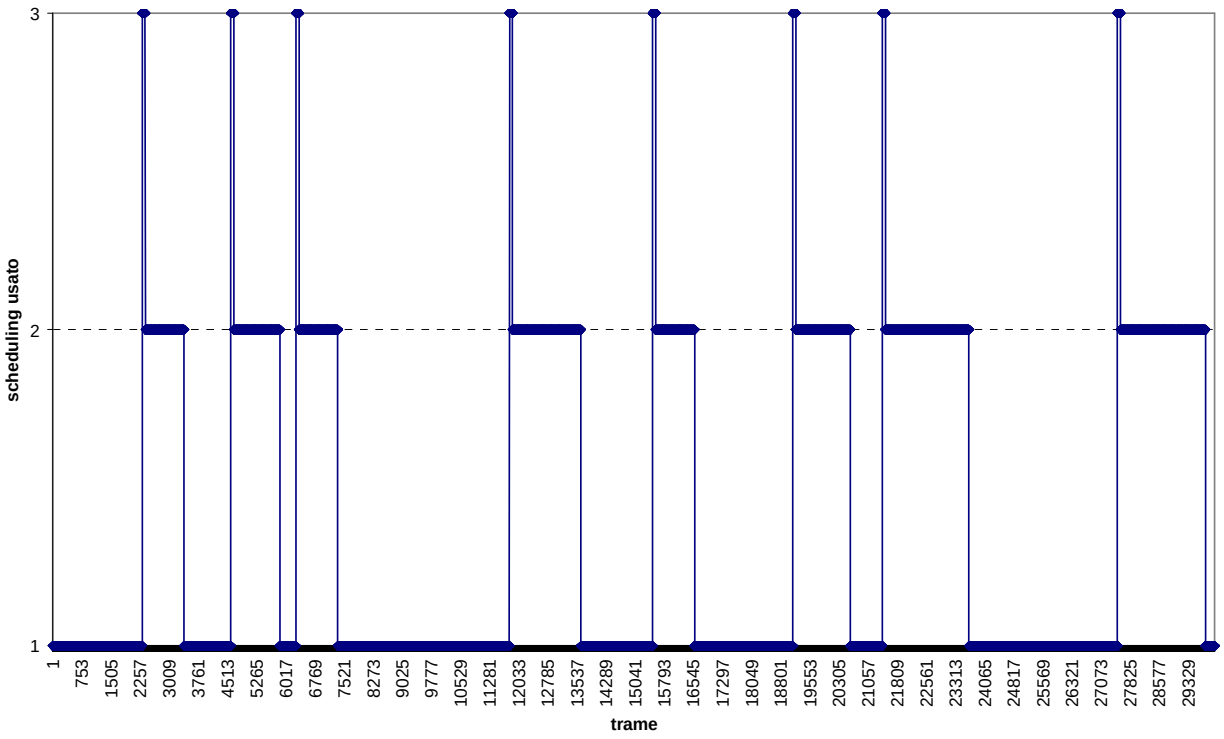


Grafico 9 – Confronto dei vari schemi di accesso: 1 rappresenta l'accesso completo, 2 rappresenta l'accesso ristretto, 3 il partizionamento completo

Ricordiamo, come mostrato anche nella didascalia del grafico 9, che nello scheduling dinamico, sono previsti 3 differenti stati: uno di partizionamento completo della risorsa, uno di accesso ristretto e l'ultimo di accesso completo (rif. paragrafo 5.3).

Dal grafico 9 si può notare come la permanenza nello stato a “partizionamento completo” sia molto breve, poche trame. Mediamente ci troviamo molto nello stato ad “accesso completo” e un po’ meno in quello ad “accesso ristretto”.
Ciò nonostante si osserva che il protocollo in certi intervalli temporali sente la necessità di garantire almeno il 20% di banda alla classe best effort.
Questo avviene senza dover rifiutare eventuali richieste GB. Il 20% precedentemente detto continua a essere garantito anche durante l’”accesso ristretto” in cui vengono concesse alla classe BE anche le risorse lasciate inutilizzate dalla classe GB. Qualora e questo accade molto spesso le stesse risorse BE non necessitino di quanto fornito si ritorna all’accesso completo, garantendo alla classe GB più di quanto avuto per diritto.

Il grafico 10 mostrato di seguito mostra la percentuale di utilizzazione di ciascun algoritmo. Come già appurato dal grafico precedente, gran parte del tempo è occupato dagli stati CA e RA: questa è una confermo del comportamento equo dell’algoritmo di scheduling nei riguardi sia delle risorse GB che delle risorse BE, che limitando leggermente l’assegnazione in eccesso fornita alle GB, ricevono una quota parte di banda.

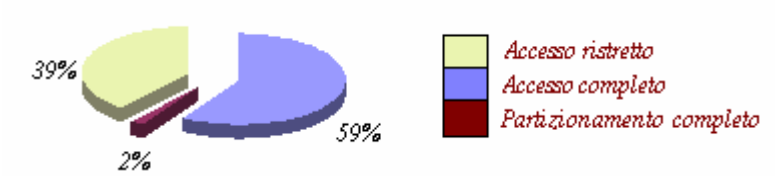


Grafico 10 - Percentuale di utilizzazione di ogni algoritmo di scheduling

Passiamo ora al commento dei ritardi medi relativi alla classe BE riscontrati per le simulazioni in ambiente statico e dinamico.

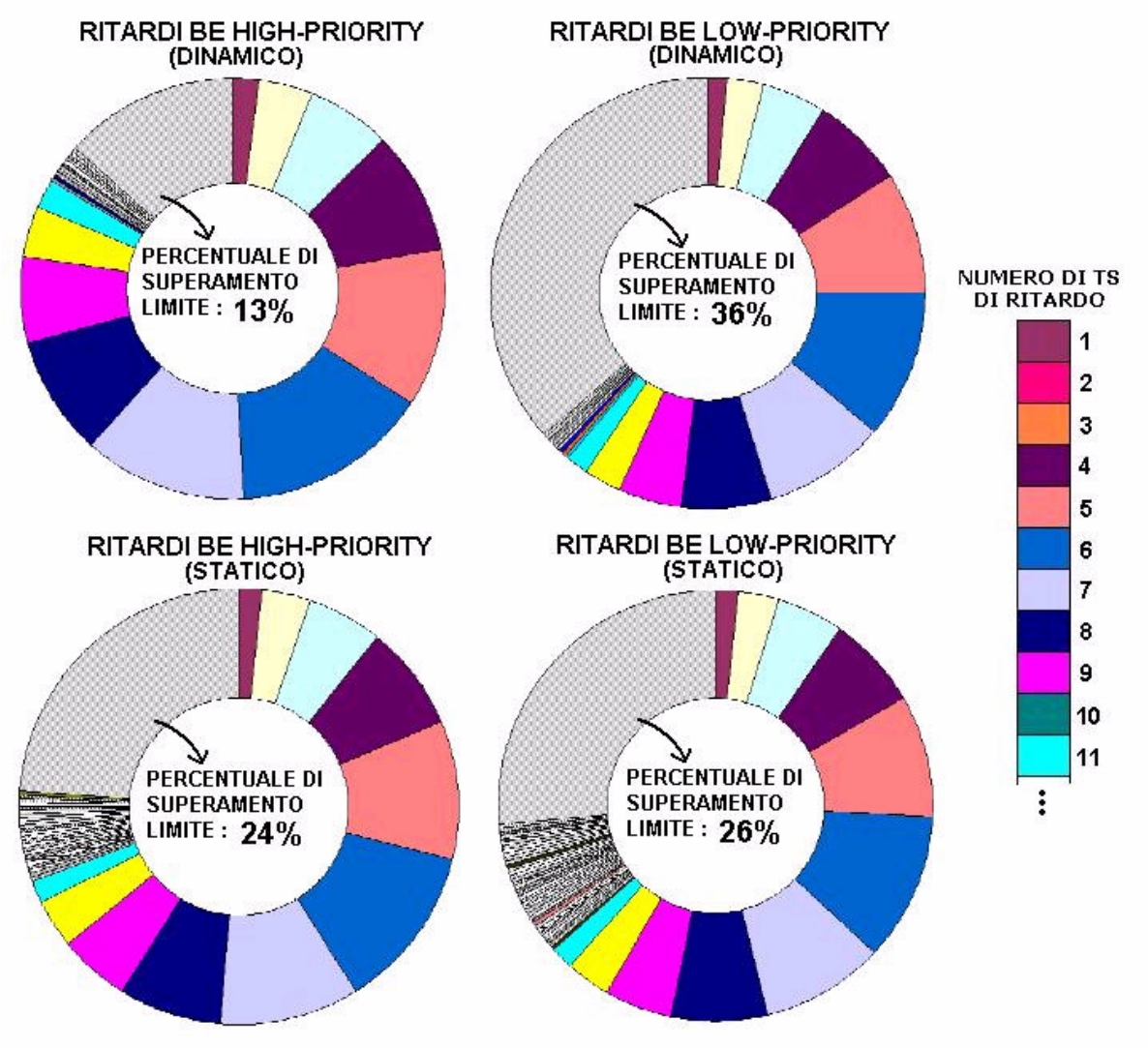


Grafico 11 – Ritardo BE Statico vs Ritardo BE Dinamico

Questi grafici vanno letti come una sorta di istogramma. Rappresentano la percentuale di volte in cui c'è stato 1 TS di ritardo, 2 TS di ritardo,…N TS di ritardo. L'unico valore di cui viene mostrata la percentuale nel grafico è il numero di volte che il traffico BE generato ha subito più di 8000 TS di ritardo. Il valore imposto come massimo ritardo non deve spaventare dal momento che i tempi in gioco sono molto brevi. Date queste informazioni, passiamo al commento del

grafico. Si può osservare che in relazione alla disposizione dei grafici si possono fare due analisi, una orizzontale, l'altra verticale.

L'analisi orizzontale permette un confronto del ritardo subito dalle MAC-PDU delle due code BE, l'una ad alta priorità e l'altra a bassa priorità nell'ambito del singolo algoritmo. L'analisi verticale consente invece di fare un confronto sui ritardi subiti dalle code BE di stessa priorità, alta e bassa, in relazione all'utilizzo dell' algoritmo statico o dinamico.

Per quanto riguarda l'analisi orizzontale, nel caso dinamico, si osserva una maggiore differenza di comportamento tra le BE ad alta priorità (HP) e le BE a bassa priorità (LP). Il protocollo dinamico è maggiormente in grado di assicurare un differente servizio alle due categorie BE, assegnando più risorse alla classe BE HP rispetto alle più penalizzate risorse BE LP. Per quanto riguarda invece l'algoritmo statico, il comportamento non mostra una differenziazione così netta tra BE HP e BE LP, che mostrano un valore di massimo ritardo decisamente confrontabile. L'Analisi Verticale, dall'altro lato, mostra come confrontando le due code nei due differenti algoritmi, particolari vantaggi dell'algoritmo dinamico rispetto a quello statico siano riscontrabili esclusivamente nel caso delle code BE ad alta priorità. Qui effettivamente si assiste ad una percentuale di ritardo massimo minore nel caso dinamico. Si comincia in sostanza ad intravedere una sorta di miglioramento introdotta con il nuovo algoritmo adattativo, anche se la stessa cosa non si può affermare per le BE a bassa priorità. Oltre ai ritardi sono state valutate anche le percentuali di overflow per la classe BE HP e BE LP, nei due scenari statico e dinamico.

La percentuale di MAC-PDU BE presenti in coda eccedenti un limite stabilito a 5000 risulta data da:

% Overflow BE HP rispetto al riferimento (caso statico)=25,6%

% Overflow BE HP rispetto al riferimento (caso dinamico)=31,50%

% Overflow BE HP rispetto al riferimento (caso statico)=53,7%

% Overflow BE LP rispetto al riferimento (caso dinamico)=46,1%

Come si può constatare dai valori calcolati, in media si ottiene un miglioramento della percentuale di MAC-PDU in coda. Questo avviene in misura maggiore per le BE ad alta priorità rispetto alle BE a bassa priorità, quasi ad indicare una sorta di relazione ritardi - overflow.

Simulazione 2

Seguendo lo stesso percorso della simulazione precedente, questa simulazione è stata fatta considerando gli stessi identici parametri dell'altra ma aumentando la durata media e quindi la dimensione del traffico di ogni tipo di sorgente generata. La simulazione ha portato a dei risultati decisamente significativi, che hanno permesso di apprezzare in misura maggiore rispetto alla prima, i vantaggi ottenibili tramite un algoritmo di scheduling adattativo. Ripercorrendo le tappe considerate nella simulazione precedente, partiamo dalle nascite, nei due algoritmi di scheduling.

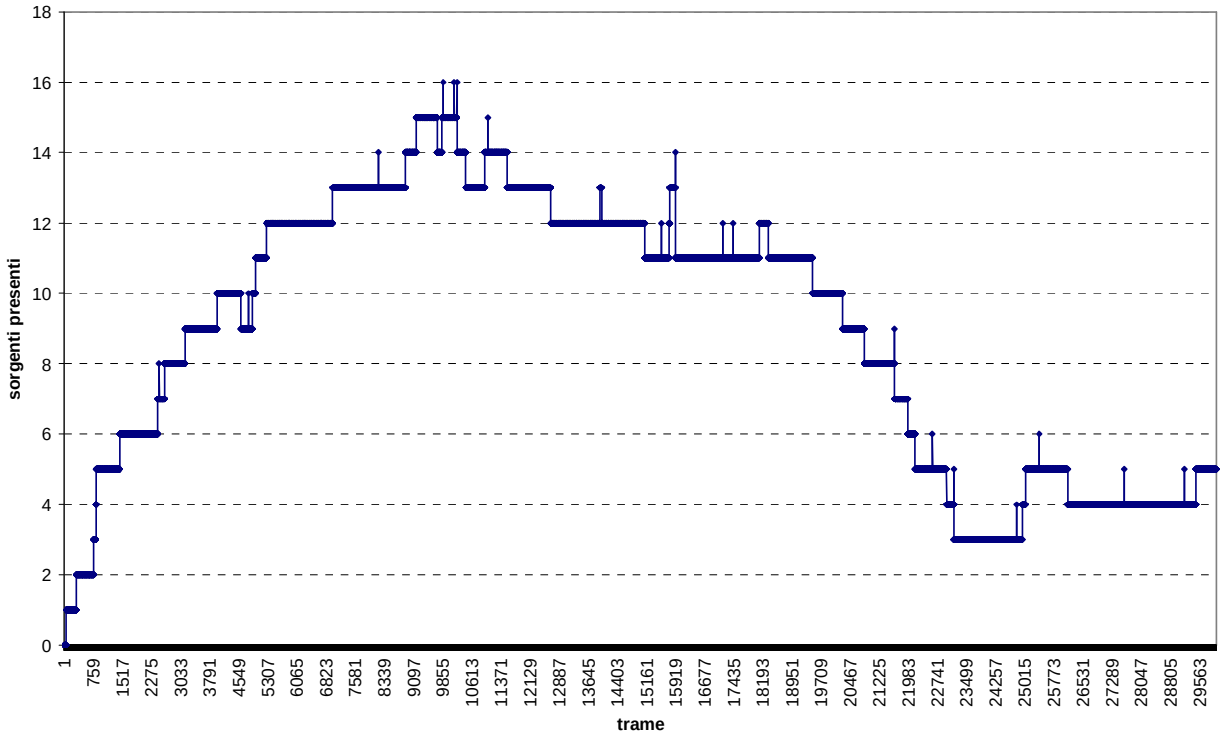


Grafico 12 – Sorgenti presenti nel sistema (caso statico)

Le sorgenti nate complessivamente nel caso statico sono 38, di cui 7 ON OFF, 9 CBR, 10 BE HP, 12 BE LP.

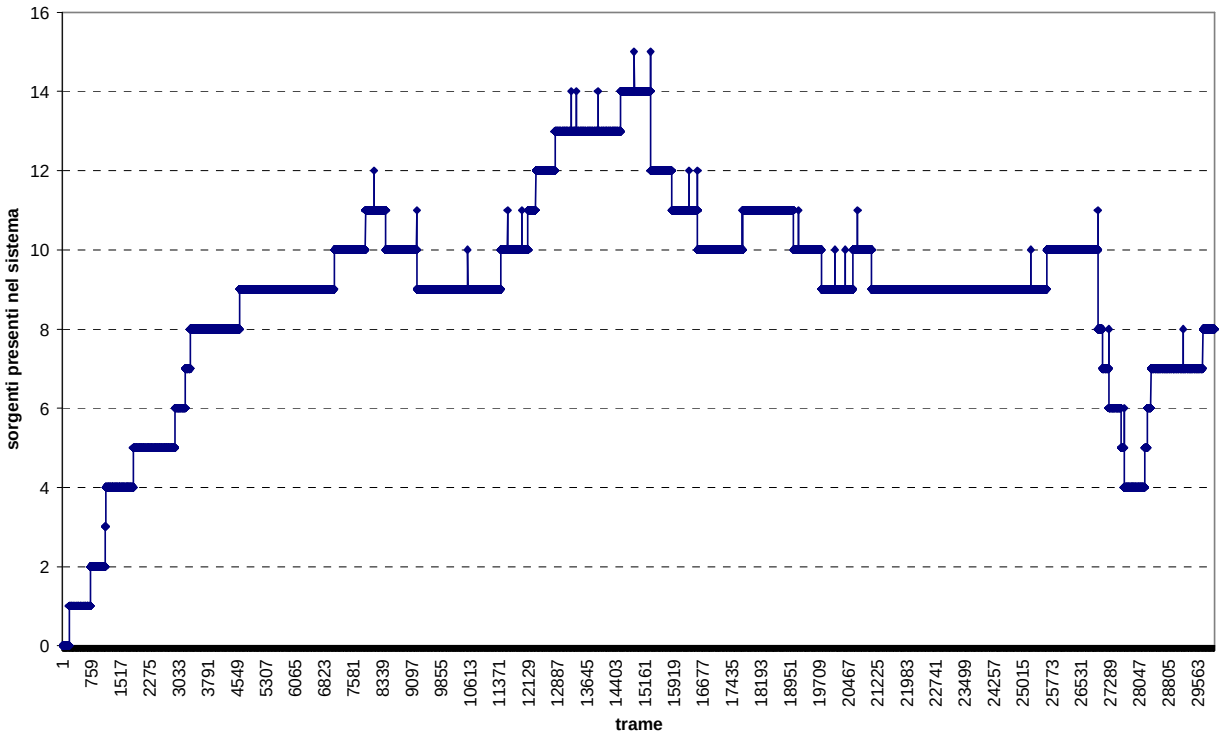


Grafico 13 - Sorgenti presenti nel sistema (caso dinamico)

Le sorgenti nate nel caso dinamico sono 39, di cui 9 ON OFF, 8 CBR, 10 BE HP e 12 BE LP. Anche in questa seconda serie di simulazioni il numero di nascite approssima abbastanza bene il valore reale, pari a 36.
Il dettaglio dell'andamento nel tempo non mostra grandi analogie tra le nascite nei due algoritmi. I risultati di questa simulazione sono sembrati, però, molto significativi dal momento che il numero di MAC PDU generate durante la simulazione nei due casi statico e dinamico è paragonabile.

Passando quindi a valutazioni prestazionali cominciamo con il mostrare i grafici relativi all'assegnazione totale della risorsa ed in particolare alla occupazione della banda GB per entrambe i casi studiati.

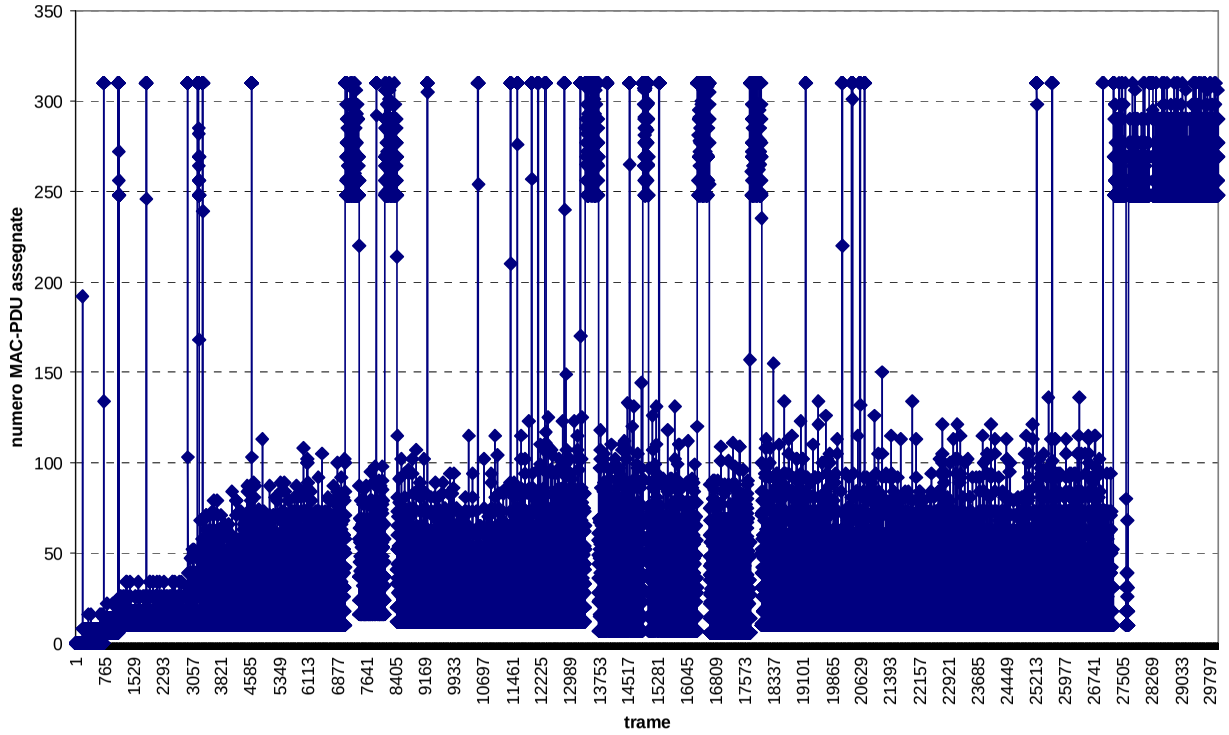


Grafico 14 – Banda assegnata totalmente (caso statico)

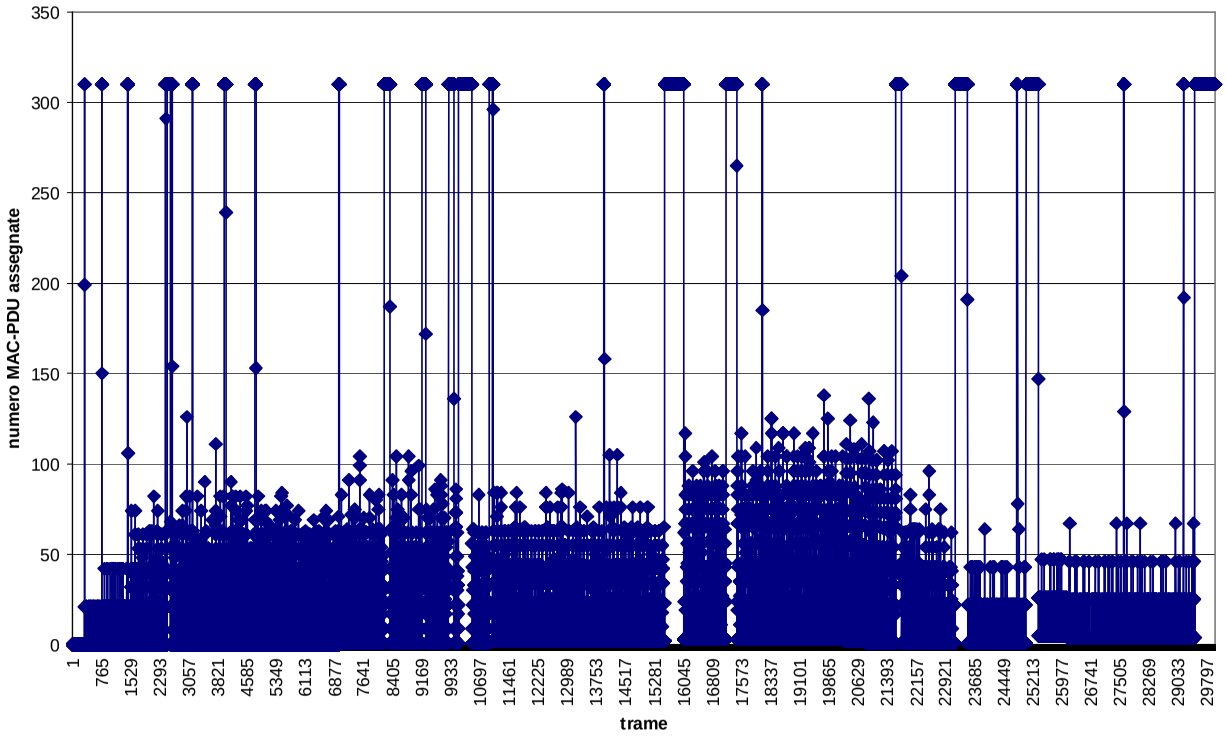


Grafico 15 – Banda assegnata totalmente (caso dinamico)

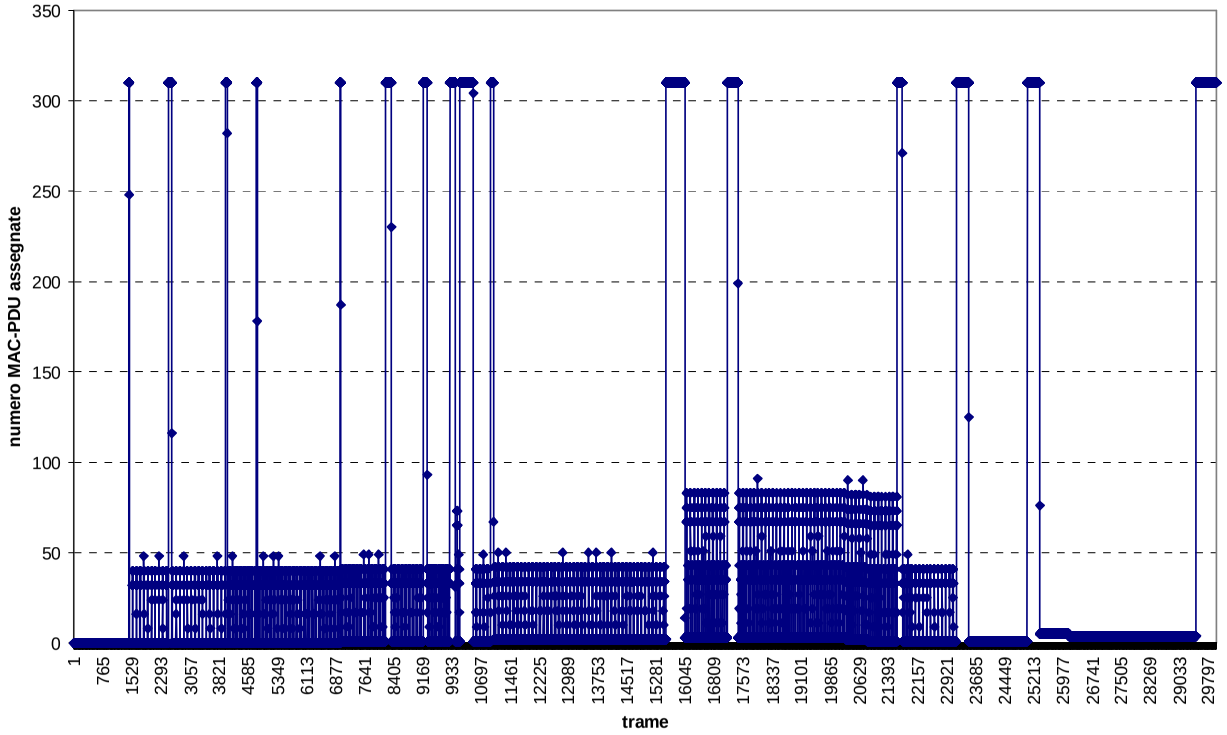


Grafico 16 - Assegnazione banda per le risorse GB (caso statico)

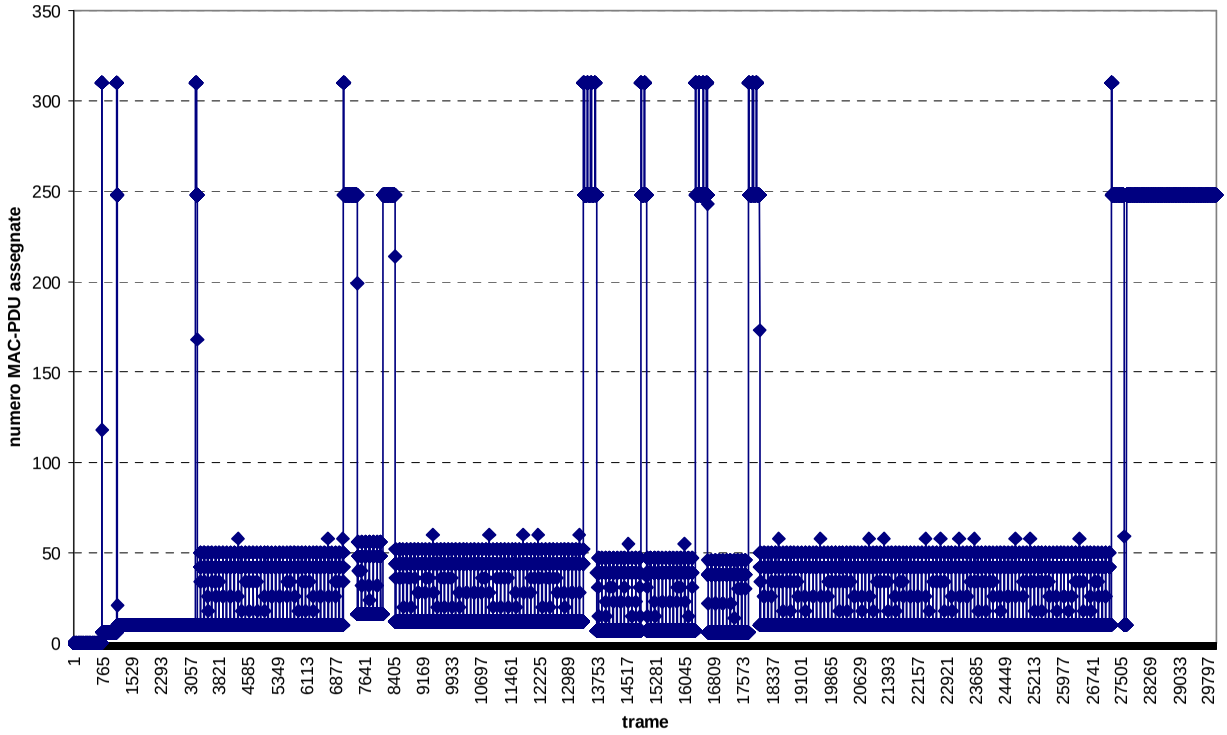


Grafico 17 – Assegnazione banda per le risorse GB (caso dinamico)

Dall'analisi che è possibile fare dei grafici di occupazione totale della banda, ovvero i grafici 14 e 15, si può notare rispetto al caso precedente un incremento nell'occupazione della banda. Il risultato è coerente con l'aver introdotto una quantità di traffico maggiore rispetto al caso precedente. Per quanto concerne le occupazioni da banda totali, queste risultano molto simili. Le occupazioni GB, invece mostrano una quantità di risorsa occupata nel caso dinamico quasi sempre minore rispetto al caso statico, tranne verso la fine del grafico, in cui si nota una impennata dell'occupazione GB fissa ad un valore pari a circa 240 MAC-PDU. Questa porzione del grafico è rappresentativa in quanto mostra come nel periodi suddetto il crescente numero di MAC-PDU GB che sforano il pattuito, non vengano soddisfatte totalmente ma risultino limitate dal massimo valore di banda consentito dagli algoritmi CP e RA e pari all'80% della risorsa disponibile nella trama, per l'appunto 240 MAC-PDU. In tale zona, come successivamente mostrato si ha effettivamente uno stato di scheduling di tipo ad accesso ristretto.

Il confronto compiuto tra le due simulazioni si è basato poi sulla valutazione di parametri globali, come avvenuto per la prima simulazione, il primo dei quali è dato dal rapporto tra le richieste effettuate dalla classe GB e quelle accettate.

Nel caso della simulazione in ambiente statico il numero di richieste è pari a 16, nel caso dinamico 17. Il numero di richieste accolte in entrambe i casi è risultato identico al numero di richieste fatte. Entrambe i sistemi, nonostante la situazione di traffico più pesante introdotta, sono stati in grado di accogliere quanto richiesto.

Gli altri due parametri valutabili in relazione alla risorsa a qualità garantita, nel caso statico ed in quello dinamico sono:

- la percentuale di MAC-PDU trasmesse in eccesso rispetto a quelle garantite per trama, il cui dettaglio è dato da:

  % Eccesso medio per trama (caso statico)=50,02%

  % Eccesso medio per trama (caso dinamico)=43%
- la percentuale di MAC-PDU presenti in coda eccedenti un limite stabilito a 5000, il cui dettaglio è dato da:

  % Overflow GB rispetto al riferimento (caso statico)=3.85%

% Overflow GB rispetto al riferimento (caso dinamico)=8,67%

La percentuale degli eccessi presenta una buona differenza nell'assegnazione di quanto presente in più nelle code GB e non contrattato, tra l'algoritmo statico e quello dinamico. Quest'ultimo sembra verificare quindi quanto ci aspettavamo, ovvero una rigidità maggiore rispetto agli eccessi pur consentendo un grado di libertà maggiore alle risorse GB rispetto alle risorse BE. La differenza di circa 7 punti percentuali tra caso statico e dinamico è un risultato discreto.

Il fatto di aver consentito meno assegnazione di banda alle risorse Gb presenti in coda è avvenuta al prezzo di un aumento di overflow GB nel caso dinamico. Sembrerebbe necessario quindi a questo punto e come proposta futura la possibilità di realizzare una bufferizzazione variabile, che nel caso si verifichi un aumento di overflow dia la possibilità al buffer di contenere comunque una quantità superiore di MAC-PDU. La ripercussione ovvia dell'aumento dell'overflow GB si rifletterà in una diminuzione dell'overflow BE, come verrà mostrato nei grafici successivi.

Passando ora allo scheduling, analizziamo il comportamento dell'algoritmo adattativo in questo caso.

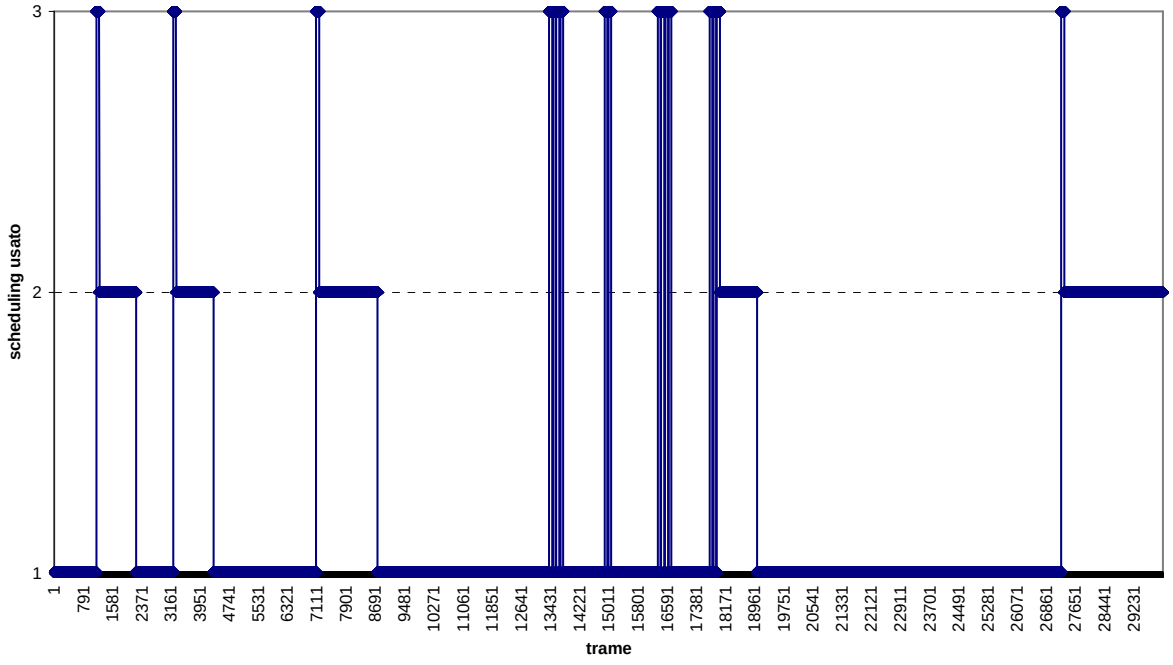


Grafico 18 – Confronto dei vari schemi di accesso:1 rappresenta l'accesso completo, 2 rappresenta l'accesso ristretto, 3 il partizionamento completo.

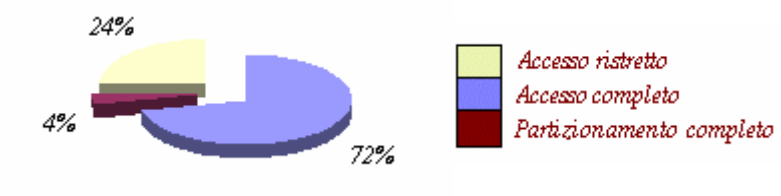


Grafico 19 – Percentuale di utilizzazione di ogni algoritmo di scheduling

Dal grafico 18 si osservano principalmente 2 fenomeni:

- a differenza di prima, dal "partizionamento completo" si ritorna direttamente all'"accesso completo". Questo probabilmente è dovuto al non rispetto dei vincoli di banda pattuiti con le GB piuttosto che alla mancanza di BE che chiedono risorse. Ciò si verifica in corrispondenza dei passaggi ripetuti dallo stato di partizionamento completo CP allo stato di accesso completo CA. Si deve notare però, facendo un parallelo con il grafico relativo all'assegnazione totale di banda, che probabilmente la quantità di MAC-PDU BE presenti in coda nel momento del passaggio dallo stato ad accesso completo a quello a partizionamento completo riesce in breve tempo ad essere smaltita, lasciando quindi della risorsa non utilizzata e permettendo il ritorno diretto allo stato di accesso completo, saltando la transizione per lo stato ad accesso ristretto.
- Verso la fine del grafico,come accennato precedentemente si nota che l'algoritmo utilizzato è di tipo ad accesso ristretto. La cosa non meraviglia dal momento che evidentemente l'occupazione della banda da parte delle risorse GB in eccesso viene limitata. Ciò mostra l'equità di assegnazione che si intende perseguire utilizzando un meccanismo di scheduling a stati.

Il grafico 19 mostra semplicemente la percentuale di occupazione di ogni schema di scheduling, relativamente alla simulazione valutata. Benché la percentuale di tempo in cui si è verificato accesso ristretto è minore rispetto al caso precedente, l'algoritmo in questa simulazione è risultato più incisivo.

Passando all'analisi dei valori considerati per le risorse BE, valutiamo i grafici relativi ai ritardi.

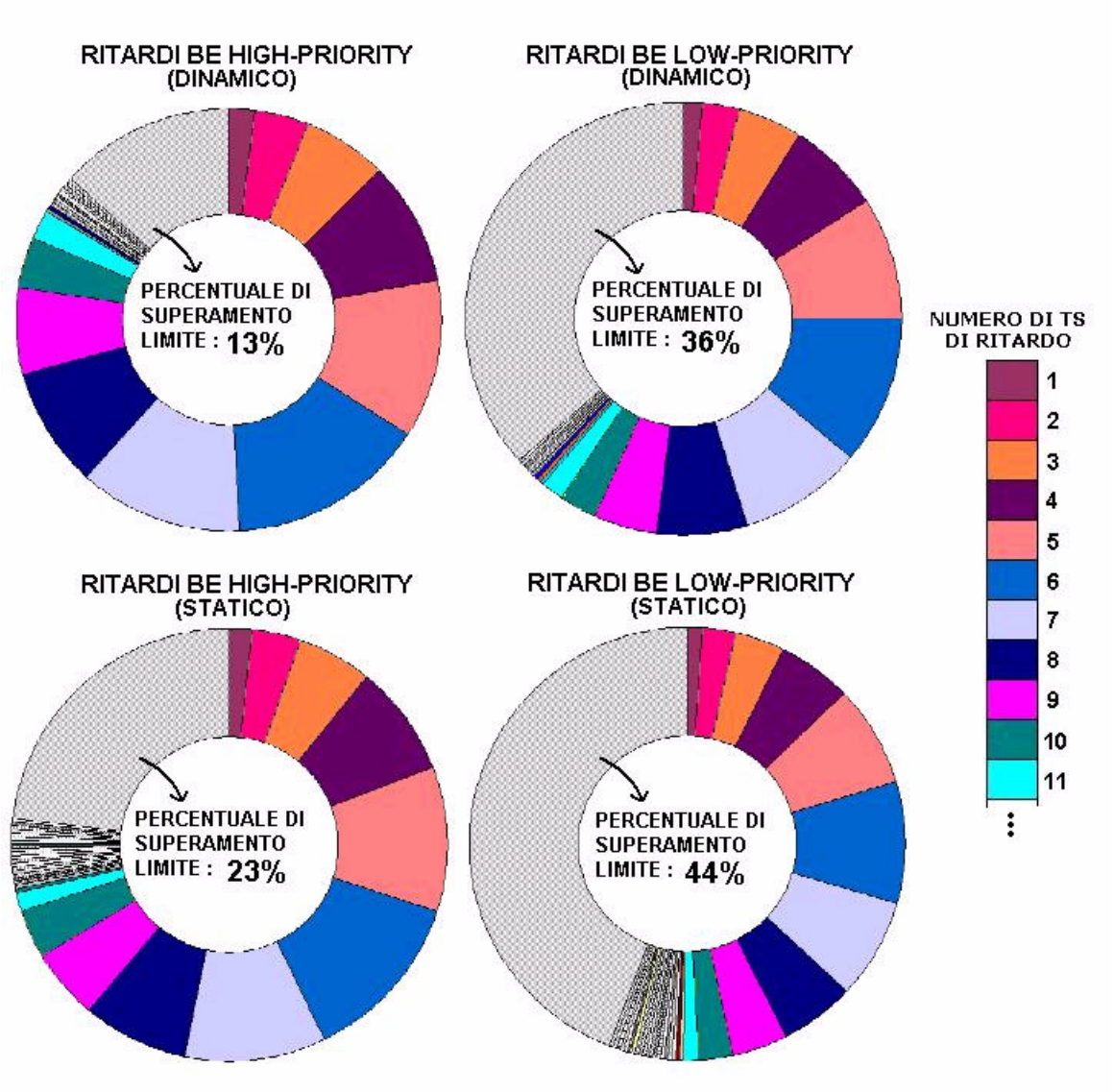


Grafico 20 – Ritardo BE statico vs Ritardo BE dinamico

Facciamo un confronto tra l'analisi orizzontale e quella verticale in entrambe i casi. In tutte e due gli scheduling è ancora più evidente la differenza di priorità che abbiamo introdotto. Le prestazioni sono superiori nel caso dinamico. Rispetto al caso precedente si assiste ad un netto miglioramento anche per quanto riguarda i ritardi a più bassa priorità, a differenza di quanto avveniva prima in cui sembrava ci fosse quasi un peggioramento per le risorse BE LP.

Osservando poi i valori di overflow delle BE ad alta priorità e quelli delle BE a bassa priorità, ci rendiamo conto che effettivamente il comportamento più rigido verso gli eccessi GB garantisce minore overflow per le PDU best effort. Ricordiamo che questo è dovuto all'aver introdotto degli stati nello scheduling nei quali è garantita almeno una parte minima alle risorse BE (20%).Entrando nel dettaglio dei valori, la percentuale di MAC-PDU BE presenti in coda eccedenti un limite stabilito a 5000 risulta data da:

% Overflow BE HP rispetto al riferimento (caso statico)=61,8%

% Overflow BE HP rispetto al riferimento (caso dinamico)=51,7%

% Overflow BE HP rispetto al riferimento (caso statico)=48,8%

% Overflow BE LP rispetto al riferimento (caso dinamico)=40,8%

Dai risultati mostrati, come ci si poteva aspettare l'introduzione di una quantità di traffico, nel sistema, maggiore, ha consentito un migliore utilizzo dell'algoritmo di scheduling dinamico introdotto decisamente migliore, rispetto alla prima simulazione. Il rapido passaggio attraverso lo stato di partizionamento completo non va visto come malfunzionento dello scheduling, bensì come stato di transizione fondamentale per consentire un'adeguata, seppur comunque contenuta, assegnazione di banda anche alle risorse di tipo BE.

Simulazione 3

Quest'ultima simulazione impone una nuova condizione al sistema: stesso traffico della simulazione 2 ma un numero di codici supportati minore. Non è stata fatta alcuna valutazione di carattere fisico circa il numero di codici considerati per la simulazione. Si voleva solamente vedere il comportamento del sistema a parità di traffico, con una risorsa a disposizione minore.

Come al solito percorrendo tutte le tappe effettuate nelle precedenti simulazioni, partiamo dalla nascita delle sorgenti.

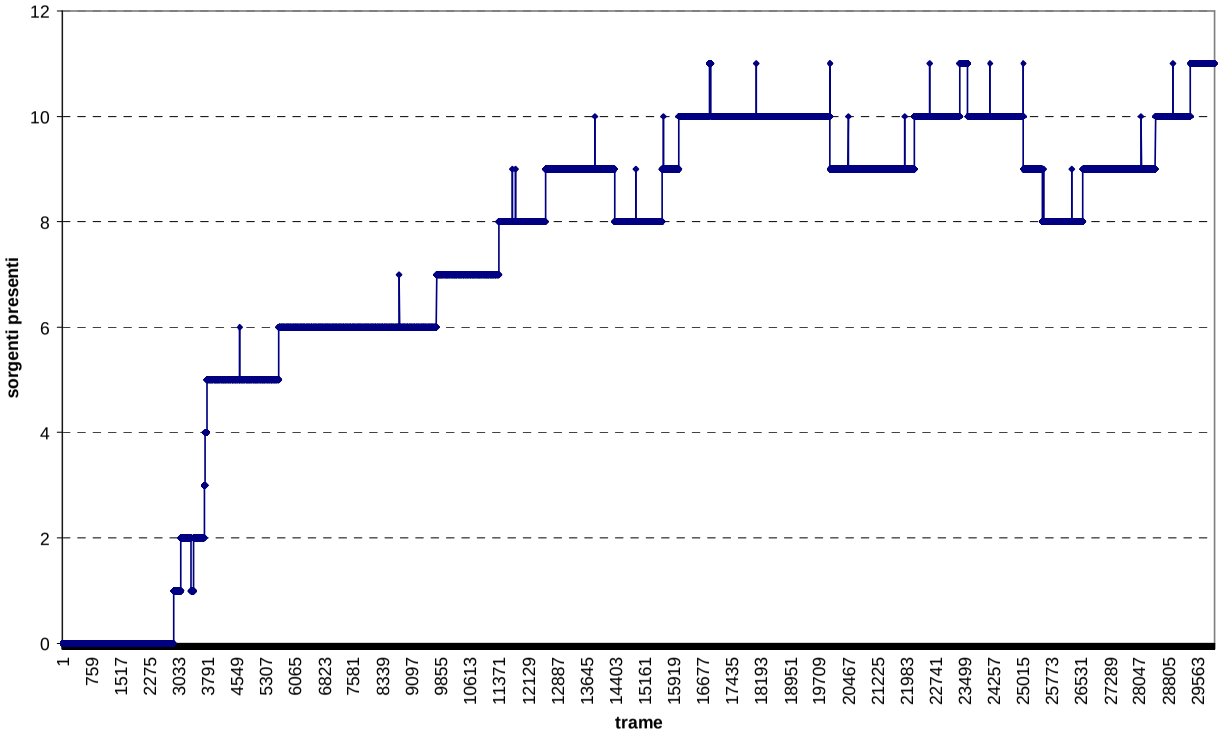


Grafico 21 – Sorgenti presenti nel sistema (caso statico)

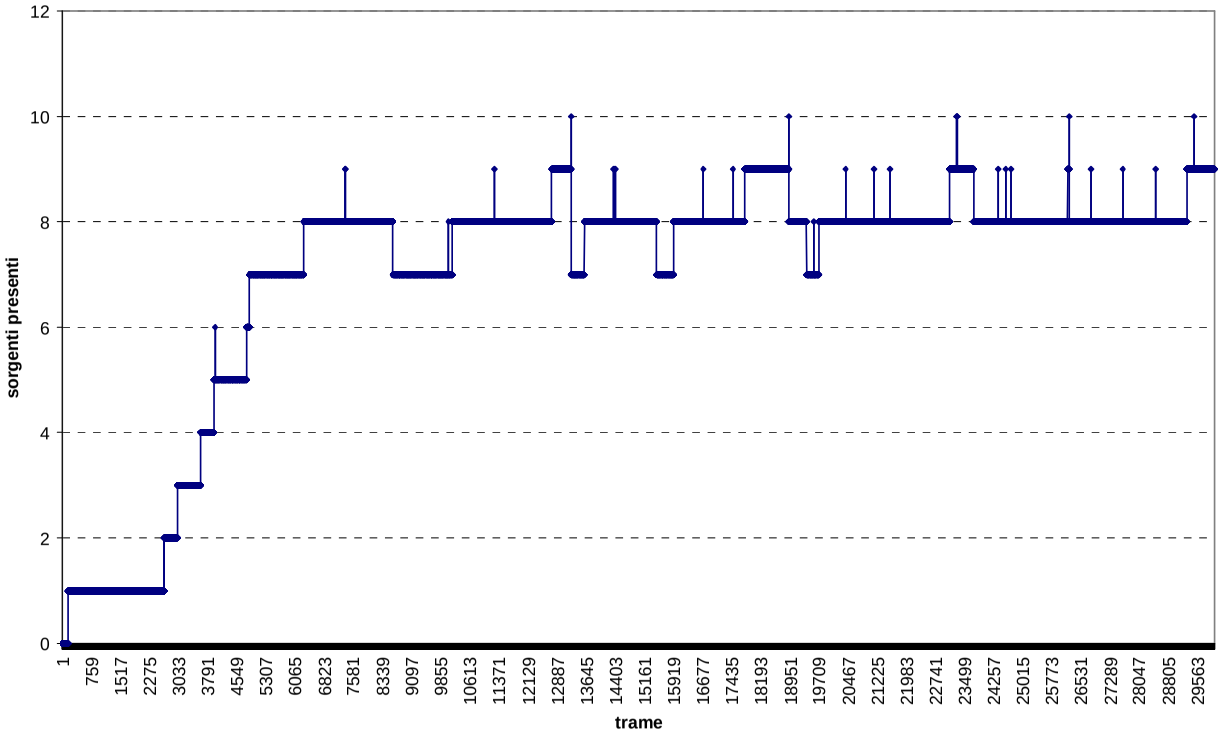


Grafico 22 – Sorgenti presenti nel sistema (caso dinamico)

Il numero di sorgenti nate nel caso statico è pari 36,ed il dettaglio delle nascite mostra 9 ON OFF, 8 CBR, 11 BE HP, 8 BE LP.

Nel caso dinamico il numero di sorgenti nate è pari a 36, anche in questo caso con un dettaglio di nascite dato da 8 ON OFF, 10 CBR, 9 BE HP, 9 BE LP. Il numero di nascite è lo stesso anche se chiaramente la distribuzione delle nascite tra le varie classi di traffico varia notevolmente. Passiamo ora alla valutazione delle prestazioni del sistema sulla base della rappresentazione delle bande assegnate.

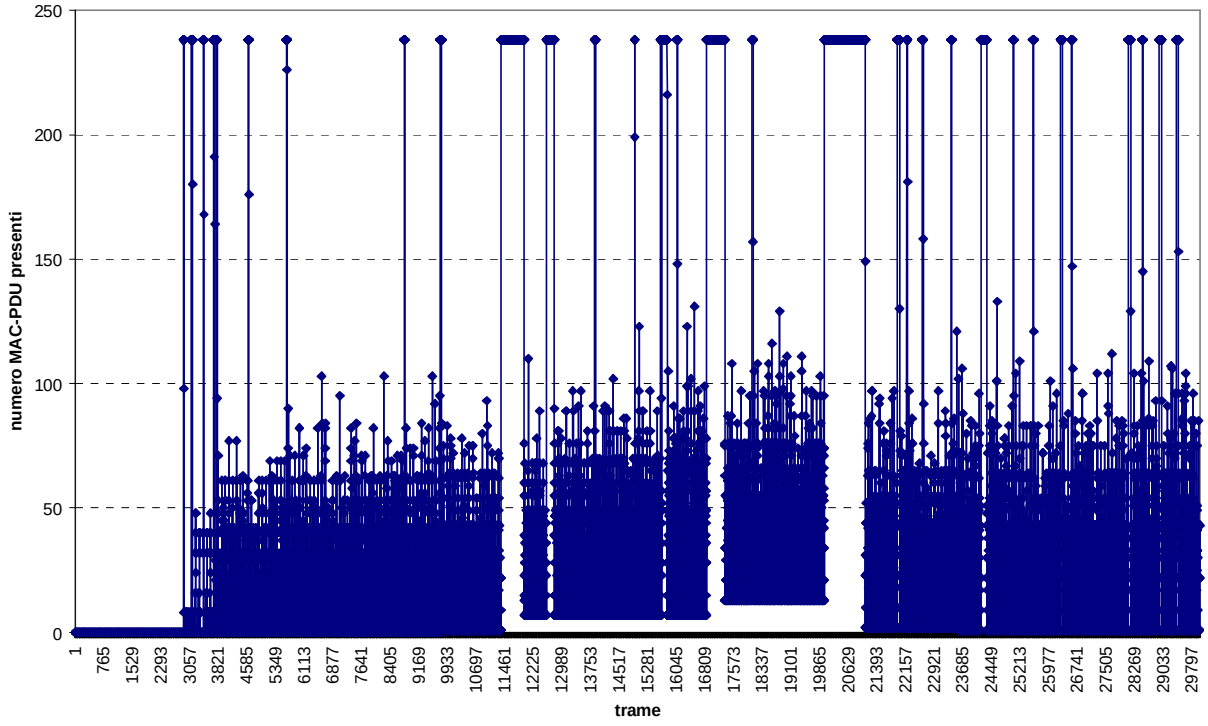


Grafico 23 – Banda assegnata totalmente (caso statico)

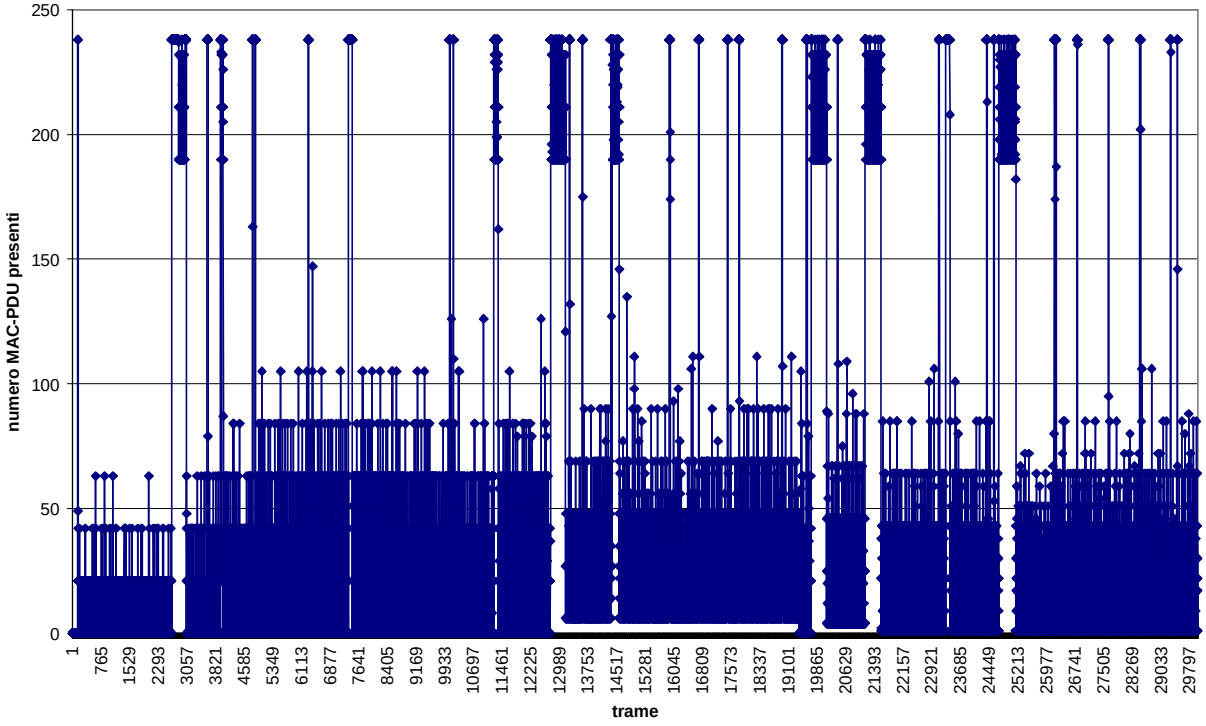


Grafico 24 – Banda assegnata totalmente (caso dinamico)

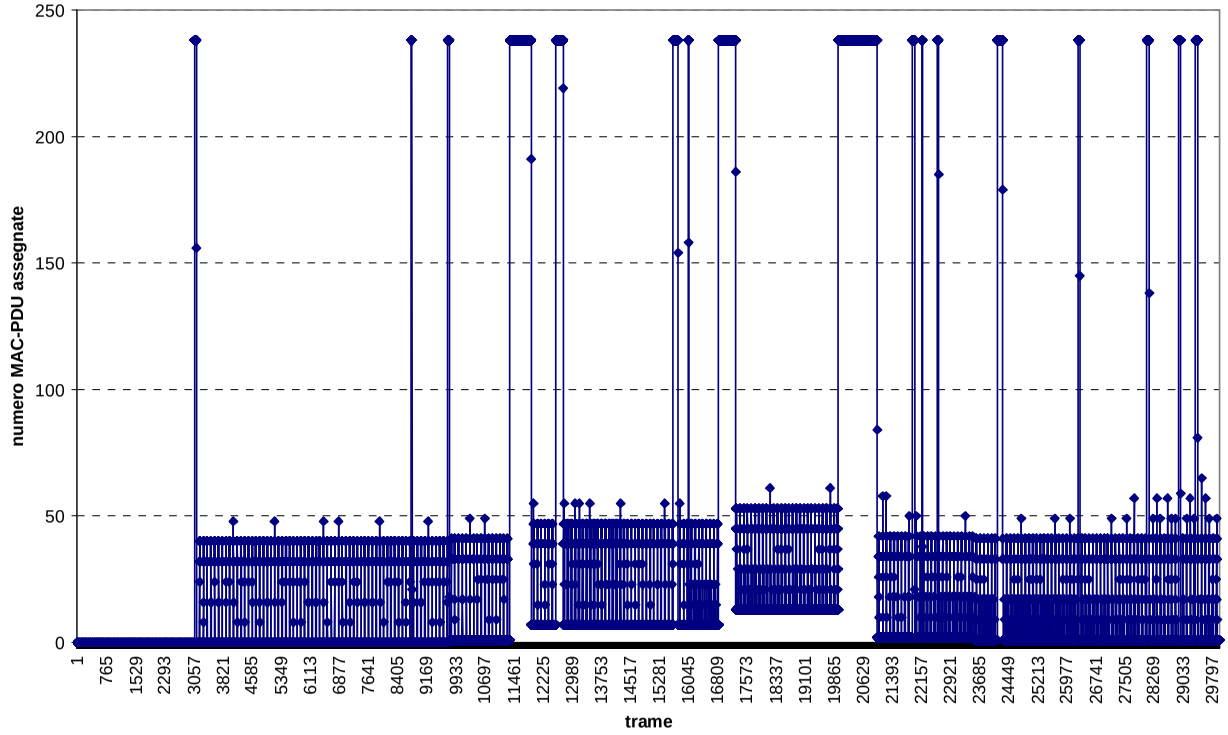


Grafico 25 – Assegnazione banda per le risorse GB (caso statico)

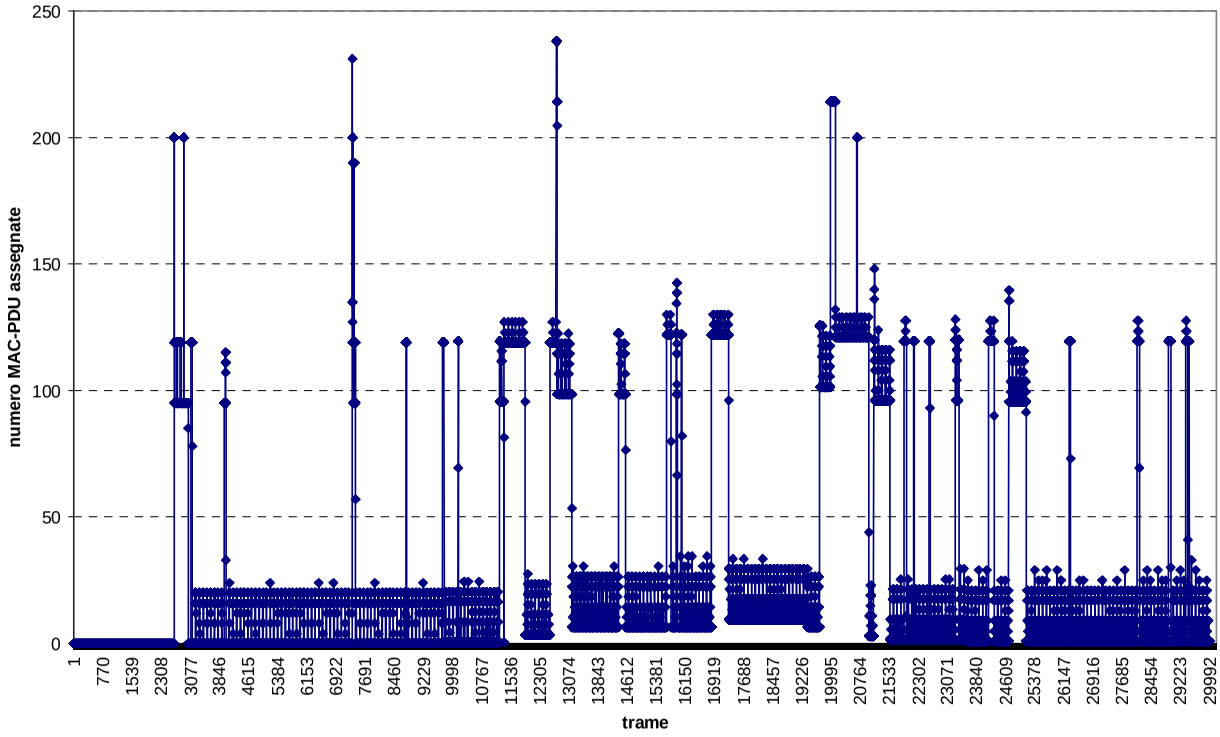


Grafico 26 – Assegnazione banda per le risorse GB (caso dinamico)

I grafici illustrati permettono di fare alcune considerazioni molto interessanti. Innanzitutto può essere facilmente fatto un confronto tra l'occupazione totale della banda in questa simulazione e l'occupazione nelle due simulazioni precedenti. L'occupazione sembra decisamente aumentata, in entrambe i casi, statico e dinamico.

Per quanto concerne in particolare l'occupazione delle risorse GB, queste nel caso statico sembrano occupare una quota parte di banda maggiore rispetto al caso dinamico. Ancora una volta l'algoritmo di scheduling adattativo implementato ha ridotto la quota parte di risorsa non pattuita assegnata in più alla classe GB. Questo ha consentito un maggiore sfruttamento della risorsa da parte delle BE. Oltre a queste considerazioni sull'occupazione della banda è necessario dire che l'aver ridotto le risorse a disposizione ha comportato il rifiuto da parte del sistema di alcune richieste effettuate dalla classe GB. Nel caso statico le richieste fatte erano 19 e di queste come facilmente riscontrabile dai primi due grafici relativi alla nascita delle sorgenti ne sono state accettate 17; nel caso dinamico invece le richieste effettuate erano 19 e quelle accolte 18. Da ciò si deduce che nel caso statico si è riscontrata una accettazione del 95% mentre nel caso dinamico una accettazione del 99%. Continuiamo con il confronto fatto anche per le simulazioni precedenti, ovvero la percentuale di banda in più assegnata alla GB rispetto a quella pattuita e la percentuale di overflow GB.

La percentuale di MAC-PDU trasmesse in eccesso rispetto a quelle garantite per trama è risultata pari a :

% Eccesso medio per trama (caso statico)=34,4%

% Eccesso medio per trama (caso dinamico)=25,3%

La percentuale di MAC-PDU presenti in coda eccedenti un limite stabilito a 5000 risulta data da:

% Overflow GB rispetto al riferimento (caso statico)=12,5%

% Overflow GB rispetto al riferimento (caso dinamico)=14,8%

La percentuale degli eccessi varia di diversi punti percentuali da caso statico a caso dinamico. Tale risultato era già visibile dal grafico dell'occupazione di banda per la

risorsa GB nel caso dinamico che risultava minore in media rispetto alla occupazione di banda nel caso statico. I valori di overflow per entrambe gli algoritmi, non sono molto differenti. Questo risultato mostra che nonostante le risorse GB siano vincolate dallo scheduling adattativo, riescono comunque ad ottenere la risorsa necessaria senza accrescere eccessivamente lo stato delle code. Passiamo ora all'algoritmo di scheduling.

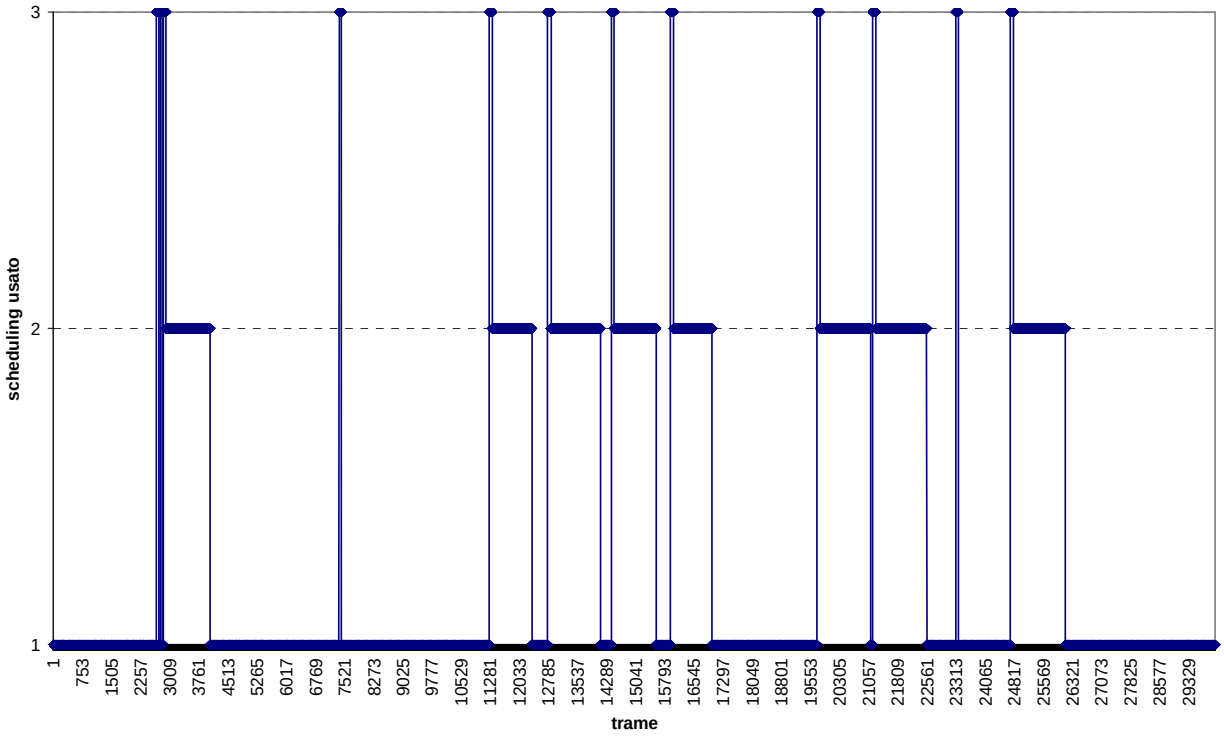
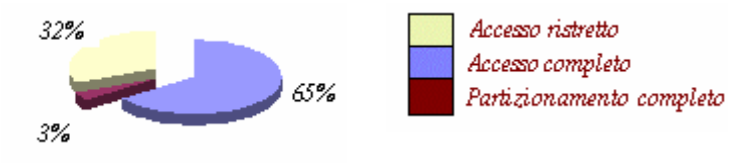


Grafico 27 – Confronto dei vari schemi di accesso: 1 rappresenta l'accesso completo, 2 l'accesso ristretto, 3 il partizionamento completo.



Grafici 28 – Percentuale di utilizzazione di ogni algoritmo di scheduling

Lo stato CP è poco utilizzato.Lo stato di accesso completo e quello di accesso ristretto sono i più utilizzati. D'altra parte non deve indurre in errore il considerare inutile lo stato CP. Questo infatti fa da tramite tra un accesso in cui è concessa tutta

la risorsa alla classe GB al caso in cui ne viene limitato l'accesso sebbene in misura non cospicua. In ogni caso l'algoritmo di scheduling mostra un contenimento maggiore degli eccessi GB rispetto all'andamento riscontrato nel caso statico.
Questo miglioramento visibile molto meglio considerando i ritardi subiti dalle MAC-PDU BE ad alta priorità ed a bassa priorità.

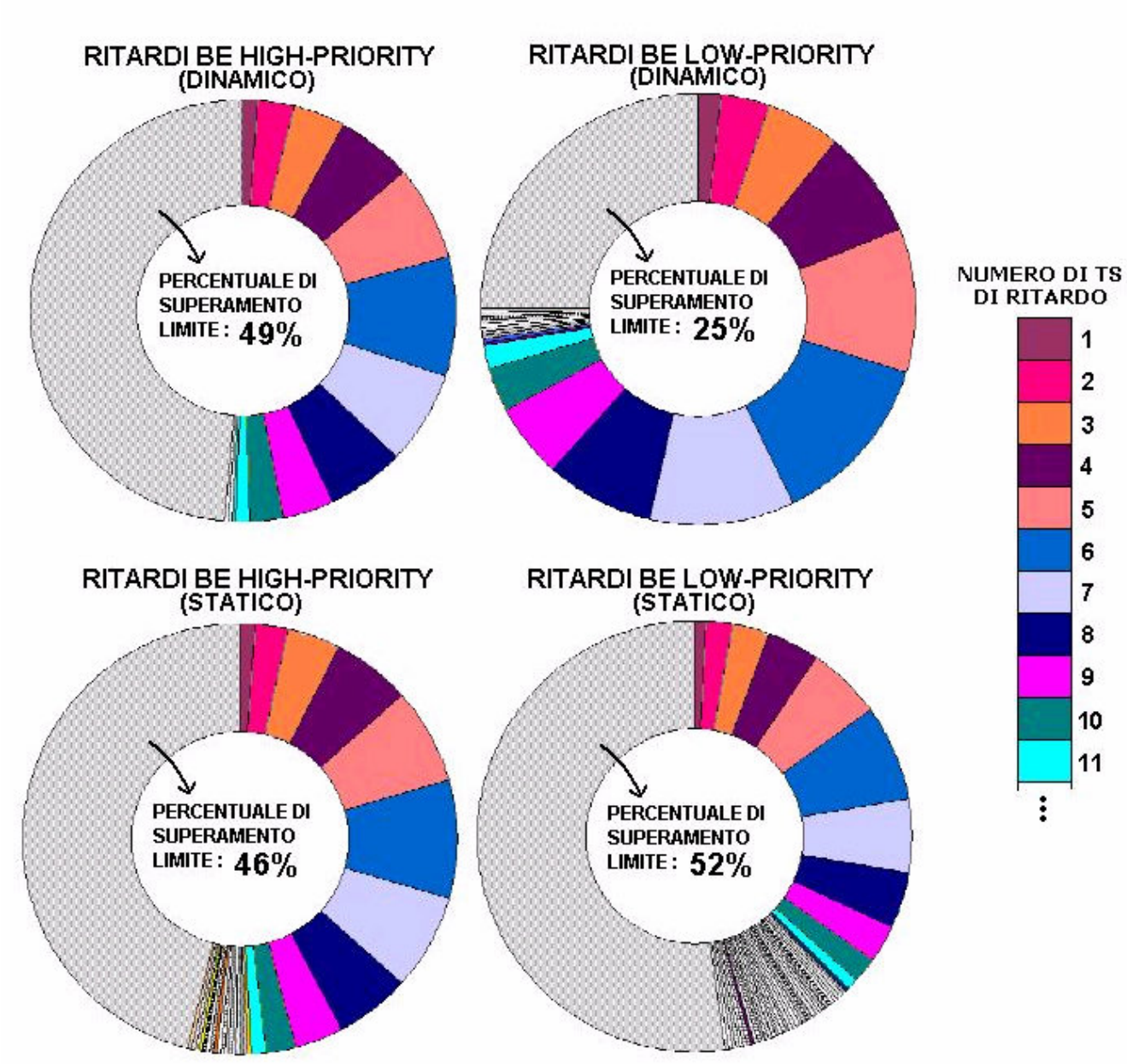


Grafico 29 – Ritardo BE Statico vs Ritardo BE Dinamico

Valutando i grafici riportati di seguito, facciamo la solita analisi orizzontale e verticale. L'analisi orizzontale dell'algoritmo dinamico mostra un grado di ritardo subito dalle MAC-PDU di classe BE molto più grande del ritardo subito dalle MAC-PDU BE a bassa priorità.

La cosa sembra un po' strana ma è spiegabile se si pensa che il passaggio da un algoritmo di accesso completo ad uno di partizionamento completo e di accesso ristretto avviene qualora le GB stiano ottenendo troppo assegnazione in più. Probabilmente nel momento di passaggio di stato le risorse BE più presenti erano quelle a bassa priorità che di conseguenza nonostante la loro posizione nella classe gerarchica sono state trattate in maniera migliore rispetto alle GB. Nonostante questo andamento si può comunque dire che in media il ritardo delle BE nel caso dinamico è discreto.

Per quanto invece riguarda le simulazioni in ambiente statico., queste presentano valori tra di loro confrontabili e comunque ritardi per le code a più alta priorità maggiori. L'analisi verticale ci fa constatare un comportamento circa uguale nel caso delle code ad alta priorità ma un miglioramento decisamente buono nel caso delle code BE LP.

Ad avallare i risultati presentati con i ritardi medi i valori di overflow BE ad alta priorità ed a bassa priorità per le simulazioni in ambiente statico ed in ambiente dinamico, i valori degli overflow nel caso delle BE ad alta priorità ed a bassa priorità.

La percentuale di MAC-PDU BE presenti in coda eccedenti il limite stabilito a 5000 risulta:

% Overflow BE HP rispetto al riferimento (caso statico)=25,4%

% Overflow BE HP rispetto al riferimento (caso dinamico)=26,19%

% Overflow BE HP rispetto al riferimento (caso statico)=49,98%

% Overflow BE LP rispetto al riferimento (caso dinamico)=37,65%

I valori trovati mostrano un miglioramento negli overflow di classe BE, superiore per le code a bassa priorità rispetto a quelle ad alta priorità. Queste ultime, d'altra parte ricevono maggiore assegnazione di banda in pieno rispetto delle priorità

attribuite alle varie classi. Nel complesso si nota che l'algoritmo di scheduling adattativo introdotto, permette una gestione più intelligente ma soprattutto più equa della quantità di traffico sia GB che BE presente nel sistema.

CONCLUSIONI

In questa tesi è stato descritto un sistema di tipo WLAN, basato sulla tecnica di modulazione OFDM-CDMA.. La configurazione per tale sistema è di tipo punto multipunto: una stazione radio base al centro di una cella, presa come scenario di riferimento, dialoga e si scambia dati con radio terminals (terminali mobili e fissi), che sono a lei connessi attraverso una configurazione a stella. Nell'ambito di questo contesto, la tesi ha affrontato diversi problemi. Dapprima è stato affrontato lo studio di un protocollo di accesso al mezzo per la tratta di downlink che fosse in grado da un lato di supportare differenti classi di servizio e, dall'altro, di garantire un assegnazione della risorsa intelligente. È stato realizzato, in concomitanza, un simulatore che nell'ambito della tesi ho denominato *statico* in grado di consentire l'accesso al mezzo a una certa quantità di utenti, e capace di assegnare le risorse disponibili basandosi su un preciso algoritmo di scheduling. Tale algoritmo è stato pensato per supportare le due classi di servizio GB e BE in maniera differente. Alla classe GB, oltre alla quantità di risorse pattuita, viene concessa ulteriore quantità di risorsa in proporzione al proprio stato delle code. La percentuale di banda rimanente viene chiaramente lasciata disponibile per la classe di traffico di tipo BE. Chiaramente una modellizzazione di questo genere, ha concesso troppe libertà alle risorse a qualità garantita lasciando poca banda alle risorse di tipo best effort. Per superare questa limitazione ho proposto un algoritmo di scheduling adattativo in grado di comportarsi in maniera più equa sia nei confronti delle GB che soprattutto nei confronti delle "più sofferenti" BE. Oltre a questo ho previsto anche un meccanismo di ingresso di sorgenti nel sistema di tipo realistico: le sorgenti nascono in un istante di tempo casuale ed, in relazione alla dimensione media di traffico emesso, muoiono. Questa caratteristica non presente nella prima versione del simulatore, ha permesso di valutare la risposta dell'algoritmo di scheduling adattativo in modo più conforme alla realtà.

Il confronto dei due algoritmi ha evidenziato un miglioramento nell'assegnazione delle risorse e nei ritardi medi relativi alle MAC-PDU BE nel caso di utilizzo di algoritmo adattativo.

Tra gli sviluppi futuri vanno considerati sicuramente l'eliminazione della doppia garanzia alle risorse a qualità garantita ed una gestione dei buffer in grado di variare la propria dimensione in condizioni di elevato carico nella rete, garantendo così minore spreco di risorsa. Il protocollo di scheduling proposto dovrebbe poi essere fuso con adeguati meccanismi di controllo d'errore e ritrasmissione per poter provare l'effettiva efficacia anche in condizioni di canale non ideale.