**Massimiliano Sbaraglia**
**Network Engineer**

# MPLS DiffServ aware Traffic Engineering

# I modelli per la QoS

In ambito IETF sono stati definiti due modelli per la QdS su reti IP:

Il modello “integrated service” separa il traffico IP nelle sue componenti end-to-end, applicando ad ogni singolo flusso politiche di QoS.
IntServ si basa sul protocollo RSVP per la segnalazione e l’allocazione di risorse relative ad ogni singolo flusso nella rete; il suo limite principale è la scalabilità.

Il modello “differentiated service” tratta il traffico IP come un insieme di aggregati di flussi chiamati behavior aggregate (BA); questo permette di applicare QoS ad insiemi di flussi anziché alla singola componente end-to-end senza ricorrrere ad un protocollo di segnalazione.

Il meccanismo DiffServ può essere riassunto nel modo seguente:

Ai bordi della rete i singoli pacchetti vengono classificati dagli *Edge Router* che marcano il campo DSCP presente nella loro intestazione in base ai requisisti prestazionali richiesti;

Ogni valore del campo DSCP corrisponde ad una classe di servizio e tutti i pacchetti aventi lo stesso campo DSCP riceveranno lo stesso trattamento all'interno della rete;

I pacchetti, una volta classificati, vengono immessi nella rete;

All'interno della rete, in ogni router vengono definiti i **PHB** (Per Hop Behaviour) ovvero i comportamenti corrispondenti alle classi di servizio con le quali sono classificati i pacchetti;

Quando un pacchetto arriva in un *Core Router*, quest'ultimo esamina il campo DSCP e tratterà il pacchetto in base alla classe di servizio corrispondente.

Per Hop behavior (PHB):
è il modo e la tecnica, in termini di politiche di QoS, attraverso il quale un router instrada un particolare aggregato di flussi; è definito dai parametri dell'algoritmo di scheduling, dalla quota di buffer e dalla quota di banda che di fatto consentono di ottenere un diverso servizio.

Il modello DiffServ prevede il riutilizzo dei primi sei bit del campo ToS dell'header IP chiamato DS field.

IETF definisce quattro PHB standard

- Default
- Class-Selector
- Expedited Forwarding (EF)
- Assured Forwarding (AF)

Class-Selector:

I valori DSCP nella forma xxx000 sono chiamati Class-Selector e sono utilizzati per il mantenimento della compatibilità con la precedente tecnica IP-Precedence.

Default PHB:

Un pacchetto marcato con DSCP 000000 è trattato in modalità best effort

Expedited Forwarding (EF)

Grazie all'utilizzo di code a priorità all'interno degli apparati DS compatibili, i pacchetti marcati con DSCP **101110** avranno diritto di precedenza.

Permette di offrire un servizio di tipo premium

Assured Forwarding (AFxy)

Ad aggregati diversi vengono assicurati diverse garanzie di forwarding

Il traffico può essere suddiviso in x class a ciascuna della quali è possibile associare y livelli di drop precedence.

Il Dominio DS è un porzione di rete amministrata da un unico gestore divisa in:

EDGE formata dai router esterni.
CORE formata dai router interni.

***Edge Router:*** sono i router ai bordi di un dominio DS.

Essi svolgono le funzioni fondamentali che prevede l'architettura DiffServ quali la classificazione dei pacchetti, la marcatura del campo DSCP e la policing-shaping del traffico.

***Core Router:*** sono router ad altissima capacità all'interno del dominio DS.

Essi hanno il solo compito di eseguire i PHB sui datagrammi già classificati e marcati dagli edge-router

Conclusioni:

La metodologia DiffServ rappresenta un primo reale passo in avanti per l'implementazione e la gestione della QoS nelle reti IP, ma è anche da considerarsi incompleta in quanto non offre le garanzie in termini di allocazione delle risorse trasmissive end-to-end al percorso effettuato dai dati dell'applicazione di nostro interesse.

Con il termine Ingegneria del Traffico (Traffic Engineering–TE) si indica un insieme di funzioni che hanno lo scopo di ottimizzare le prestazioni di una rete (RFC 2702)

Gli obiettivi prestazionali associati alle funzioni di TE riguardano

- Il TRAFFICO trasportato da una rete
- Le RISORSE di cui dispone una rete

Gli obiettivi che riguardano il TRAFFICO, si riferiscono al controllo e al soddisfacimento dei requisiti di QoS dei flussi di traffico che attraversano la rete

- Minimizzazione della perdita di pacchetti
- Massimizzazione del throughput
- Minimizzazione del ritardo

Gli obiettivi che riguardano le RISORSE, si riferiscono alla distribuzione del carico e all' ottimizzazione dell'uso delle risorse di rete (es. banda dei link)

- Minimizzazione della congestione
- Efficiente uso della banda

# Limitazione dei protocolli IGP



I protocolli di routing IGP (ISIS, OSPF) non sono adeguati alle esigenze di una politica di TE

I protocolli IGP :

- utilizzano algoritmi Shortest Path First (SPF)
- utilizzano metriche additive
- sono topology-oriented

I protocolli IGP non considerano:

- la disponibilità di banda sui vari link (stato di congestione dei link)
- le caratteristiche del traffico

# Fattori chiave: Constraint-based Routing

***Constraint-based Routing*** (CR)
L'idea alla base del TE sta nel fatto che per avere un routing efficiente bisogna instradare il traffico lungo percorsi a costo minimo rispettando contemporaneamente determinati vincoli (*constraints*) sullo sfruttamento delle risorse di rete.
Per poter effettuare il CR sono necessarie funzionalità aggiuntive nella rete, tra le quali la presenza di protocolli di routing opportunamente estesi (come ad esempio OSPF-TE o ISIS-TE) per portare le informazioni relative al TE.

Uno volta determinato il percorso ottimo serve poi un protocollo di segnalazione tramite il quale informare tutti gli LSR interessati della necessità di allocare un tunnel TE con determinati requisiti di banda e/o amministrativi e quindi di associare delle etichette MPLS per definire il/gli LSP corrispondenti.

In ambito IETF sono stati standardizzati due protocolli di segnalazione:

RSVP-TE (*ReSerVation Protocol with Tunneling Extension*).
CR-LDP (*Constraint-based LSP setup using LDP*)

Entrambi sono estensioni dei protocolli RSVP e LDP

Il protocollo MPLS è in grado di fornire funzionalità di TE analoghe a quelle fornite dal modello overlay:

Gestione degli explicit routing LSP (tecnica quasi a circuito)

Possibilità di mantenere e proteggere gli LSP

Mappaggio dei flussi di traffico sugli LSP

Caratterizzazione dei flussi e delle risorse di rete mediante attributi

Facilità di integrazione con meccanismi di Constraint-Based Routing (CBR)

Bassa complessità realizzativa

Si definisce ***Traffic Trunk*** un aggregato di flussi di traffico caratterizzati dall'appartenenza alla medesima classe di servizio ed alla stessa FEC.

Un traffic Trunk viene instradato su un collegamento virtuale detto tunnel TE, tipicamente costituito da un singolo LSP

Un Traffic Trunk (TT) è caratterizzato da:

- gli LSR di ingresso e di uscita
- la FEC in cui è mappato
- gli attributi che ne caratterizzano il comportamento

Gli attributi di base dei traffic trunk sono relativi a:

- Parametri di traffico
- Policing
- Selezione e gestione dei cammini
- Priorità
- Preemption
- Resilienza

Gli attributi dei trunk e delle risorse e i parametri associati alla tecnica di routing rappresentano le variabili di controllo che possono essere modificate dal gestore per agire sullo stato della rete

La funzionalità previste in MPLS per il supporto delle funzionalità TE sono

- La definizione di un insieme di attributi associati ai traffic trunk che ne caratterizzano il comportamento
- La definizione di un insieme di attributi associati alle risorse di rete che ne vincolano l'utilizzazione
- Un instradamento vincolato (Constraint-based routing) usato per la scelta del cammino per i traffic trunk soggetto ai vincoli fissati dagli attributi delle risorse

# Attributi dei Trunk



I parametri di traffico

sono usati per riassumere le caratteristiche statistiche del flusso di traffico trasportato dal TT (profilo di traffico)

Peak rate
Average rate
Burst size

Servono a valutare le risorse di rete (es. banda) necessarie al trasporto del TT

Funzione di allocazione delle risorse

Gli attributi di policing

stabiliscono il trattamento da riservare alla quota parte di traffico non conforme al profilo permesso

Scarto
Marcatura

# Attributi dei Trunk – selezione e gestione cammini



Definiscono le regole per la scelta dei cammini per il supporto dei TT e per le alternative per la loro gestione

I cammini possono essere

- Calcolati automaticamente attraverso i protocolli di routing (es. OSPF)
- Determinati off-line dall'operatore di rete
  - Cammini completamente specificati, intero insieme degli LSR
  - Cammini parzialmente specificati, insieme parziale degli LSR

La gestione dei cammini riguarda

- l'adattività ai cambiamenti di stato della rete
  - Re-ottimizzazione permessa
  - Re-ottimizzazione non permessa
- le regole di distribuzione del traffico su cammini alternativi
  - Definizione delle percentuali di traffico da inviare sui cammini alternativi

# Attributi dei Trunk – priorità e preemption



La priorità definisce l'ordine con il quale avviene la selezione del cammino al momento dell'instaurazione dei TT

L'attributo di preemption (prelazione) determina se un TT può sottrarre banda ad un altro TT già attivo su un particolare cammino

La preemption può essere usata per assicurare che TT ad alta priorità vengano sempre instradati sui cammini più favorevoli

La preemption può essere usata per realizzare varie politiche di riconfigurazione dei cammini in caso di guasto

Ciascun TT ha due tipi di priorità associate:

Set-up priority

è il livello di priorità che caratterizza un TT nel prendere le risorse all'atto della sua instaurazione

Holding priority

è il livello di priorità che caratterizza un TT nel mantenere le risorse successivamente alla sua instaurazione

Entrambe le priorità possono assumere valori nell'intervallo [0-7]; il valore zero è la priorità più alta

All'atto dell'instaurazione, la banda richiesta da un TT può essere allocata su un link se:

La banda è disponibile (libera)

Mediante l'abbattimento di TT già instaurati che hanno un valore della holding priority inferiore alla set-up priority del TT da instaurare

# Attributi dei Trunk - resilenza



Determina il comportamento del TT in presenza di guasti

Sono possibili diverse politiche

- Nessun reinstradamento
- Reinstradamento solo su cammini con risorse sufficienti
- Reinstradamento su qualsiasi cammino con qualsiasi livello di risorse
- Qualsiasi combinazione delle precedenti

Conclusioni:

Il TE permette di bilanciare il traffico in una rete in modo da non avere link congestionati né scarsamente utilizzati.
Inoltre il TE consente l'allocazione delle risorse trasmissive per aggregati di flussi, rendendo possibile lo sviluppo di meccanismi di QoS end-to-end.

I meccanismi di Fast Reroute, fornendo un efficace meccanismo di protezione locale del tunnel TE, aumentano la robustezza e la resilienza della rete

Il TE nella sua versione base (non DiffServ aware) non è in grado di creare un LSP allocando risorse trasmissive sulla base della classe di servizio (CoS aware).

Per supportare la QoS le infrastrutture di rete devono soddisfare due vincoli fondamentali:

L'allocazione delle risorse di rete deve essere garantita lungo tutto il percorso effettuato dal flusso di traffico sia in condizioni normali che in condizioni di congestione o malfunzionamento degli apparati di rete.

Ad ogni pacchetto di un flusso di traffico, in ogni nodo della rete, deve essere applicato un trattamento che permetta il soddisfacimento dei parametri di QoS.

L 'architettura MPLS grazie all'estensione TE dei protocolli di routing IGP può offrire un ambiente connection-oriented idoneo per l'implementazione della QoS, garantendo le opportune e richieste risorse trasmissive agli aggregati di traffico.

L'architettura DiffServ, invece, permette la classificazione dei pacchetti in BA e il loro trattamento diversificato rappresentato dai PHB nodo per nodo.

L'integrazione delle due architetture porta allo sviluppo di una meccanismo per gestire in modo completo la QoS. Si parla in questo caso di MPLS DiffServ-aware Traffic Engineering.

Il Traffic Engineering nella sua versione evoluta (DiffServ aware) è grado di creare un LSP allocando risorse trasmissive sulla base della classe di servizio con il quale in traffico può essere classificato.

Grazie al DiffServ TE siamo oggi in grado fornire tutti i meccanismi a disposizione del Traffic Engineering classico (resource reservation e fault-tolerance properties) con un granuralità spinta al livello CoS

Affinché MPLS e quindi il TE possa supportare DiffServ, è necessario che gli LSR riescano a distinguere i vari pacchetti in base al loro DSCP per inoltrarli secondo il PHB corrispondente.

Gli LSR si basano esclusivamente sulla label dello shim header MPLS per il forwarding dei pacchetti, e non esaminano l'header IP. Le soluzioni che sono state proposte sono:

**E-LSP (Experimental bit inferred LSP):**

si copia il DSCP nel campo EXP dell'intestazione MPLS.
Conveniente per reti che offrono poche classi di servizio (max. 8)
Un solo LSP per tutte le classi di servizio

**L-LSP (Label inferred LSP):**

Un LSP per ciascuna classe di servizio
Conveniente per classi di servizio con diversi livelli di "drop precedence"
L'inoltro sulle relative code avviene attraverso l'etichetta MPLS di livello più elevato

Prima di introdurre i modelli utilizzati per allocare le risorse di banda sui link di una rete MPLS fissiamo alcuni concetti chiave:

Class Type (CT)

E' un insieme di Traffic Trunk (TT) governati dai medesimi BW constraints (BC)

Un Diffserv-TE-LSP può trasportare il traffico di un solo CT (regular LSP)

Un Diffserv-TE-LSP può trasportare il traffico di differenti CT (Multiclass LSP)

I percorsi seguiti dal traffico appartenete a ciascun CT definito all'interno del medesimo multiclass Diffserv-TE-LSP possono essere gli stessi o differenti.

Sono supportate al massimo 8 CT (CT0 ÷ CT7)

La mappatura tra PHB Diffserv e CT è definita dall'amministratore di rete in base alle particolari esigenze di servizio.

Ad ogni CT può essere associato ad un differente livello di priorità (0÷7). Questa occorrenza genera 64 combinazioni differenti. Il TE-Class è definito come la combinazione di (CT, priority)

Bandwidth Constraint (BC)

Rappresenta il limite percentuale di risorse di banda di un link che una CT (oppure un gruppo di CT) può utilizzare.

La configurazione di un regular DiffServ-aware TE LSP prevede l'attribuzione del CT e l'assegnazione della corrispondente BC.

La configurazione di un multiclass DiffServ-aware TE LSP prevede l'attribuzione di un insieme di CT e l'assegnazione dei corrispondente BC.

E' possibile la convivenza sulla stessa rete di regular e multiclass DiffServ-aware TE LSP

Maximum Reservable Bandwidth (MaxRBW)

La banda complessiva massima che è possibile riservare su un link di capacità C (bit/s)

# Modelli di allocazione delle risorse

Per default un regular o multiclass DiffServ-aware TE LSP è segnalato con una "setup priority" pari a 7 ed una "holding priority" pari a 0. Questo significa che il tunnel all'atto dell'instaurazione non può prendere le risorse di rete di un tunnel pre-esistente e che successivamente alla sua instaurazione nessun nuovo tunnel può prendere le sue.

E' possibile la convivenza sulla stessa rete di TE LSP e DiffServ-aware TE LSP.

Maximum Allocation Model (MAM)

RFC 4125

Russian Dolls Model (RDM)

RFC 4127

# Maximum Allocation Model (MAM)

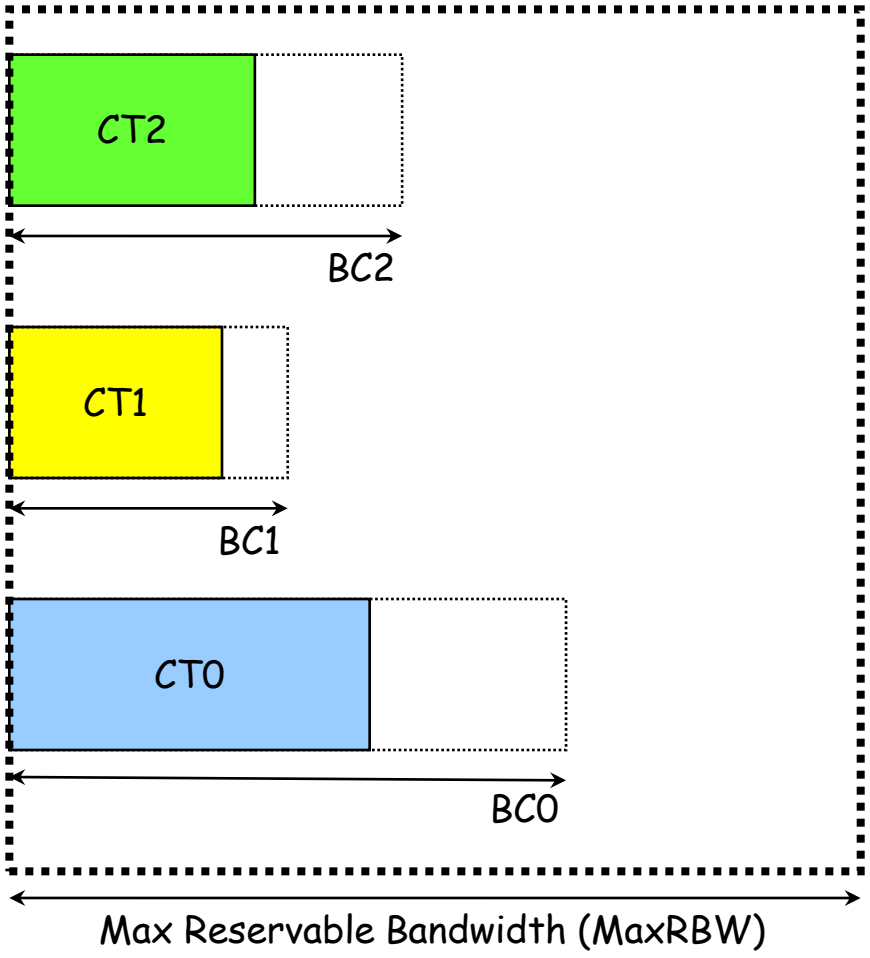


La banda complessiva (MaxRBW) è rigidamente divisa fra i CT

Ogni CT ha il proprio vincolo di banda massima (BC).

La somma delle bande riservate a tutte le CT deve essere minore o uguale della MaxRBW.

CT isolation: ogni CT ha la sua porzione di banda assegnata BC condizione che non consente condivisione tra le CT e quindi l'utilizzo di eventuale banda non utilizzata.

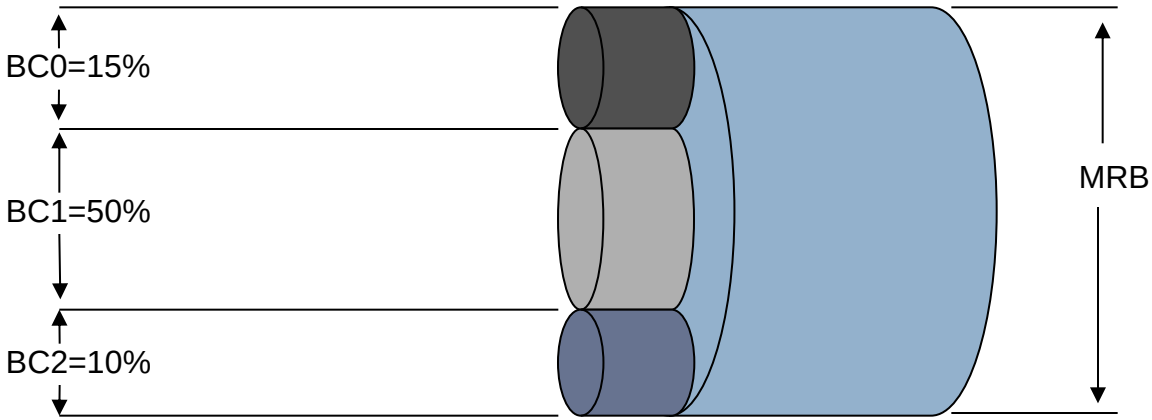


# Maximum Allocation Model (MAM)

massimiliano sbaraglia

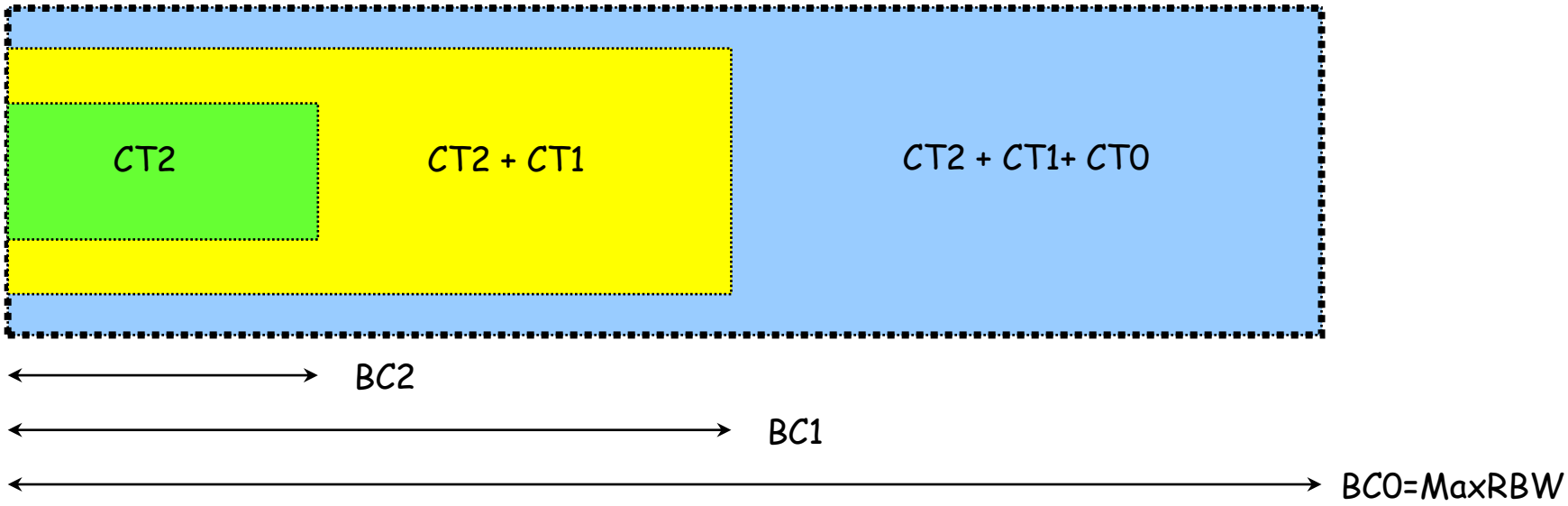
| Bandwidth Constraint | Maximum Bandwidth Allocation For |
|---|---|
| BC7 | CT7 |
| BC6 | CT6 |
| BC5 | CT5 |
| BC4 | CT4 |
| BC3 | CT3 |
| BC2 | CT2 |
| BC1 | CT1 |
| BC0 | CT0 |

Figureshows an example of a set of BCs using MAM. This DS-TE configuration uses three CTs with their corresponding BCs. In this case, BC0 limits CT0 bandwidth to 15 percent of the maximum reservable bandwidth. BC1 limits CT1 to 50 percent, and BC2 limits CT2 to 10 percent. The sum of BCs on this link is less than its maximum reservable bandwidth. Each CT will always receive its bandwidth share without the need for preemption. Preemption will not have an effect on the bandwidth that a CT can use. This predictability comes at the cost of no bandwidth sharing between CTs. The lack of bandwidth sharing can force some TE LSPs to follow longer paths than necessary.



# Russian Dolls Model (RDM)






Ogni CT ha a disposizione una banda uguale alla somma di due componenti

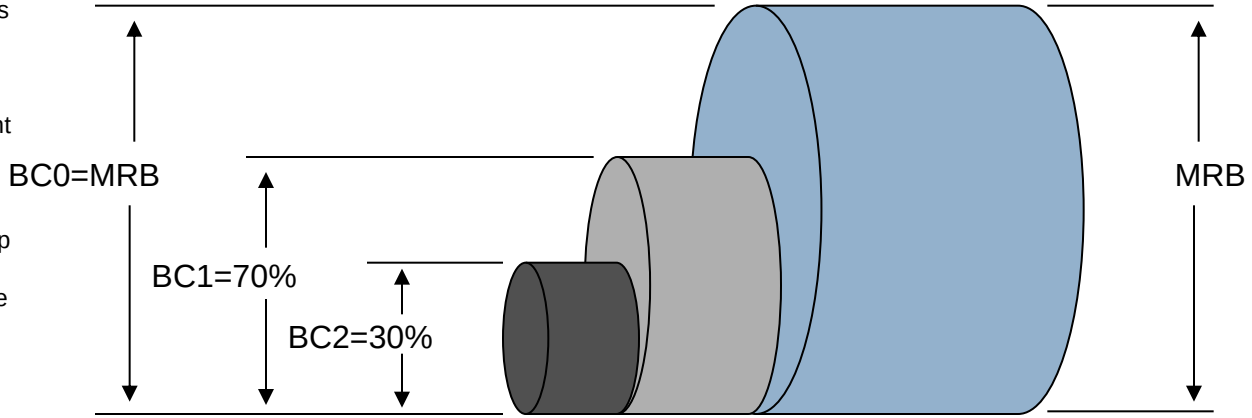
La somma delle bande a disposizione delle classi inferiori

Una propria componente

# Russian Dolls Model (RDM)

| Bandwidth Constraint | Maximum Bandwidth Allocation For |
|---|---|
| BC7 | CT7 |
| BC6 | CT7+CT6 |
| BC5 | CT7+CT6+CT5 |
| BC4 | CT7+CT6+CT5+CT4 |
| BC3 | CT7+CT6+CT5+CT4+CT3 |
| BC2 | CT7+CT6+CT5+CT4+CT3+CT2 |
| BC1 | CT7+CT6+CT5+CT4+CT3+CT2+CT1 |
| BC0 | CT7+CT6+CT5+CT4+CT3+CT2+CT1+CT0 |

Figure shows an example of a set of BCs using RDM. This DS-TE implementation uses three CTs with their corresponding BCs. In this case, BC2 limits CT2 to 30 percent of the maximum reservable bandwidth. BC1 limits CT2+CT1 to 70 percent. BC0 limits CT2+CT1+CT0 to 100 percent of the maximum reservable bandwidth, as is always the case with RDM. CT0 can use up to 100 percent of the bandwidth in the absence of CT2 and CT1 TE LSPs. Similarly, CT1 can use up to 70 percent of the bandwidth in the absence of TE LSPs of the other two CTs. CT2 will always be limited to 30 percent when no CT0 or CT1 TE LSPs exist. The maximum bandwidth that a CT receives on a particular link depends on the previously signaled TE LSPs, their CTs, and the preemption priorities of all TE LSPs



# MAM vs RDM

Il metodo MAM assicura:

Estrema semplicità
Scarsa efficienza nell'uso della banda
Ogni CT ha la sua porzione di banda assegnata (CT isolation), condizione che non consente condivisione di banda tra le CT (scarsa efficienza).

Il metodo RDM assicura:

Elevata complessità
Elevata efficienza nell'uso della banda
Un basso grado di isolamento tra le classi
Protezione contro la degradazione della QoS in tutte le classi
La CT isolation può essere raggiunta se si utilizza la preemption tra le classi