Capítol 3. Xarxes troncals

3.1 Commutació de trames
3.2 Commutació de cel·les
3.3 Commutació d'etiquetes
3.4 Carrier Ethernet
3.5 Control de la congestió

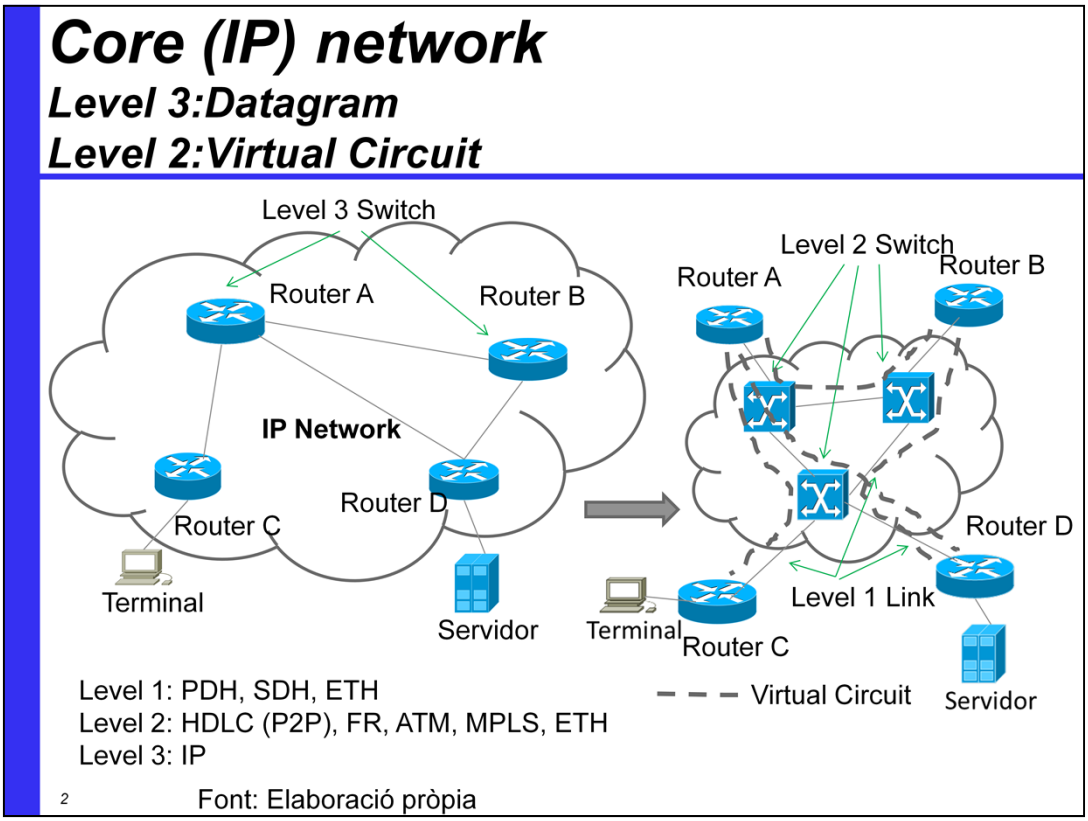
1

Tecnologies per a les xarxes troncals (Core networks).

Tot són tecnologies de nivell 2 (o 2,5) i el nom identifica el que commuta el commutador (trames, cel·les, etiquetes, trames ethernet).

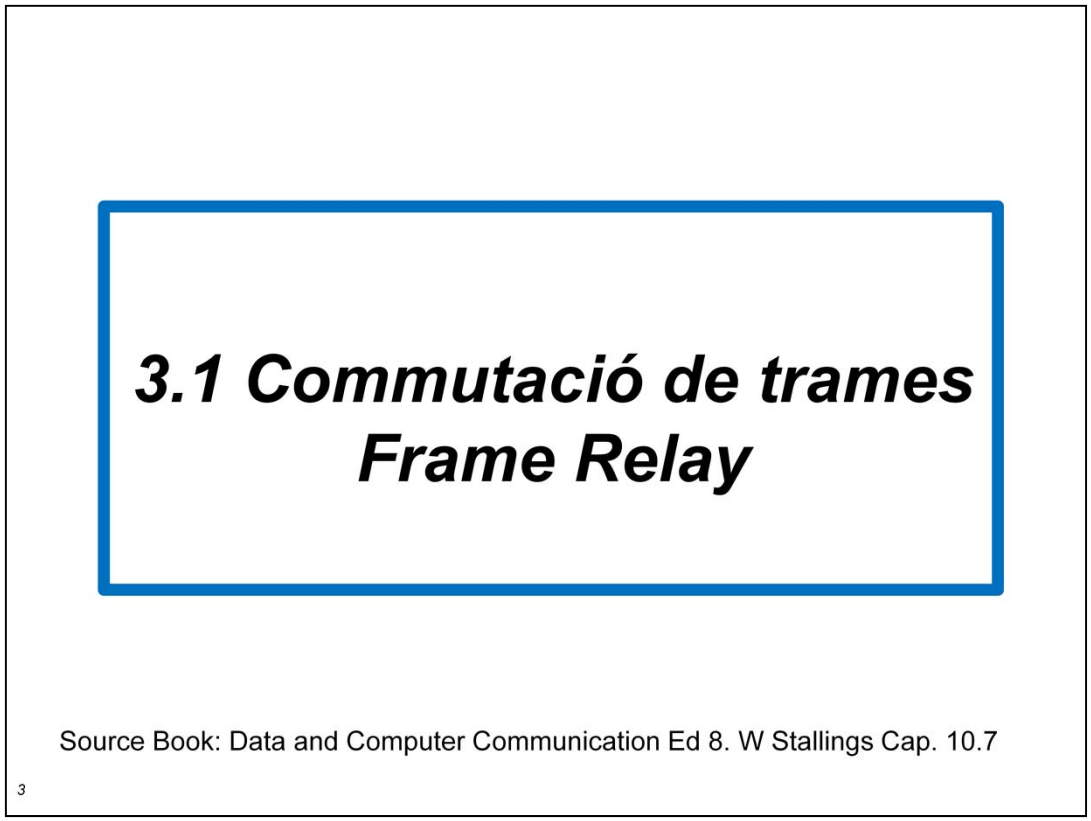
Al final veurem el control de la congestió aplicable a totes les tecnologies.

T



Les taules d'enrutament de la xarxa IP diuen el camí del terminal al servidor passant pels routers C, A i D. Caldrà seguir els circuits virtuals de la xarxa Level 2

T



Sistema de commutació a nivell 2. És una modificació de l'HDLC simplificada sense camp de control.

Frame Relay

- Developed to take advantage of high data rates and low error rates
- Operates at data rates of up to 2 Mbps
- Key to achieving high data rates is to strip out most of the overhead involved with error control

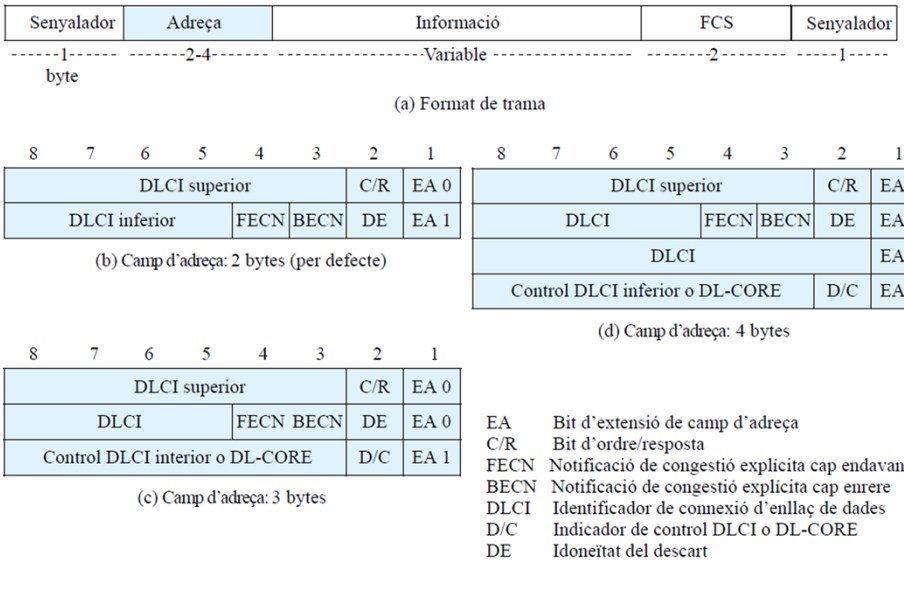
4

Packet switching was developed at a time when digital long-distance transmission facilities exhibited a relatively high error rate compared to today's facilities. As a result, there is a considerable amount of overhead built into packet-switching schemes to compensate for errors. The overhead includes additional bits added to each packet to introduce redundancy and additional processing at the end stations and the intermediate switching nodes to detect and recover from errors. With modern high-speed telecommunications systems, this overhead is unnecessary and counterproductive. It is unnecessary because the rate of errors has been dramatically lowered and any remaining errors can easily be caught in the end systems by logic that operates above the level of the packet-switching logic. It is counterproductive because the overhead involved soaks up a significant fraction of

the high capacity provided by the network. Frame relay was developed to take advantage of these high data rates and low error rates. Whereas the original packet-switching networks were designed with a data rate to the end user of about 64 kbps, frame relay networks are designed to operate efficiently at user data rates of up to 2 Mbps. The key to achieving these high data rates is to strip out most of the overhead involved with error control.

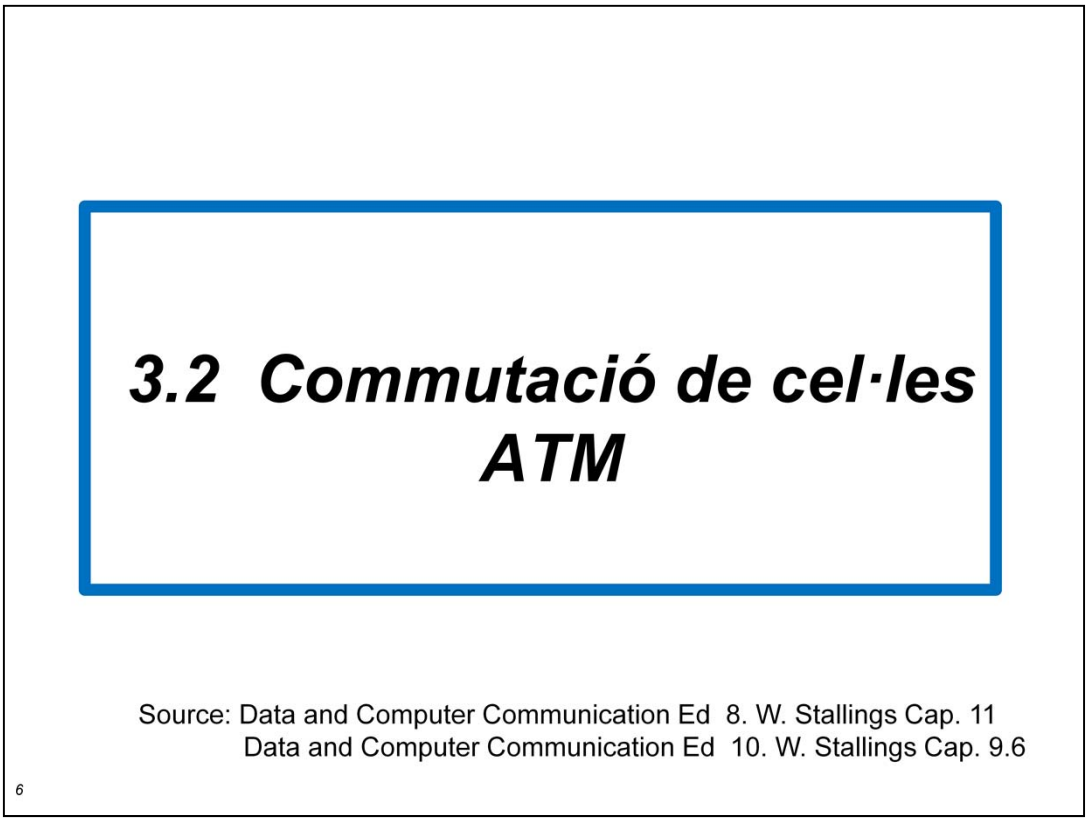
T

Frame relay: LAPF Core

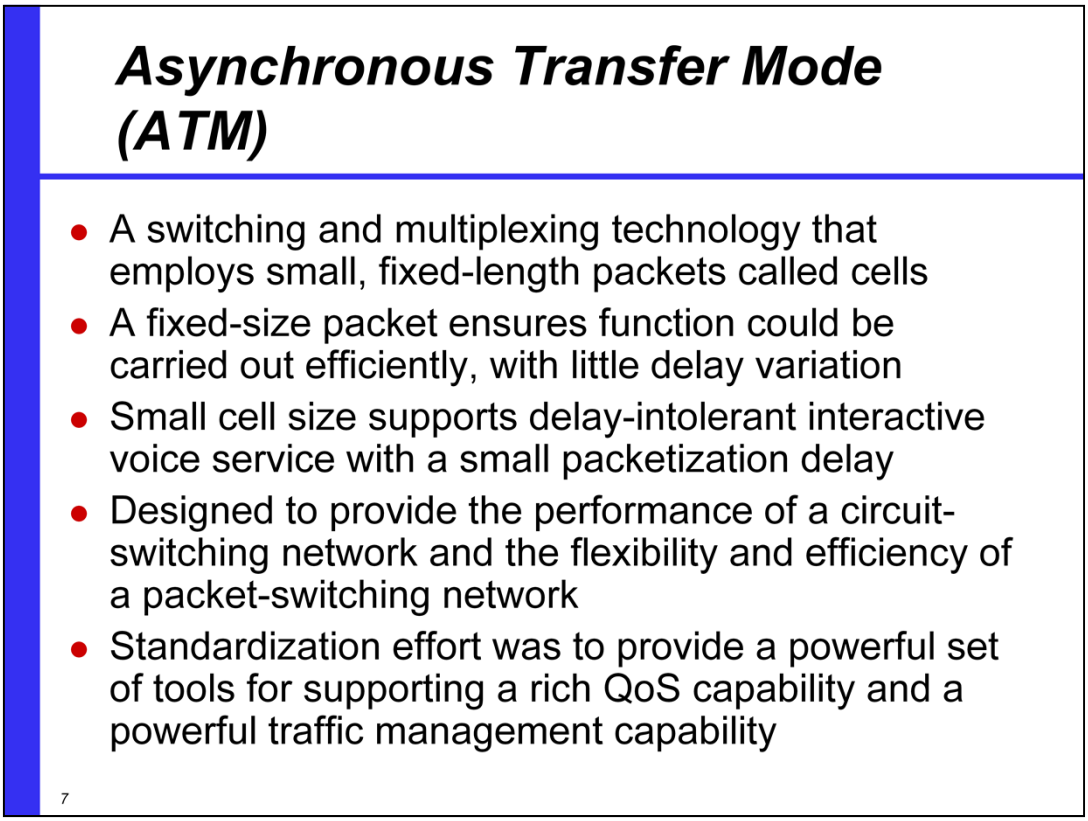


LAP F Core és el protocol de transmissió de dades a nivell 2 que prové de HDLC eliminant el camp de control, i per tant de les funcions implícites a ell, i introdueix el concepte de circuit virtual dins del camp d'adreça.

Els camps de seqüència de comprovació de senyaladors i trames (FCS) funcionen com a l'HDLC. El camp d'informació transporta les dades de les capes superiors. Si l'usuari tria implementar funcions de control d'enllaç de dades addicionals d'extrem a extrem, es pot transportar una trama d'enllaç de dades en aquest camp. De forma específica, una selecció comuna serà utilitzar tot el protocol LAPF (anomenat protocol de control LAPF) per realitzar funcions sobre les funcions centrals de LAPF. Teniu en compte que el protocol implementat d'aquesta manera s'estableix estrictament entre els abonats finals i és transparent a la xarxa de retransmissió de trama. El camp d'adreça té una longitud per defecte de 2 bytes i es pot ampliar a 3 o 4 bytes. Transporta un identificador de connexió d'enllaç de dades (DLCI) de 10, 16 o 23 bits. El DLCI té la mateixa funció que el número de circuit virtual a l'X.25: permet que diverses connexions de retransmissió de trama es multiplexin en un sol canal. Com a l'X.25, l'identificador de connexió només té un significat: cada extrem de la connexió lògica assigna un DLCI propi de l'agrupació de números no utilitzats localment i la xarxa n'ha de mapar un amb l'altre. L'alternativa, utilitzant el mateix DLCI als dos extrems, exigiria algun tipus de gestió global de valors de DLCI. La longitud del camp d'adreça i, per tant, del DLCI, es determina mitjançant els bits d'extensió de camp d'adreça (EA). El bit C/R és específic d'aplicació i no l'utilitza el protocol de retransmissió de trama estàndard. La resta de bits del camp d'adreça tenen a veure amb el control la congestió.

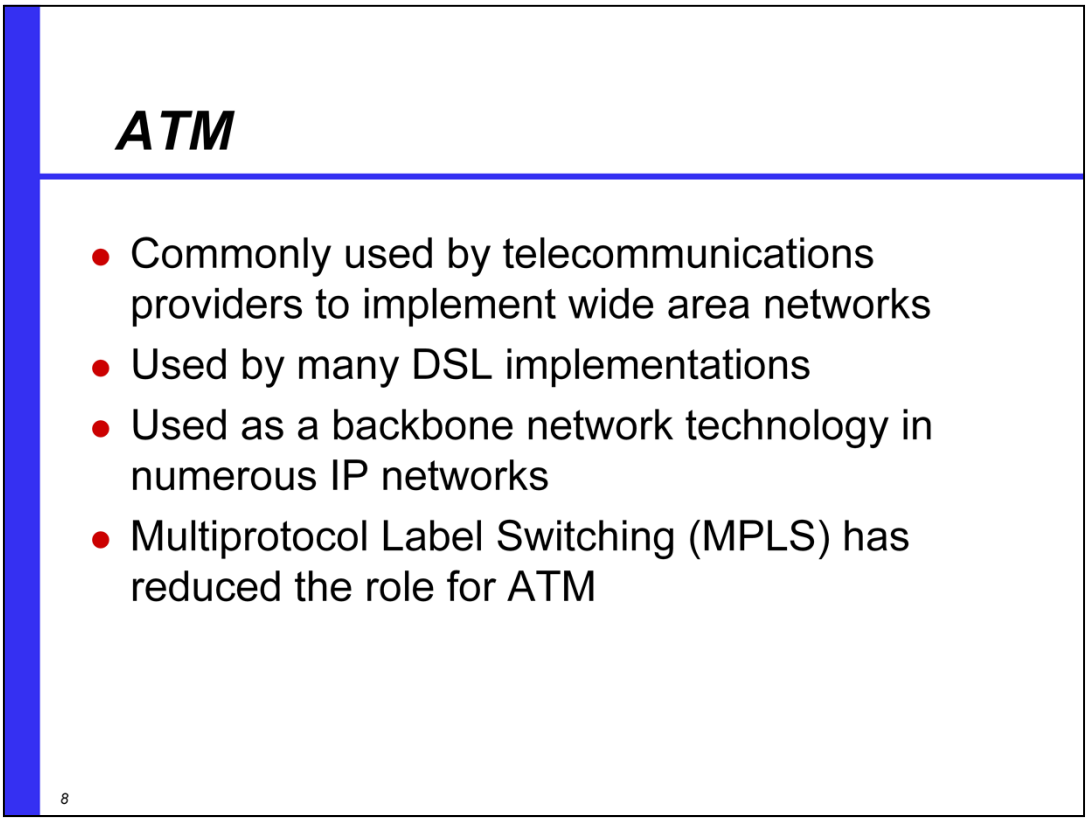


ATM no s'utilitza directament a l'accés, encara que si indirectament a ADSL. Avui dia només s'utilitza en xarxes troncals.



Asynchronous transfer mode is a switching and multiplexing technology that employs small, fixed-length packets called cells . A fixed-size packet was chosen to ensure that the switching and multiplexing function could be carried out efficiently, with little delay variation. A small cell size was chosen primarily to support delay-intolerant interactive voice service with a small packetization delay. ATM is a connection-oriented packet-switching technology that was designed to provide the performance of a circuit-switching network and the flexibility and efficiency of a packet-switching network. A major thrust of the ATM standardization effort was to

provide a powerful set of tools for supporting a rich QoS capability and a powerful traffic management capability. ATM was intended to provide a unified networking standard for both circuit-switched and packet-switched traffic, and to support data, voice, and video with appropriate QoS mechanisms. With ATM, the user can select the desired level of service and obtain guaranteed service quality. Internally, the ATM network makes reservations and preplans routes so that transmission allocation is based on priority and QoS characteristics.



ATM was intended to be a universal networking technology, with much of the switching and routing capability implemented in hardware, and with the ability to support IP-based networks and circuit-switched networks. It was also anticipated that ATM would be used to implement local area networks. ATM never achieved this comprehensive deployment. However, ATM remains an important technology. ATM is commonly used by telecommunications providers to implement wide area networks. Many DSL implementations use ATM over the basic DSL hardware for multiplexing and switching, and ATM is used as a backbone network technology in

numerous IP networks and portions of the Internet. A number of factors have led to this lesser role for ATM. IP, with its many associated protocols, provides an integrative technology that is more scalable and

less complex than ATM. In addition, the need to use small fixed-sized cells to reduce jitter has disappeared as transport speeds have increased. The development of voice and video over IP protocols has provided an integration capability at the IP level. Perhaps the most significant development related to the reduced role for ATM is the widespread acceptance of Multiprotocol Label Switching (MPLS). MPLS is a layer-2 connection-oriented packet-switching protocol that, as the name suggests, can provide a switching service for a variety of protocols and applications, including IP, voice, and video. We introduce MPLS in Chapter 23.

Virtual Channel Connection (VCC)

- Logical connection in ATM
- Analogous to a virtual circuit
- Basic unit of switching in an ATM network
- Set up between two end users through the network, and a variable-rate, full duplex flow of fixed-size cells is exchanged over the connection
- Also used for user-network exchange and network-network exchange

9

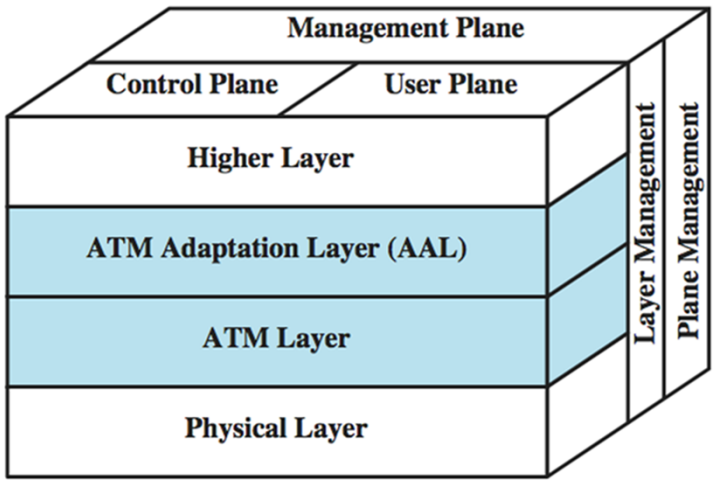
ATM is a packet-oriented transfer mode. It allows multiple logical connections to be multiplexed over a single physical interface. The information flow on each logical connection is organized into fixed-size packets called cells. Logical connections in ATM are referred to as virtual channel connections (VCCs) . A VCC is analogous to a virtual circuit; it is the basic unit of switching in an ATM network. A VCC is set up between two end users through the network, and a variable-rate, full-duplex flow of fixed-size cells is exchanged over the connection. VCCs are also used for user–network exchange (control signaling) and network–network exchange (network

management and routing). For ATM, a second sublayer of processing has been introduced that deals with the concept of virtual path (Figure 9.16). A virtual path connection (VPC) is a bundle of VCCs that have the same endpoints. Thus, all of the cells flowing over alof the VCCs in a single VPC are switched together.

The virtual path concept was developed in response to a trend in high-speed networking in which the control cost of the network is becoming an increasingly higher proportion of the overall network cost. The virtual path technique helps contain the control cost by grouping connections sharing common paths through the network into a single unit. Network management actions can then be applied to a small number of groups of connections instead of a large number of individual connections.

.

Protocol Architecture



10

The standards issued for ATM by ITU-T are based on the protocol architecture shown figure, which illustrates the basic architecture for an interface between user and network. The physical layer involves the specification of a transmission medium and a signal encoding scheme. The data rates specified at the physical layer range from 25.6 Mbps to 622.08 Mbps. Other data rates, both higher and lower, are possible. Two layers of the protocol architecture relate to ATM functions. There is an ATM layer common to all services that provides packet transfer capabilities, and an ATM adaptation layer (AAL) that is service dependent. The ATM layer defines the transmission of data in fixed-size cells and defines the use of logical connections. The use of ATM creates the need for an adaptation layer to support information transfer protocols not based on ATM. The AAL maps higher-layer information into ATM cells to be transported over an ATM network, then collects information from ATM cells for delivery to higher layers.

T

Delay in ATM networks

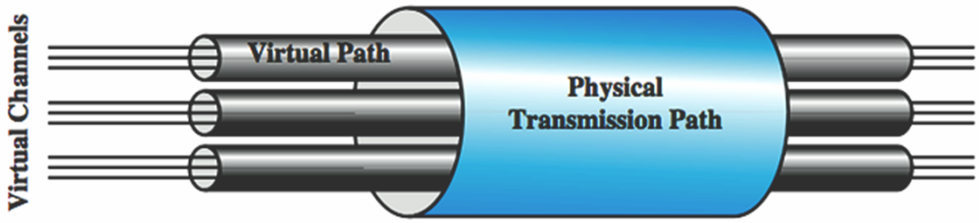
- End to end delay R = Rp + Rt
- Rp (packet delay) = $48x8 / V_{ts}$
- Rt (transfer delay) = Tt + Tp + W
  - *Tt (transmission time)* Σt_t ($t_t = 53x8 / V_{tn}$)
  - *Tp (propagation time)* Σt_p ($t_p = d / V_p$)
  - *W (queue waiting time)* Σw ($w = nxt_t$)

11

El retard en ATM es pot dividir en el retard extrem extrem + retard de paquitització. El primer depen dels temps de transmissió i de propagació. El segon depen del temps d'omplir una cel·la amb dades ja que les cel·les han de anar plenes, a ser possible.

ATM Virtual Path Connection

- virtual path connection (VPC)
  - bundle of VCC with same end points



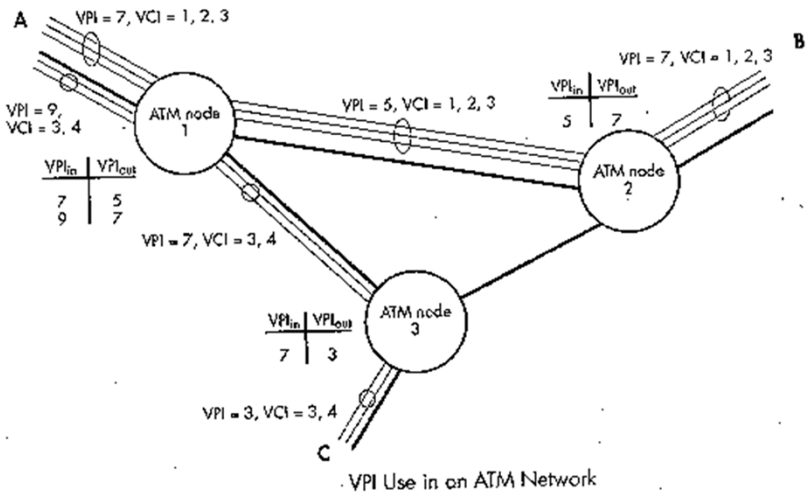
12

For ATM, a second sublayer of processing has been introduced that deals with the concept of virtual path (Stallings DCC9eFigure 11.4). A **virtual path connection** (VPC) is a bundle of VCCs that have the same endpoints. Thus, all of the cells flowing over all of the VCCs in a single VPC are switched together. The virtual path concept was developed in response to a trend in high-speed networking in which the control cost of the network is becoming an increasingly higher proportion of the overall network cost. The virtual path technique helps contain the control cost by grouping connections sharing common paths through the network into a single unit. Network management actions can then be applied to a small number of groups of connections instead of a large number of individual connections.

T

Virtual Channel and Virtual Path

- VPI and VCI relation



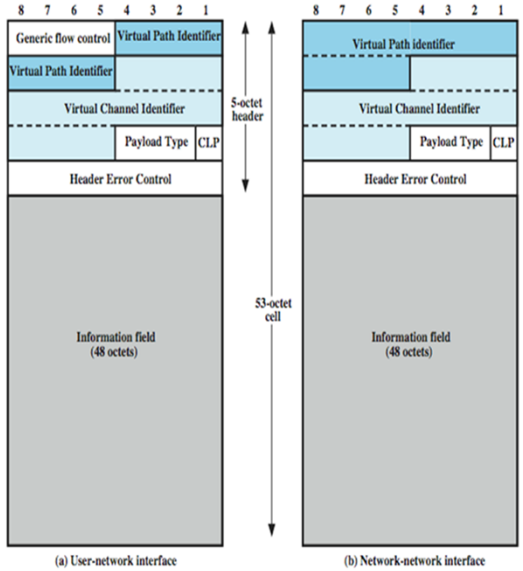
VPI Use in an ATM Network

13

Els circuits virtuals en ATM es concentren actualment en els VP ja que els VC no es fan servir.

Els nodes tenen taules on s'indica #VP #I/F entrada / #VP #I/F sortida. Es creen manualment ja que estan pensats per a xarxes troncals permanents.

ATM Cells



14

The asynchronous transfer mode makes use of fixed-size cells, consisting of a 5-octet header and a 48-octet information field. There are several advantages to the use of small, fixed-size cells. First, the use of small cells may reduce queuing delay for a high-priority cell, because it waits less if it arrives slightly behind a lower-priority cell that has gained access to a resource (e.g., the transmitter). Second, it appears that fixed-size cells can be switched more efficiently, which is important for the very high data rates of ATM [PARE88]. With fixed-size cells, it is easier to implement the switching mechanism in hardware.

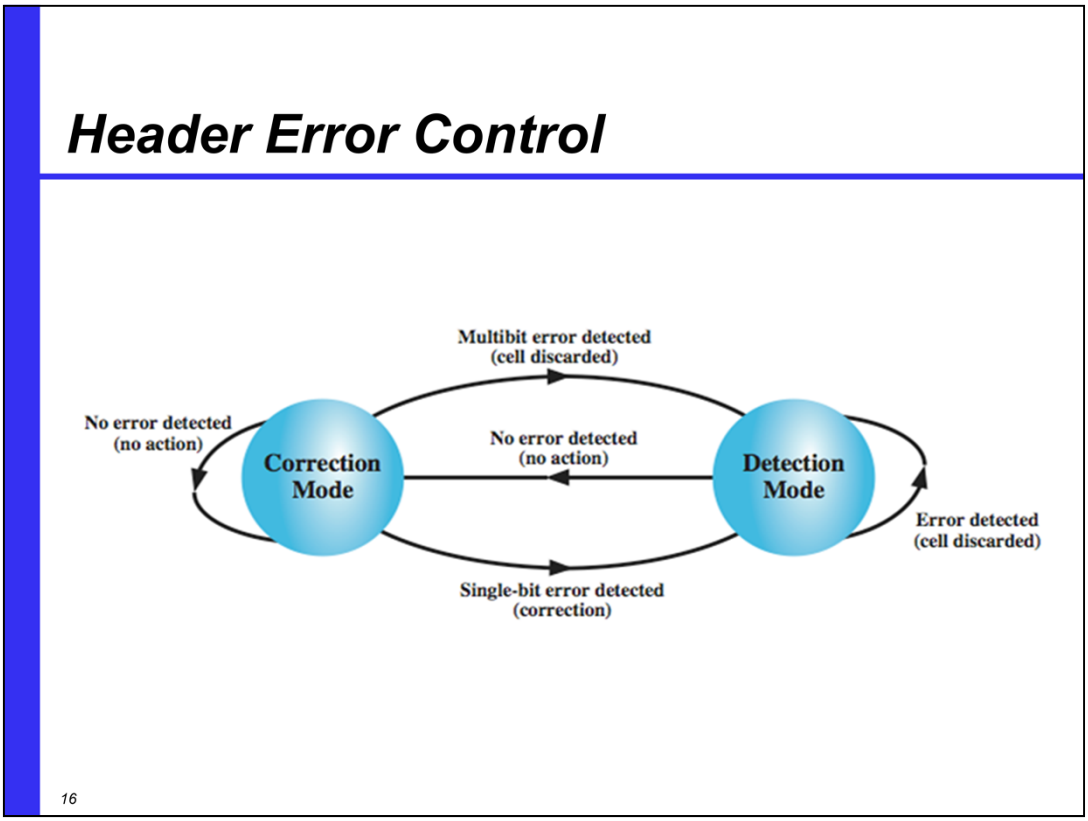
ATM Cells Format

- Payload Type indicator

Payload type	Meaning
000	User data cell, no congestion, cell type 0
001	User data cell, no congestion, cell type 1
010	User data cell, congestion experienced, cell type 0
011	User data cell, congestion experienced, cell type 1
100	Maintenance information between adjacent switches
101	Maintenance information between source and destination switches
110	Resource Management cell (used for ABR congestion control)
111	Reserved for future function

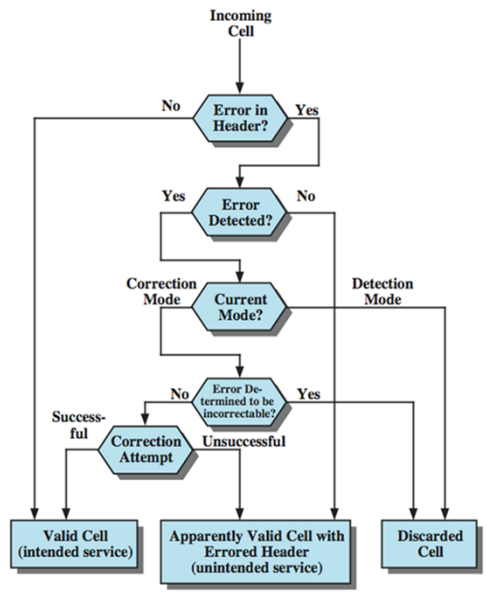
15

El camp Tipus de càrrega útil (PT) indica el tipus d'informació del camp d'informació. La taula 11.2 mostra la interpretació dels bits PT. Un valor 0 al primer bit indica informació d'usuari (és a dir, informació de la següent capa més alta). En aquest cas, el segon bit indica si s'ha experimentat congestió. El tercer bit, conegut com a bit de tipus d'unitat de dades de servei (SDU)1, és un camp d'un bit que es pot utilitzar per discriminar dos tipus d'SDU de l'ATM associats amb una connexió. El terme SDU fa referència a la càrrega útil de 48 bytes de la cel·la. Un valor d'1 al primer bit del camp Tipus de càrrega útil indica que aquesta cel·la transporta informació de gestió o manteniment de la xarxa. Aquesta indicació permet la inserció de cel·les de gestió de xarxa a la VCC d'un usuari, sense que això afecti les dades de l'usuari. Per tant, el camp PT pot proporcionar informació de control en banda.



Each ATM cell includes an 8-bit HEC field that is calculated based on the remaining 32 bits of the header. The polynomial used to generate the code is $X8 + X2 + X + 1$. In most existing protocols that include an error control field, such as HDLC, the data that serve as input to the error code calculation are in general much longer than the size of the resulting error code. This allows for error detection. In the case of ATM, the input to the calculation is only 32 bits, compared to 8 bits for the code. The fact that the input is relatively short allows the code to be used not only for error detection but also, in some cases, for actual error correction. This is because there is sufficient redundancy in the code to recover from certain error patterns. Figure depicts the operation of the HEC algorithm at the receiver. At initialization, the receiver's error correction algorithm is in the default mode for single-bit error correction. As each cell is received, the HEC calculation and comparison is performed. As long as no errors are detected, the receiver remains in error correction mode. When an error is detected, the receiver will correct the error if it is a single-bit error or will detect that a multibit error has occurred. In either case, the receiver now moves to detection mode. In this mode, no attempt is made to correct errors. The reason for this change is a recognition that a noise burst or other event might cause a sequence of errors, a condition for which the HEC is insufficient for error correction. The receiver remains in detection mode as long as errored cells are received. When a header is examined and found not to be in error, the receiver switches back to correction mode.

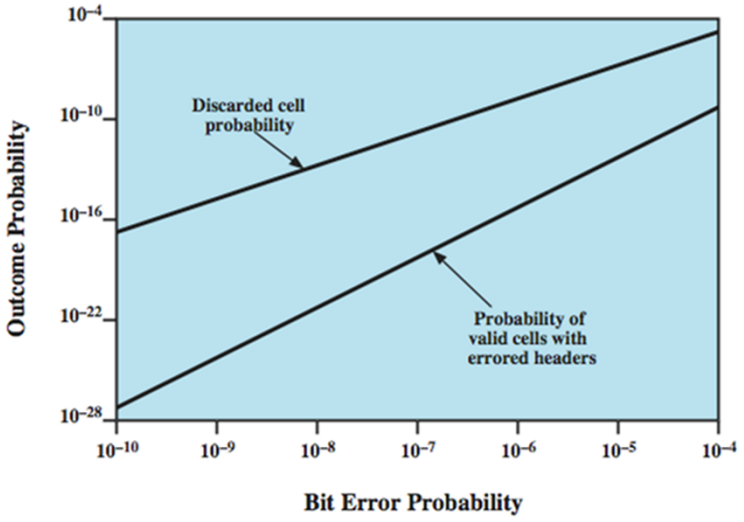
Effect of Error Cell Header



17

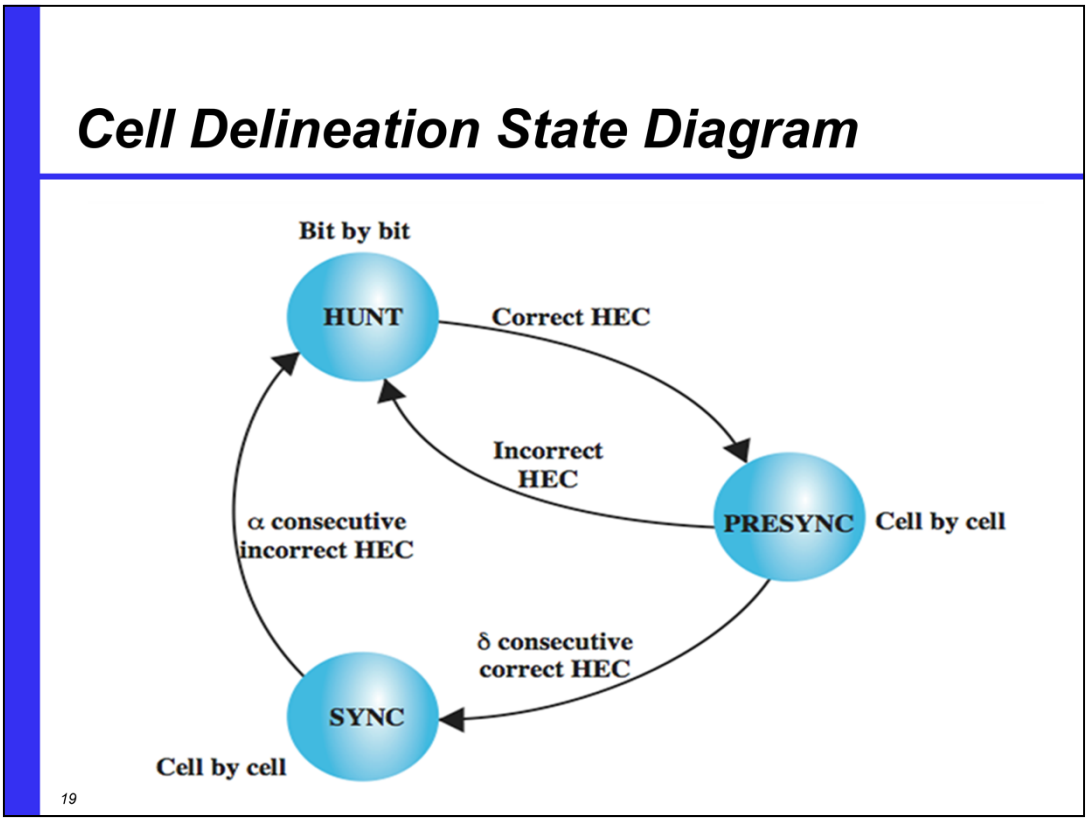
The flowchart of figure shows the consequence of errors in the cell header. The error-protection function provides both recovery from single-bit header errors and a low probability of the delivery of cells with errored headers under bursty error conditions. The error characteristics of fiber-based transmission systems appear to be a mix of single-bit errors and relatively large burst errors. For some transmission systems, the error correction capability, which is more time-consuming, might not be invoked.

Impact of Random Bit Errors on HEC Performance



18

Based on one in ITU-T I.432, indicates how random bit errors impact the probability of occurrence of discarded cells and valid cells with errored headers when HEC is employed.



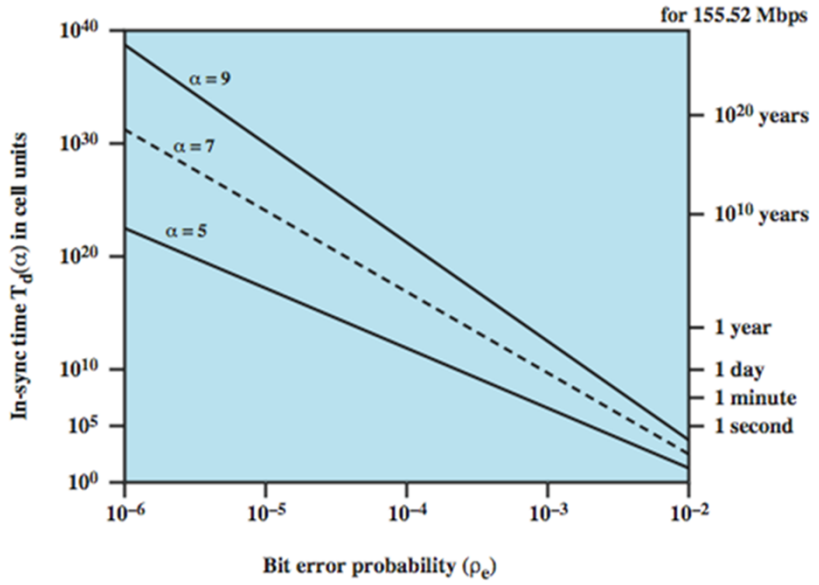
1. In the HUNT state, a cell delineation algorithm is performed bit by bit to determine if the HEC coding law is observed (i.e., match between received HEC and calculated HEC). Once a match is achieved, it is assumed that one header has been found, and the method enters the PRESYNC state.

2.In the PRESYNC state, a cell structure is now assumed. The cell delineation algorithm is performed cell by cell until the encoding law has been confirmed consecutively d times.

3.In the SYNC state, the HEC is used for error detection and correction (see Figure 11.7). Cell delineation is assumed to be lost if the HEC coding law is recognized consecutively as incorrect *a* times.

The values of a and d are design parameters. Greater values of d result in longer delays in establishing synchronization but in greater robustness against false delineation. Greater values of a result in longer delays in recognizing a misalignment but in greater robustness against false misalignment.

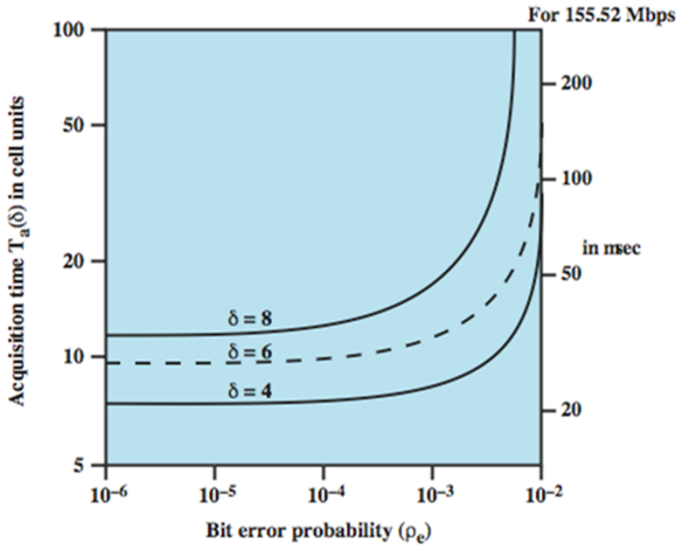
Impact of Random Bit Errors on Cell Delineation Performance



20

Figure based on I.432, show the impact of random bit errors on cell delineation performance for various values of a and d. The first figure shows the average amount of time that the receiver will maintain synchronization in the face of errors, with a as a parameter.

Acquisition Time vs. Bit Error Rate



21

Figure , based on I.432, show the impact of random bit errors on cell delineation performance for various values of δ. The second figure shows the average amount of time to acquire synchronization as a function of error rate, with d as a parameter.

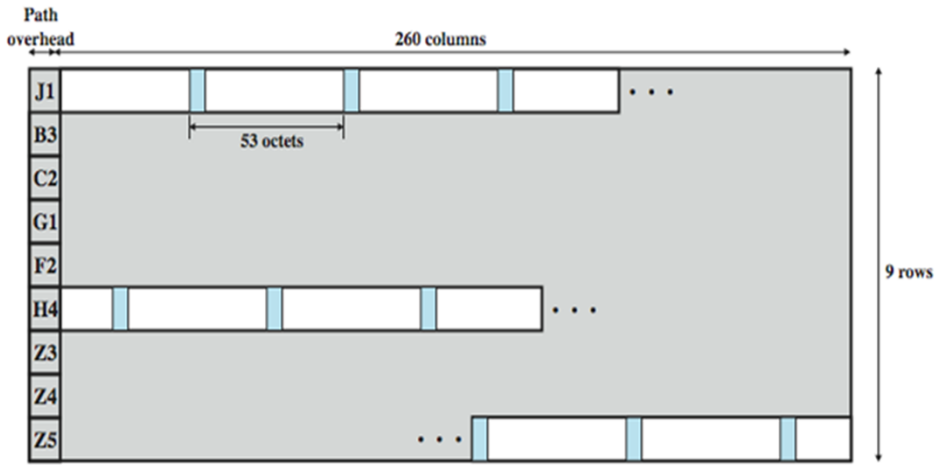
SDH Based Physical Layer

- imposes structure on ATM stream
  - eg. for 155.52Mbps
  - use STM-1 (STS-3) frame
- can carry ATM and STM payloads
- specific connections can be circuit switched using SDH channel
- SDH multiplexing techniques can combine several ATM streams

22

The SDH-based physical layer imposes a structure on the ATM cell stream. For the SDH-based physical layer, framing is imposed using the STM-1 (STS-3) frame. Stallings DCC9e Figure 11.13 shows the payload portion of an STM-1 frame (see Stallings DCC 9eFigure 8.11). This payload may be offset from the beginning of the frame, as indicated by the pointer in the section overhead of the frame. As can be seen, the payload consists of a 9-octet path overhead portion and the remainder, which contains ATM cells. Because the payload capacity (2340 octets) is not an integer multiple of the cell length (53 octets), a cell may cross a payload boundary.

STM-1 Payload for SDH-Based ATM Cell Transmission



23

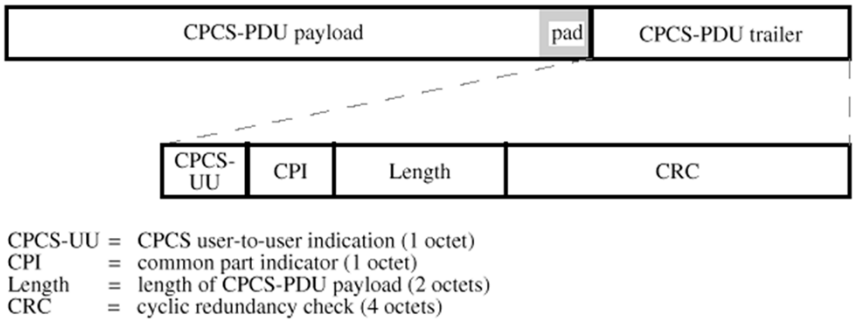
For the SDH-based physical layer, framing is imposed using the STM-1 (STS-3) frame. Stallings DCC9e Figure 11.13 shows the payload portion of an STM-1 frame (for comparison, see Stallings DCC9e Figure 8.11).

The H4 octet in the path overhead is set at the sending side to indicate the next occurrence of a cell boundary. That is, the value in the H4 field indicates the number of octets to the first cell boundary following the H4 octet. The permissible range of values is 0 to 52.The advantages of the SDH-based approach include: It can be used to carry either ATM-based or STM-based (synchronous transfer mode) payloads, making it possible to initially deploy a high-capacity fiber-based transmission infrastructure for a variety of circuit-switched and dedicated applications and then readily migrate to the support of ATM.

Some specific connections can be circuit switched using an SDH channel. For example, a connection carrying constant-bit-rate video traffic can be mapped into its own exclusive payload envelope of the STM-1 signal, which can be circuit switched. This may be more efficient than ATM switching. Using SDH synchronous multiplexing techniques, several ATM streams can be combined to build interfaces with higher bit rates than those supported by the ATM layer at a particular site. For example, four separate ATM streams, each with a bit rate of 155 Mbps (STM-1), can be combined to build a 622-Mbps (STM-4) interface. This arrangement may be more cost effective than one using a single 622-Mbps ATM stream.

T

AAL5 CPCS-PDU



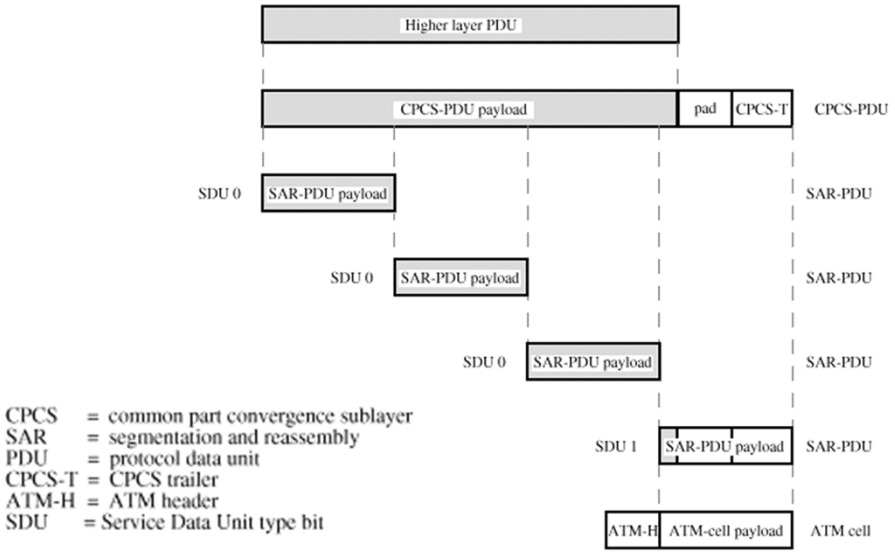
CPCS-UU = CPCS user-to-user indication (1 octet)
CPI = common part indicator (1 octet)
Length = length of CPCS-PDU payload (2 octets)
CRC = cyclic redundancy check (4 octets)

24

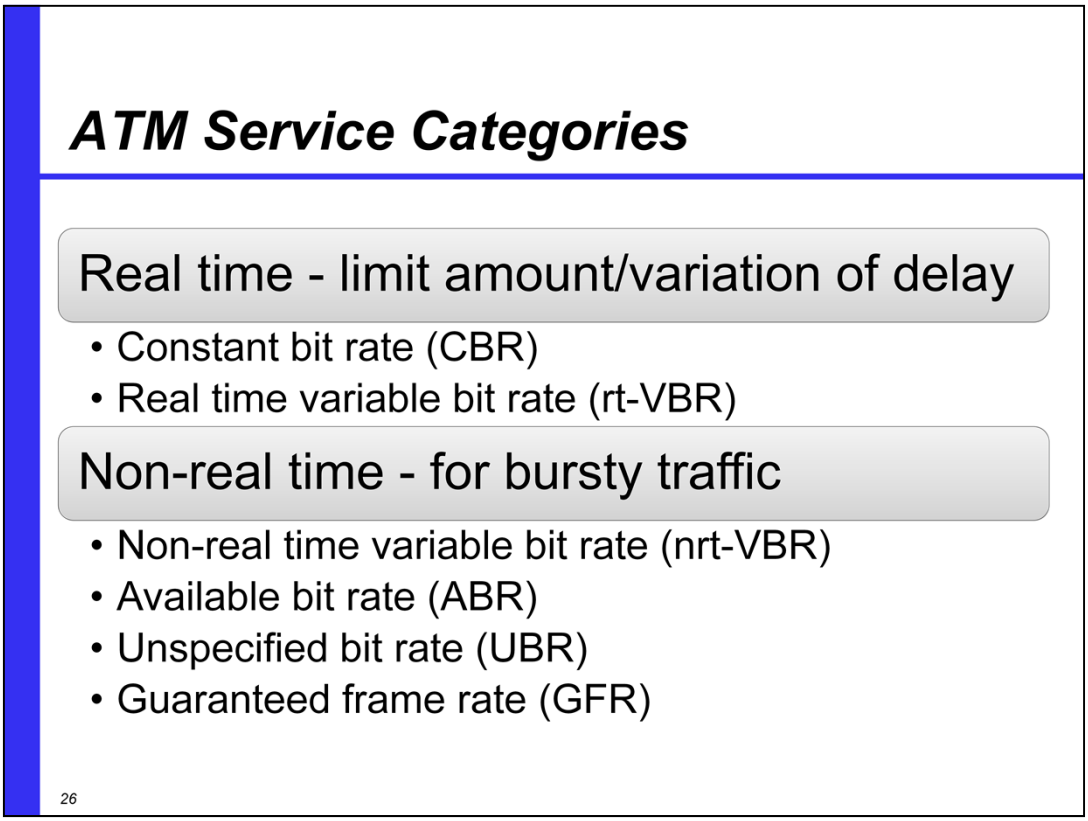
El protocol AAL5 introdueix un trailer que indica la llargària del payload. El PAD es un farcit per fer la llargària del total múltiple de 48 octets, ja que ha d'anar dins de cel·les ATM.

T

AAL5 transmission

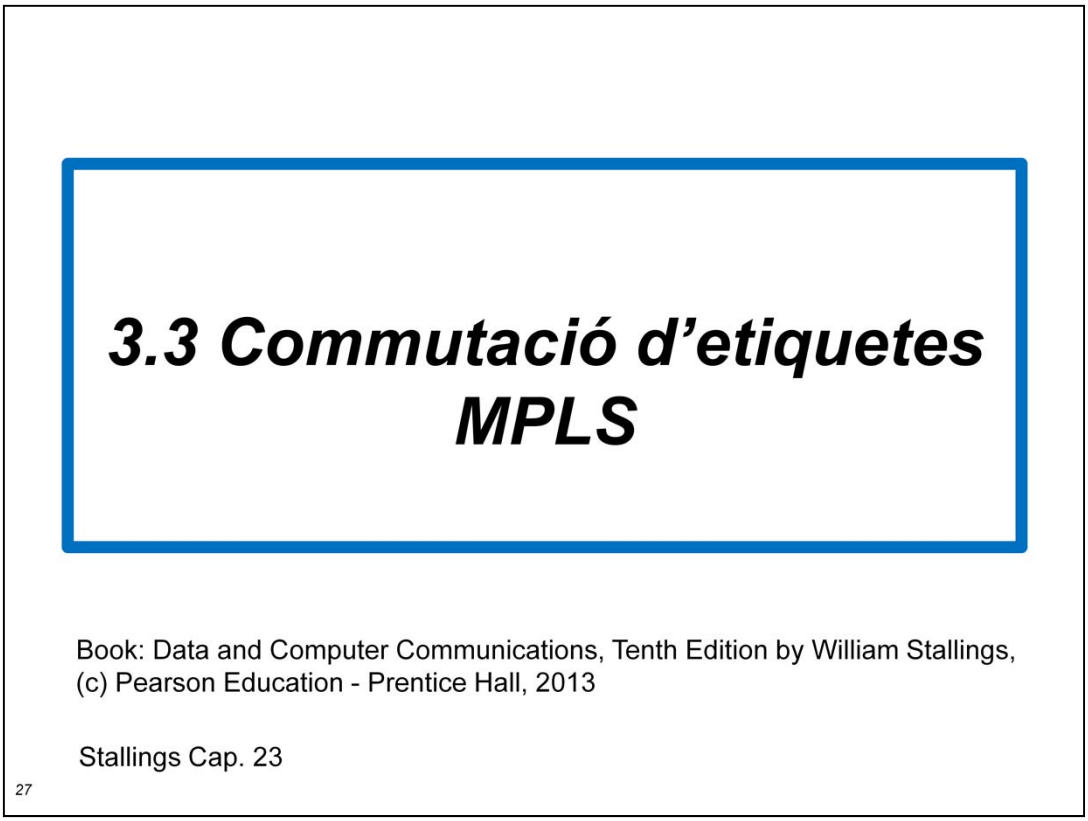


El total AAL5-PDU es divideix per 48, s'introdueix el PAD i cada tros es posa dins una cel·la ATM. La darrera cel·la porta el tercer bit del PTI a 1. Així el receptor pot recomposar el CPCS-PDU original.



An ATM network is designed to be able to transfer many different types of traffic simultaneously, including real-time flows such as voice, video, and bursty TCP flows. Although each such traffic flow is handled as a stream of 53-octet cells traveling through a virtual channel, the way in which each data flow is handled within the network depends on the characteristics of the traffic flow and the requirements of the application. For example, real-time video traffic must be delivered within minimum variation in delay. We examine the way in which an ATM network handles different types of traffic flows in Chapter 13. In this section, we summarize ATM service categories, which are used by an end system to identify the type of service required.

T



Es considera nivell 2.5. Es una forma de crear circuits virtuals pel nivell IP.

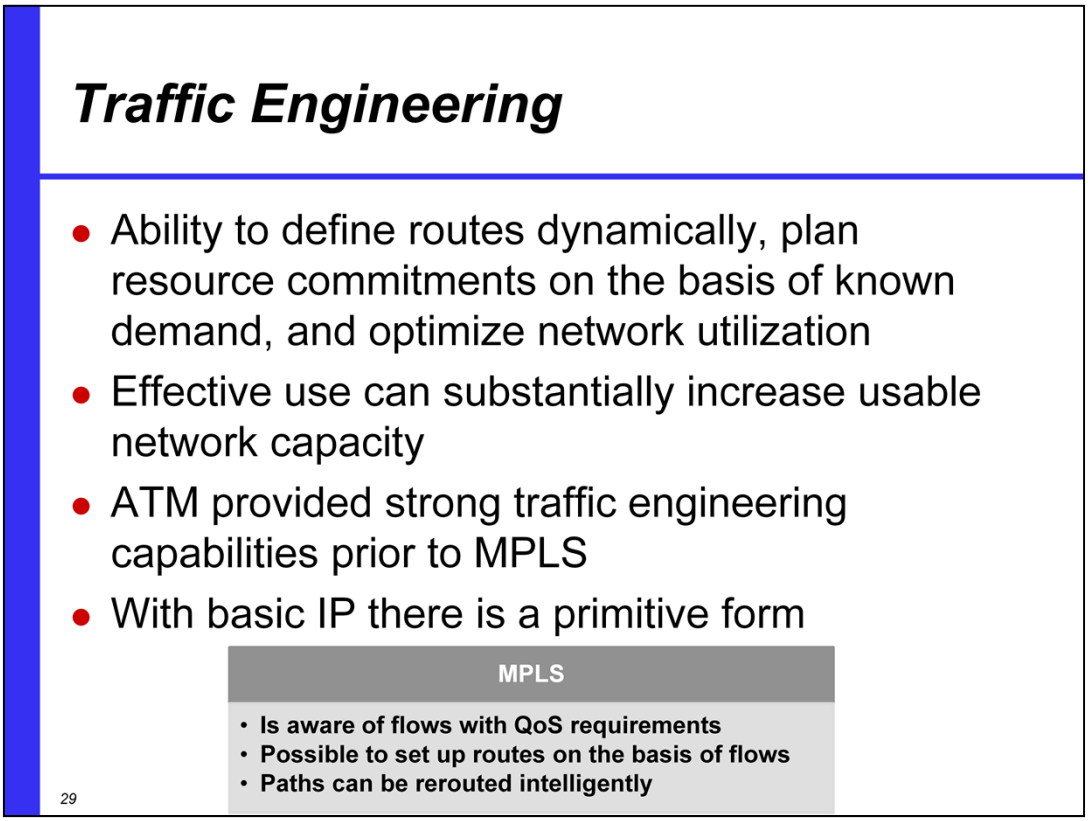
Capítol 23 Ed. 10 Stallings (english)

Role of MPLS

- Efficient technique for forwarding and routing packets
- Designed with IP networks in mind
  - *Can be used with any link-level protocol*
- Fixed-length label encapsulates an IP packet or a data link frame
- MPLS label contains all information needed to perform routing, delivery, QoS, and traffic management functions
- Is connection oriented

28

In essence, MPLS is an efficient technique for forwarding and routing packets. MPLS was designed with IP networks in mind, but the technology can be used without IP to construct a network with any link-level protocol, including ATM and frame relay. In an ordinary packet-switching network packet switches must examine various fields within the packet heard to determine destination, route, quality of service (QoS), and any traffic management functions (such as discard or delay) that may be supported. Similarly, in an IP-based network, routers examine a number of fields in the IP header to determine these functions. In an MPLS network, a fixed-length label encapsulates an IP packet or a data link frame. The MPLS label contains all the information needed by an MPLS-enabled router to perform routing, delivery, QoS, and traffic management functions. Unlike IP, MPLS is connection oriented.



MPLS makes it easy to commit network resources in such a way as to balance the load in the face of a given demand and to commit to differential levels of support to meet various user traffic requirements. The ability to define routes dynamically, plan resource commitments on the basis of known demand, and optimize network utilization is referred to as **traffic engineering**. Prior to the advent of MPLS, the one networking technology that provided strong traffic engineering capabilities was ATM. With the basic IP mechanism, there is a primitive form of automated traffic engineering. Specifically, routing protocols such as OSPF enable routers to dynamically change the route to a given destination on a packet-by-packet basis to try to balance load. But such dynamic routing reacts in a very simple manner to congestion and does not provide a way to support QoS. All traffic between two endpoints follows the same route, which may be changed when congestion occurs. MPLS, on the other hand, is aware of not just individual packets but flows of packets in which each flow has certain QoS requirements and a predictable traffic demand. With MPLS, it is possible to set up routes on the basis of these individual flows, with two different flows between the same endpoints perhaps following different routers. Further, when congestion threatens, MPLS paths can be rerouted intelligently. That is, instead of simply changing the route on a packet-by-packet basis, with MPLS, the routes are changed on a flow-by-flow basis, taking advantage of the known traffic demands of each flow. Effective use of traffic engineering can substantially increase usable network capacity.

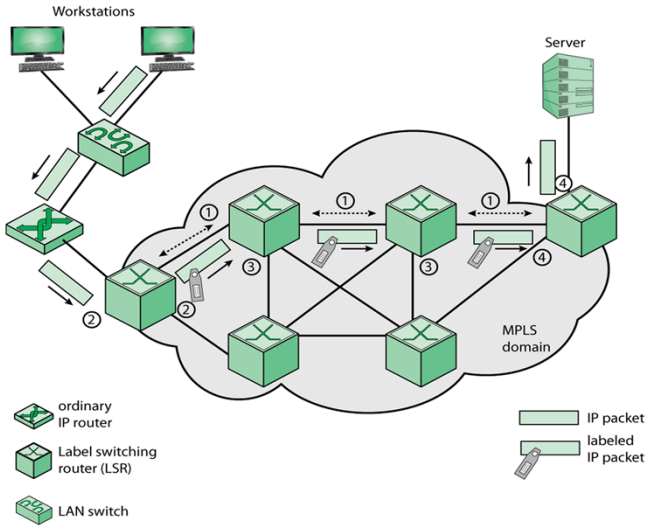
MPLS Operation

- Label switching routers (LSRs)
  - *Nodes capable of switching and routing packets on the basis of label*
- Labels define a flow of packets between two endpoints
- Assignment of a particular packet is done when the packet enters the network of MPLS routers
- Connection-oriented technology

30

An MPLS network or internet consists of a set of nodes, called **label switching routers** (LSRs) capable of switching and routing packets on the basis of which a label has been appended to each packet. Labels define a flow of packets between two endpoints or, in the case of multicast, between a source endpoint and a multicast group of destination endpoints. For each distinct flow, called a **forwarding equivalence class** (FEC), a specific path through the network of LSRs is defined, called a **label switched path** (LSP). In essence, an FEC represents a group of packets that share the same transport requirements. All packets in an FEC receive the same treatment en route to the destination. These packets follow the same path and receive the same QoS treatment at each hop. In contrast to the forwarding in ordinary IP networks, the assignment of a particular packet to a particular FEC is done just once, when the packet enters the network of MPLS routers. Thus, MPLS is a connection-oriented technology. Associated with each FEC is a traffic characterization that defines the QoS requirements for that flow. The LSRs need not examine or process the IP header but rather simply forward each packet based on its label value. Each LSR builds a table, called a **label information base** (LIB), to specify how a packet must be treated and forwarded. Thus, the forwarding process is simpler than with an IP router.

MPLS Operation



31

Figure 21.1 MPLS Operation

Figure depicts the operation of MPLS within a domain of MPLS-enabled routers. The following are key elements of the operation:

1.Prior to the routing and delivery of packets in a given FEC, a path through the network, known as a **label switched path** (LSP), must be defined and the QoS parameters along that path must be established. The QoS parameters determine (1) how much resources to commit to the path, and (2) what queuing and discarding policy to establish at each LSR for packets in this FEC. To accomplish these tasks, two protocols are used to exchange the necessary information among routers: **(a)**An interior routing protocol, such as OSPF, is used to exchange reachability and routing information. **(b)**Labels must be assigned to the packets for a particular FEC. Because the use of globally unique labels would impose a management burden and limit the number of usable labels, labels have local significance only, as discussed subsequently. A network operator can specify explicit routes manually and assign the appropriate label values. Alternatively, a protocol is used to determine the route and establish label values between adjacent LSRs. Either of two protocols can be used for this purpose: the Label Distribution Protocol (LDP) or an enhanced version of RSVP. LDP is now considered the standard technique, with the RSVP approach deprecated.

2.A packet enters an MPLS domain through an ingress edge LSR, where it is processed to determine which network-layer services it requires, defining its QoS. The LSR assigns this packet to a particular FEC, and therefore a particular LSP; appends the appropriate label to the packet; and forwards the packet. If no LSP yet exists for this FEC, the edge LSR must cooperate with the other LSRs in defining a new LSP.

3.Within the MPLS domain, as each LSR receives a labeled packet, it **(a)**Removes the incoming label and attaches the appropriate outgoing label to the packet **(b)** Forwards the packet to the next LSR along the LSP

4.The egress edge LSR strips the label, reads the IP packet header, and forwards the packet to its final destination.

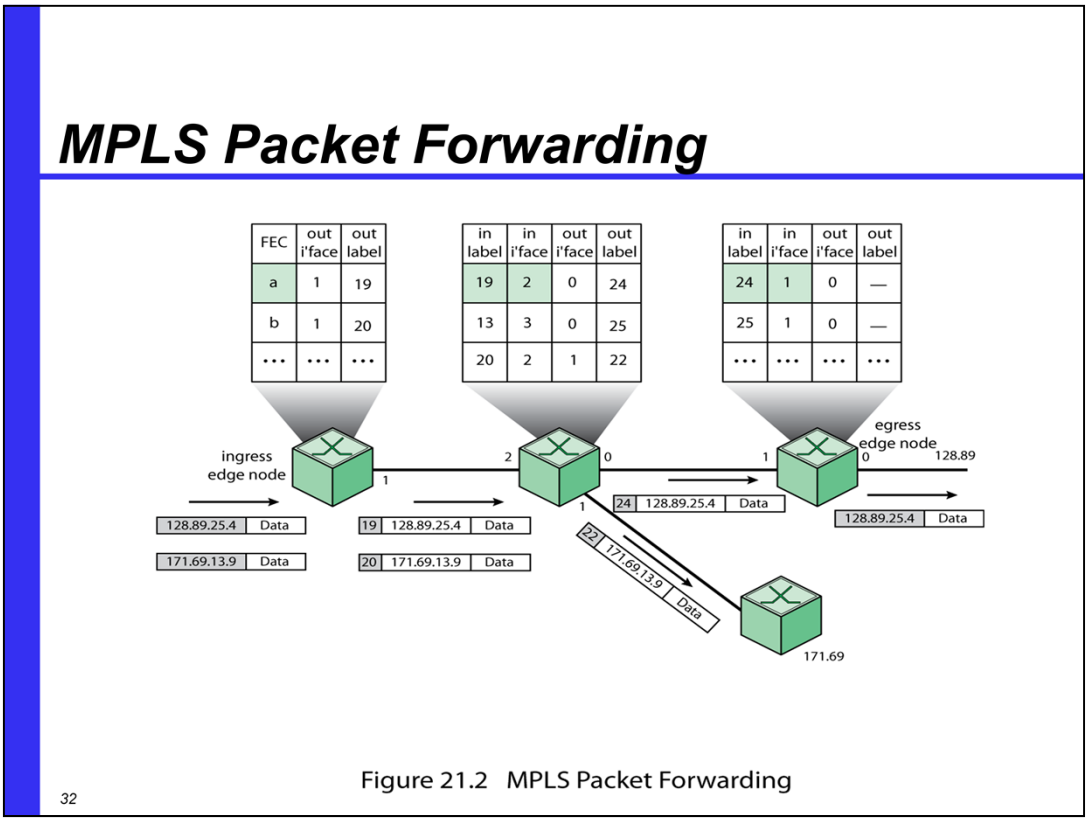
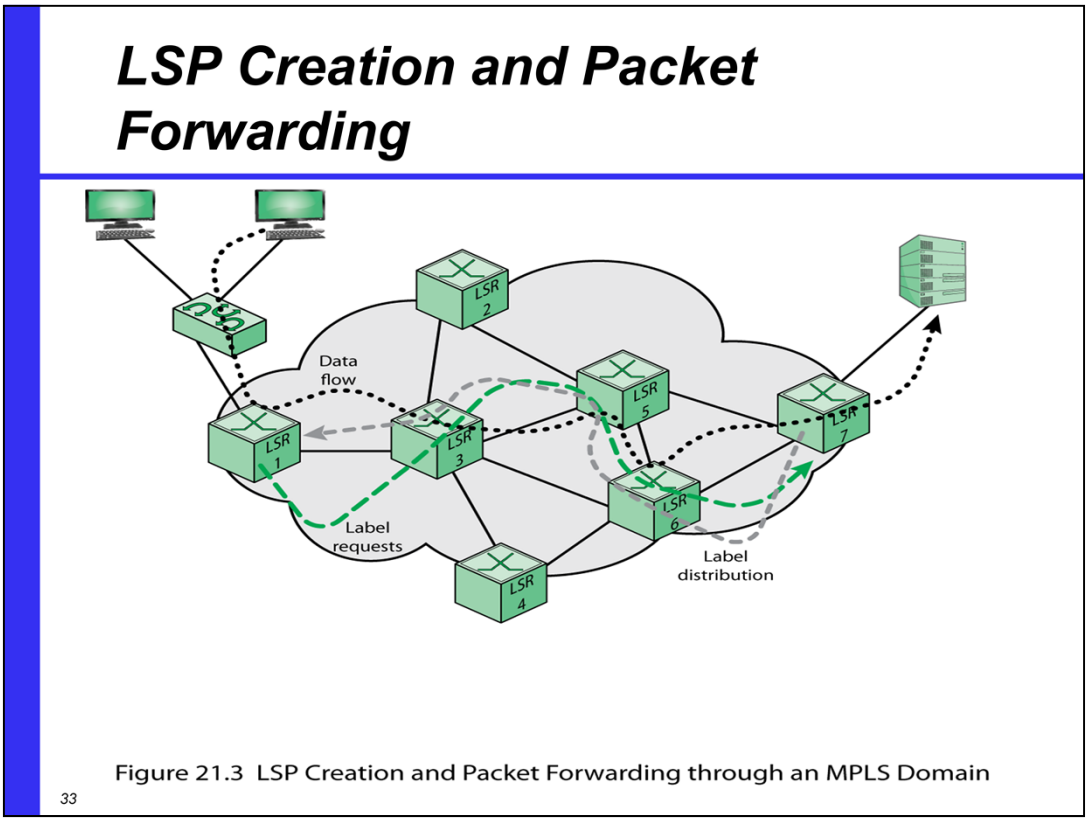
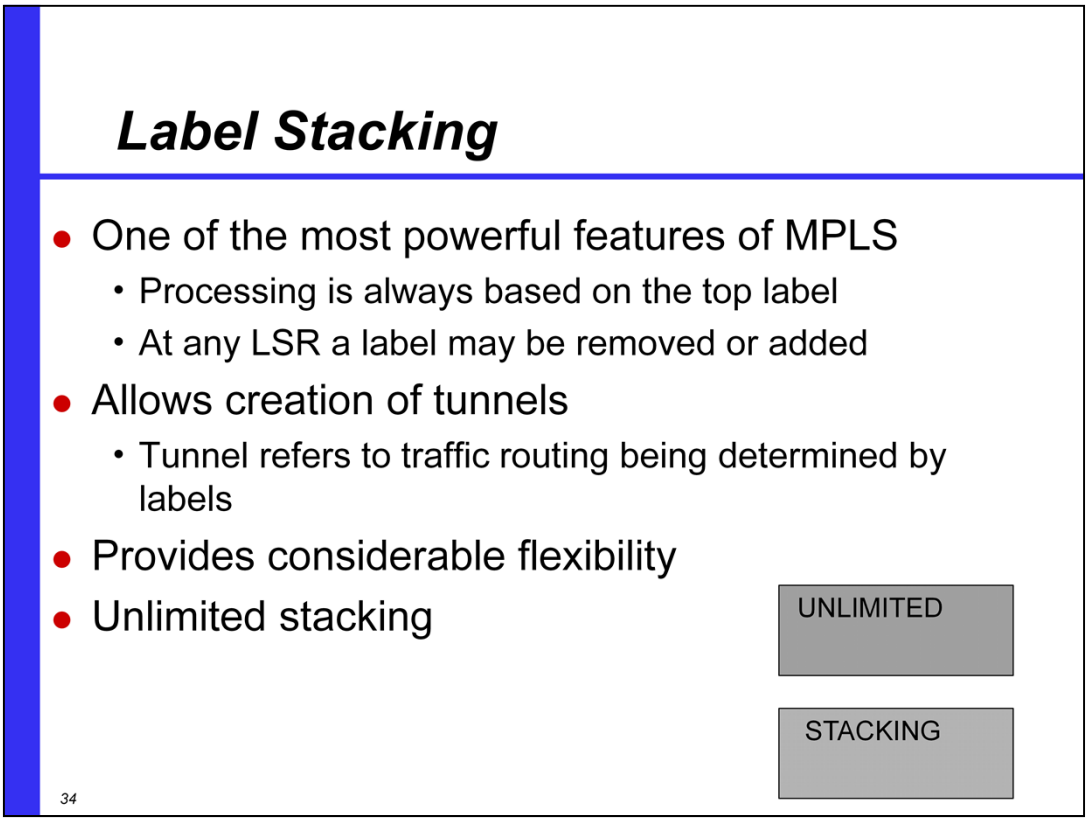


Figure 21.2 MPLS Packet Forwarding

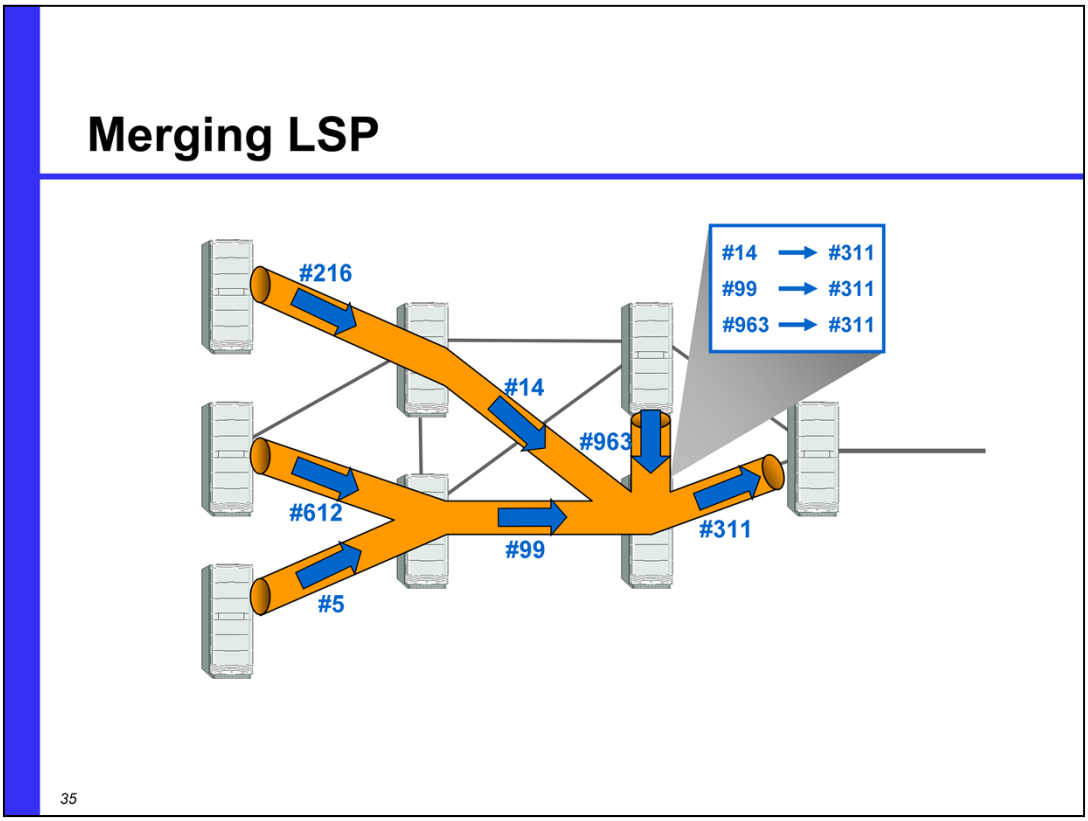
Stallings DCC9e Figure 21.2 shows the label-handling and forwarding operation in more detail. Each LSR maintains a forwarding table for each LSP passing through the LSR. When a labeled packet arrives, the LSR indexes the forwarding table to determine the next hop. For scalability, as was mentioned, labels have local significance only. Thus, the LSR removes the incoming label from the packet and attaches the matching outgoing label before forwarding the packet. The ingress edge LSR determines the FEC for each incoming unlabeled packet and, on the basis of the FEC, assigns the packet to a particular LSP, attaches the corresponding label, and forwards the packet. In this example, the first packet arrives at the edge LSR, which reads the IP header for the destination address prefix, 128.89. The LSR then looks up the destination address in the switching table, inserts a label with a 20-bit label value of 19, and forwards the labeled packet out interface 1. This interface is attached via a link to a core LSR, which receives the packet on its interface 2. The LSR in the core reads the label and looks up its match in its switching table, then replaces label 19 with label 24, and forwards it out interface 0. The egress LSR reads and looks up label 4 in its table, which says to strip the label and forward the packet out interface 0.



Let us now look at an example that illustrates the various stages of operation of MPLS, Figure 21.3. We examine the path of a packet as it a source workstation to a destination server. Across the MPLS network, the packet enters at egress node LSR 1. Assume that this is the first occurrence of a packet on a new flow of packets, so that LSR 1 does not have a label for the packet. LSR 1 consults the IP header to find the destination address and then determine the next hop. Assume in this case that the next hop is LSR 3. Then, LSR 1 initiates a label request toward LSR 3. This request propagates through the network as indicated by the dashed green line.Each intermediate router receives a label from its downstream router starting from LSR 7 and going upstream until LSR 1, setting up an LSP. The LSP setup is indicated by the dashed grey line. The setup can be performed using LDP and may or may not involve traffic engineering considerations. LSR 1 is now able to insert the appropriate label and forward the packet to LSR 3. Each subsequent LSR (LSR 5, LSR 6, LSR 7) examines the label in the received packet, replaces it with the outgoing label, and forwards it. When the packet reaches LSR 7, the LSR removes the label because the packet is departing the MPLS domain and delivers the packet to the destination.



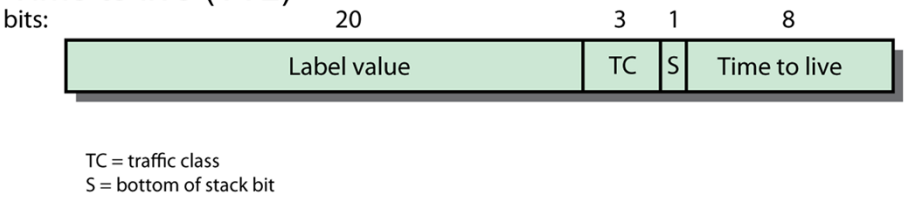
One of the most powerful features of MPLS is label stacking. A labeled packet may carry a number of labels, organized as a last-in-first-out stack. Processing is always based on the top label. At any LSR, a label may be added to the stack (push operation) or removed from the stack (pop operation). Label stacking allows the aggregation of LSPs into a single LSP for a portion of the route through a network, creating a tunnel. The term *tunnel* refers to the fact that traffic routing is determined by labels, and is exercised below normal IP routing and filtering mechanisms. At the beginning of the tunnel, an LSR assigns the same label to packets from a number of LSPs by pushing the label onto each packet's stack. At the end of the tunnel, another LSR pops the top element from the label stack, revealing the inner label. This is similar to ATM, which has one level of stacking (virtual channels inside virtual paths) but MPLS supports unlimited stacking. Label stacking provides considerable flexibility. An enterprise could establish MPLS-enabled networks at various sites and establish a number of LSPs at each site. The enterprise could then use label stacking to aggregate multiple flows of its own traffic before handing it to an access provider. The access provider could aggregate traffic from multiple enterprises before handing it to a larger service provider. Service providers could aggregate many LSPs into a relatively small number of tunnels between points of presence. Fewer tunnels means smaller tables, making it easier for a provider to scale the network core.



Els LSR es pode agregar en una ruta. Per fer això es fa servir les taules I/O del circuit virtual.

Label Format

- defined in RFC 3032
- 32-bit field consisting of:
  - Label value
  - Traffic class (TC)
  - S
  - Time to live (TTL)



TC = traffic class
S = bottom of stack bit

36

Figure 21.4 MPLS Label Format

An MPLS label is a 32-bit field consisting of the following elements (Stallings DCC9e Figure 21.4), defined in RFC 3032:

Label value: Locally significant 20-bit label. Values 0 through 15 are reserved.

Traffic class (TC): 3 bits used to carry traffic class information.

S: Set to one for the oldest entry in the stack, and zero for all other entries. Thus, this bit marks the bottom of the stack.

Time to live (TTL): 8 bits used to encode a hop count, or time to live, value.

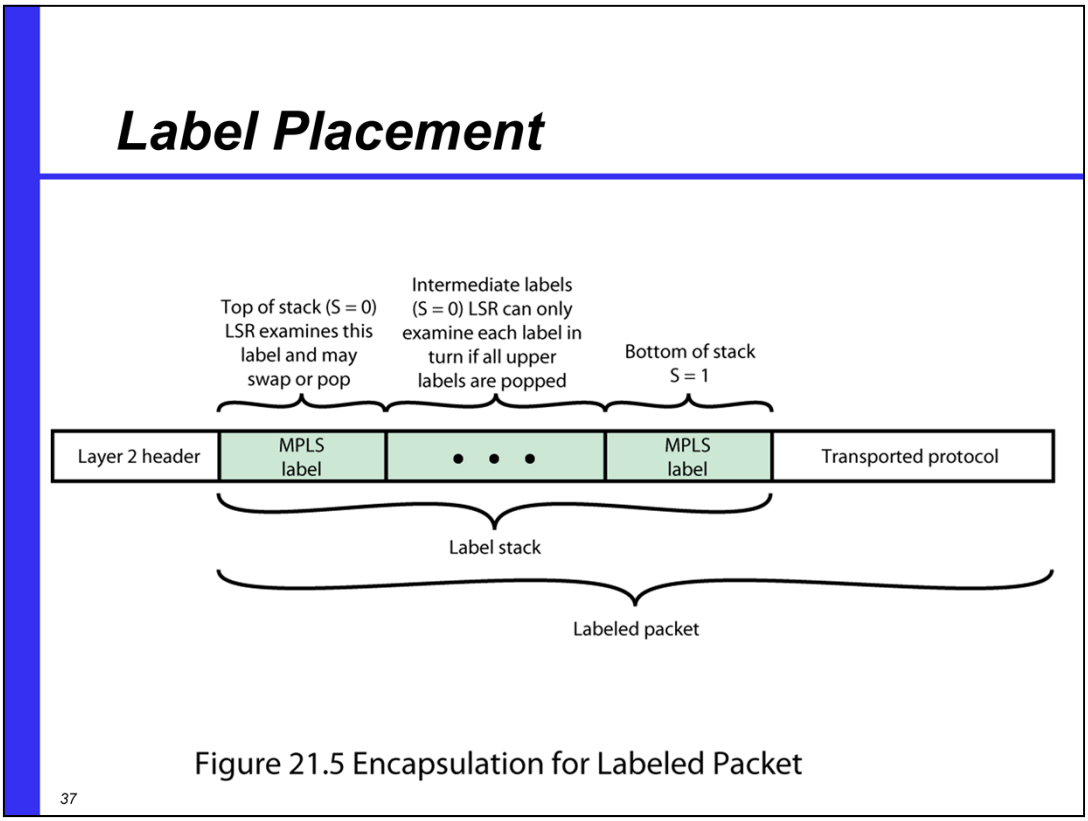


Figure 21.5 Encapsulation for Labeled Packet

The label stack entries appear after the data link layer headers, but before any network layer headers. The top of the label stack appears earliest in the packet (closest to the data link header), and the bottom appears latest (closest to the network layer header), as shown in Figure. The network layer packet immediately follows the label stack entry that has the S bit set.

Label Stack

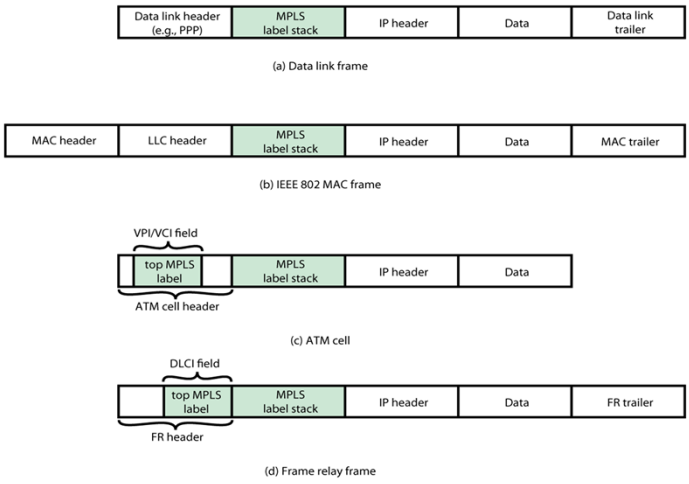
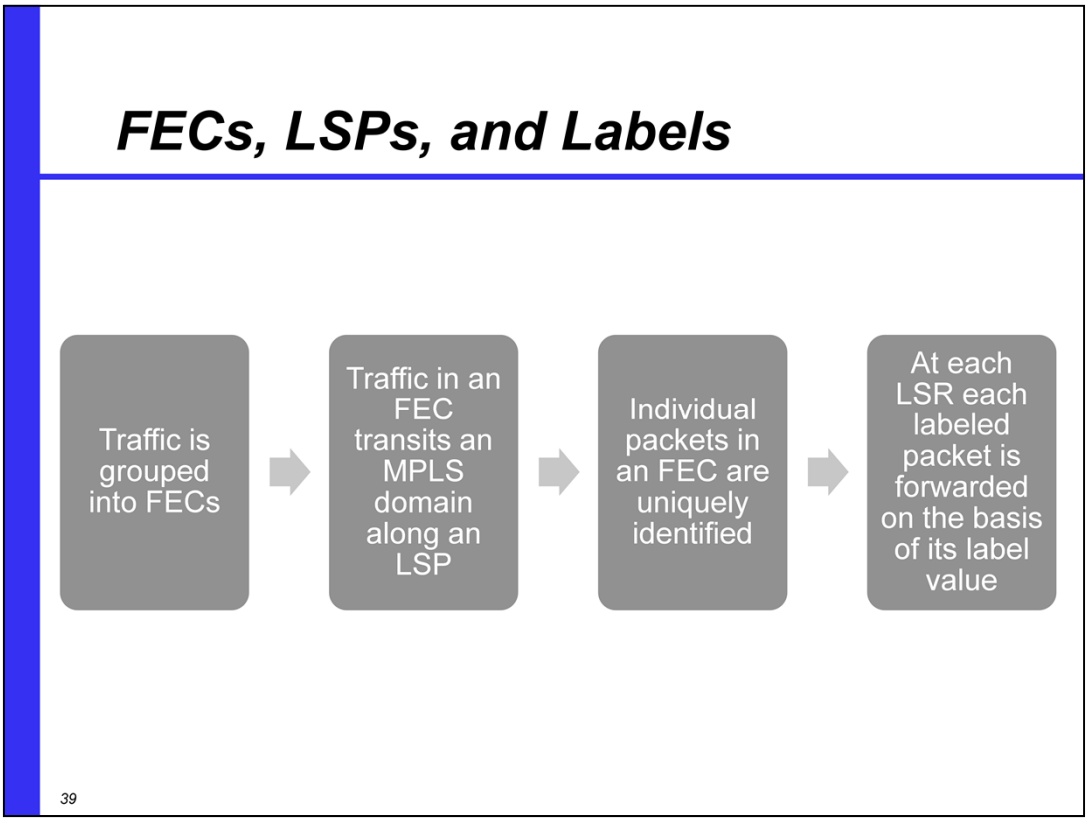


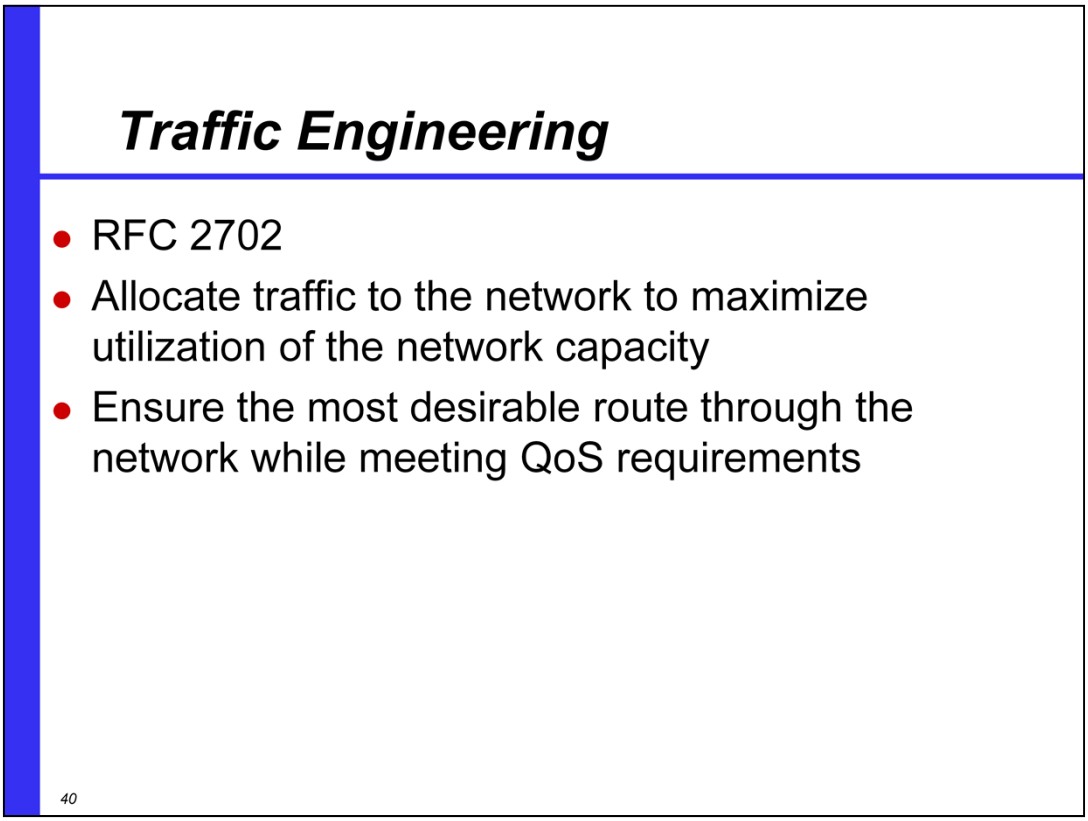
Figure 21.6 Position of MPLS Label Stack

38

In data link frame, such as for PPP (point-to-point protocol), the label stack appears between the IP header and the data link header (Stallings DCC9e Figure 21.6a). For an IEEE 802 frame, the label stack appears between the IP header and the LLC (logical link control) header (Figure 21.6b). If MPLS is used over a connection-oriented network service, a slightly different approach may be taken, as shown in Figures 21.6c and d. For ATM cells, the label value in the topmost label is placed in the VPI/VCI field in the ATM cell header. The entire top label remains at the top of the label stack, which is inserted between the cell header and the IP header. Placing the label value in the ATM cell header facilitates switching by an ATM switch, which would, as usual, only need to look at the cell header. Similarly, the topmost label value can be placed in the DLCI (data link connection identifier) field of a frame relay header. Note that in both these cases, the Time to Live field is not visible to the switch and so is not decremented. The reader should consult the MPLS specifications for the details of the way this situation is handled.



To understand MPLS, it is necessary to understand the operational relationship among FECs, LSPs, and labels. The specifications covering all of the ramifications of this relationship are lengthy. In the remainder of this section, we provide a summary. The essence of MPLS functionality is that traffic is grouped into FECs. The traffic in an FEC transits an MPLS domain along an LSP. Individual packets in an FEC are uniquely identified as being part of a given FEC by means of a locally significant label. At each LSR, each labeled packet is forwarded on the basis of its label value, with the LSR replacing the incoming label value with an outgoing label value. The overall scheme described in the previous paragraph imposes a number of requirements. Specifically, **1.** Each traffic flow must be assigned to a particular FEC. **2.** A routing protocol is needed to determine the topology and current conditions in the domain so that a particular LSP can be assigned to an FEC. The routing protocol must be able to gather and use information to support the QoS requirements of the FEC. **3.** Individual LSRs must become aware of the LSP for a given FEC, must assign an incoming label to the LSP, and must communicate that label to any other LSR that may send it packets for this FEC. The first requirement is outside the scope of the MPLS specifications. The assignment needs to be done either by manual configuration, or by means of some signaling protocol, or by an analysis of incoming packets at ingress LSRs.



RFC 2702 (Requirements for Traffic Engineering Over MPLS) describes traffic engineering as follows: Traffic Engineering (TE) is concerned with performance optimization of operational networks. In general, it encompasses the application of technology and scientific principles to the measurement, modeling, characterization, and control of Internet traffic, and the application of such knowledge and techniques to achieve specific performance objectives. The aspects of Traffic Engineering that are of interest concerning MPLS are measurement and control.The goal of MPLS traffic engineering is twofold. First, traffic engineering seeks to allocate traffic to the network to maximize utilization of the network capacity. And second, traffic engineering seeks to ensure the most desirable route through the network for packet traffic, taking into account the QoS requirements of the various packet flows. In performing traffic engineering, MPLS may override the shortest path or least-cost route selected by the interior routing protocol for a given source-destination flow.

Example of Traffic Engineering

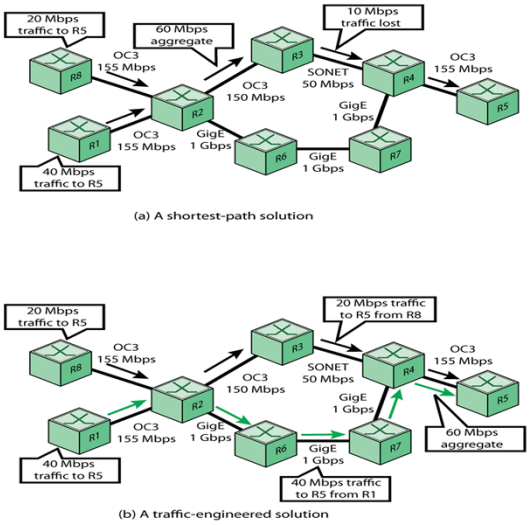
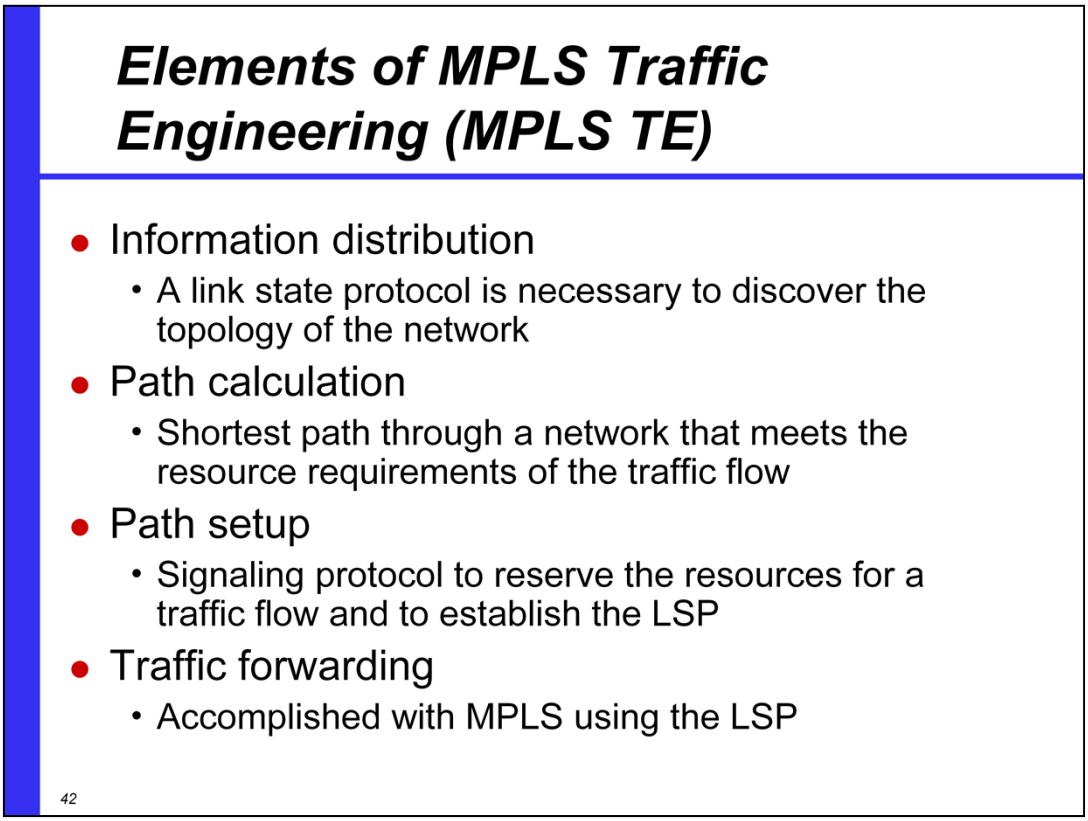


Figure 21.9 Traffic Engineering Example

41

Figure provides a simple example of traffic engineering. Both R1 and R8 have a flow of packets to send to R5. Using OSPF or some other routing protocol, the shortest path is calculated as R2-R3-R4. However, if we assume that R8 has a steady-state traffic flow of 20 Mbps and R1 has a flow of 40 Mbps, then the aggregate flow over this route will be 60 Mbps, which will exceed the capacity of the R3-R4 link. As an alternative, a traffic engineering approach is to determine a route from source to destination ahead of time and reserve the required resources along the way by setting up a LSP and associating resource requirements with that LSP. In this case, the traffic from R8 to R5 follows the shortest route, but the traffic from R1 to R5 follows a longer route that avoids overloading the network.



MPLS TE works by learning about the topology and resources available in a network. It then maps the traffic flows to a particular path based on the resources that the traffic flow requires and the available resources. MPLS TE builds unidirectional LSPs from a source to the destination, which are then used for forwarding traffic. The point where the LSP begins is called LSP headend or LSP source, and the node where the LSP ends is called LSP tailend or LSP tunnel destination. LSP tunnels allow the implementation of a variety of policies related to network performance optimization. For example, LSP tunnels can be automatically or manually routed away from network failures, congestion, and bottlenecks. Furthermore, multiple parallel LSP tunnels can be established between two nodes, and traffic between the two nodes can be mapped onto the LSP tunnels according to local policy. The following components work together to implement MPLS TE: **Information distribution:** A link state protocol, such as Open Shortest Path First (OSPF), is necessary to discover the topology of the network. OSPF is enhanced to carry additional information related to TE, such as bandwidth available and other related parameters. OSPF uses Type 10 (Opaque) Link State Advertisements (LSAs) for this purpose. **Path calculation:** Once the topology of the network and the alternative routes are known, a constraint-based routing scheme is used for finding the shortest path through a particular network that meets the resource requirements of the traffic flow. The Constrained Shortest Path First (CSPF) algorithm (discussed subsequently), which operates on the LSP headend is used for this functionality. **Path setup:** is a signaling protocol to reserve the resources for a traffic flow and to establish the LSP for a traffic flow. IETF has defined two alternative protocols for this purpose. The Resource Reservation Protocol (RSVP) has been enhanced with TE extensions for carrying labels and building the LSP. The other approach is an enhancement to LDP known as Constraint-based Routing Label Distribution Protocol (CR-LDP). **Traffic forwarding:** This is accomplished with MPLS, using the LSP set up by the traffic engineering components just described.

RSVP – TE Operation

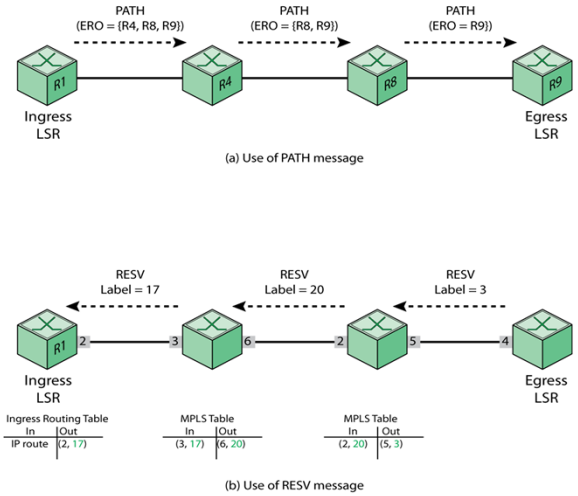


Figure 21.11 RSVP-TE Operation

43

Early in the MPLS standardization process, it became clear that a protocol was needed that would enable providers to set up LSPs that took into account QoS and traffic engineering parameters. Development of this type of signaling protocol proceeded on two different tracks: Extensions to RSVP for setting up MPLS tunnels, known as RSVP-TE [RFC3209] Extensions to LDP for setting constraint based LSPs [RFC3212]. The motivation for the choice of protocol in both cases was straightforward. Extending RSVP-TE to do in an MPLS environment what it already was doing (handling QoS information and reserving resources) in an IP environment is comprehensible; you only have to add the label distribution capability. Extending a native MPLS protocol like LDP, which was designed to do label distribution, to handle some extra TLVs with QoS information is also not revolutionary. Ultimately, the MPLS working group announced, in RFC 3468, that RSVP-TE is the preferred solution. In general terms, RSVP-TE operates by associating an MPLS label with an RSVP flow. RSVP is used to reserve resources and to define an explicit router for an LSP tunnel. Stallings DCC9e Figure 21.11 illustrates the basic operation of RSVP-TE. An ingress node uses the RSVP PATH message to request an LSP to be defined along an explicit route. The PATH message includes a label request object and an explicit route object (ERO). The ERO defines the explicit route to be followed by the LSP. The destination node of a label-switched path responds to a LABEL_REQUEST by including a LABEL object in its response RSVP Resv message. The LABEL object is inserted in the filter spec list immediately following the filter spec to which it pertains. The Resv message is sent back upstream towards the sender, following the path state created by the Path message, in reverse order.

Virtual Private Network (VPN)

- Private network configured within a public network in order to take advantage of management facilities of larger networks
- Traffic designated as VPN traffic can only go from a VPN source to a destination in the same VPN

Widely used by enterprises to:

- Create wide area networks (WANs)
- Provide site-to-site communications to branch offices
- Allow mobile user to dial up their company LANs

44

A virtual private network (VPN) is a private network that is configured within a public network (a carrier's network or the Internet) in order to take advantage of the economies of scale and management facilities of large networks. VPNs are widely used by enterprises to create wide area networks (WANs) that span large geographic areas, to provide site-to-site connections to branch offices, and to allow mobile users to dial up their company LANs. From the point of view of the provider, the pubic network facility is shared by many customers, with the traffic of each customer segregated from other traffic. Traffic designated as VPN traffic can only go from a VPN source to a destination in the same VPN. It is often the case that encryption and authentication facilities are provided for the VPN.

Layer 2 VPN

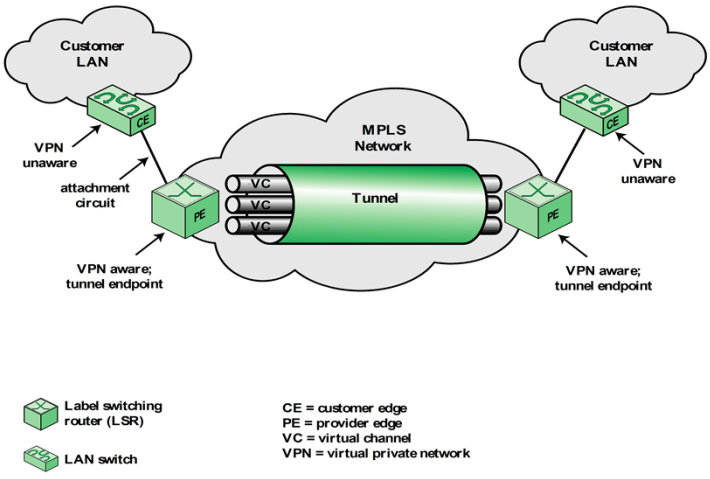
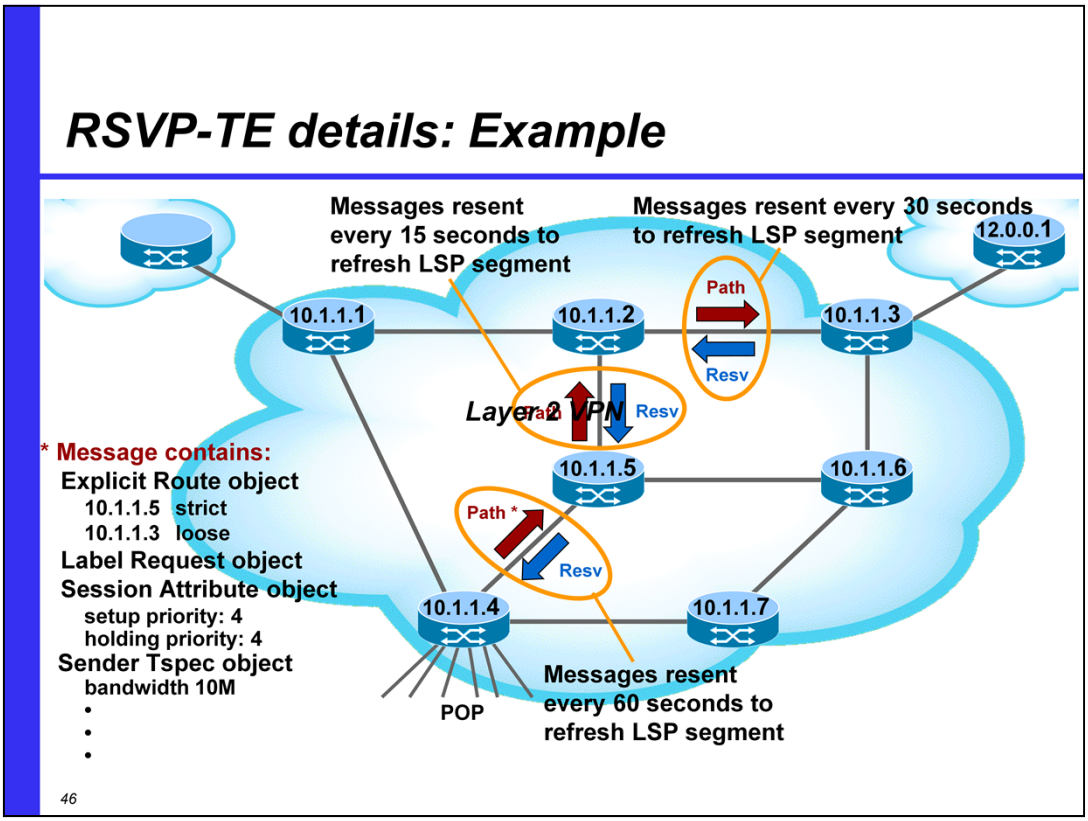


Figure 23.12 Layer 2 VPN Concepts

45

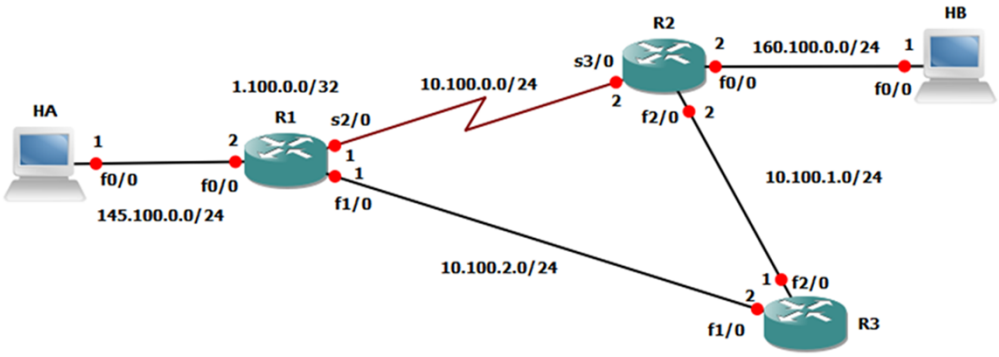
With a layer 2 VPN, there is mutual transparency between the customer network and the provider network. In effect, the customer requests a mesh of unicast LSPs among customer switches that attach to the provider network. Each LSP is viewed as a layer 2 circuit by the customer. In a L2VPN, the provider's equipment forwards customer data based on information in the Layer 2 headers, such as an Ethernet MAC address, an ATM virtual channel identifier, or a frame relay data link connection identifier.



Les rutes poden ser explícites, indicant exactament per on passa (estricte) i dinàmiques on el camí es busca per OSPF. Per establir una ruta s'indica la prioritat i un cop feta s'indica la prioritat de manteniment. En l'establiment s'indica el Bw demanat. Les prioritats tenen un valor contrari al valor del número.

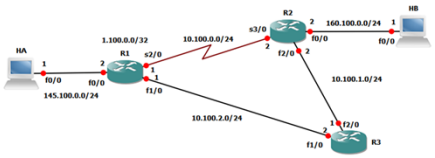
Real example MPLS

Font: Elaboració pròpia



47

Programació Router



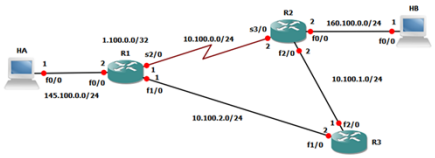
R1

```
ip cef
mpls label protocol ldp
interface Loopback0
 ip address 1.100.1.1 255.255.255.255
interface FastEthernet0/0
 ip address 145.100.0.2 255.255.255.0
interface FastEthernet1/0
 ip address 10.100.2.1 255.255.255.0
mpls ip
interface Serial2/0
 ip address 10.100.0.1 255.255.255.0
 mpls ip
router ospf 1
passive-interface FastEthernet0/0
 network 1.100.1.1 0.0.0.0 area 0
 network 10.100.0.0 0.0.0.255 area 0
 network 10.100.2.0 0.0.0.255 area 0
 network 145.100.0.0 0.0.0.255 area 0
```

R2

```
ip cef
mpls label protocol ldp
interface Loopback0
 ip address 1.100.1.2 255.255.255.255
interface FastEthernet0/0
 ip address 160.100.0.2 255.255.255.0
interface FastEthernet2/0
 ip address 10.100.1.2 255.255.255.0
mpls ip
interface Serial3/0
 ip address 10.100.0.2 255.255.255.0
 mpls ip
router ospf 1
 passive-interface FastEthernet0/0
 network 1.100.1.2 0.0.0.0 area 0
 network 10.100.0.0 0.0.0.255 area 0
 network 10.100.1.0 0.0.0.255 area 0
 network 160.100.0.0 0.0.0.255 area 0
```

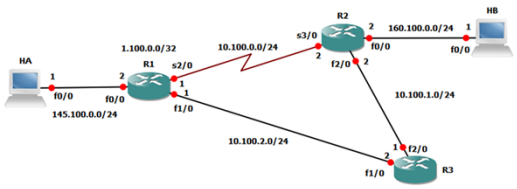
Routing table MPLS



R1#show ip route

1.0.0.0/32 is subnetted, 3 subnets
C 1.100.1.1 is directly connected, Loopback0
O 1.100.1.2 [110/3] via 10.100.2.2, 00:00:59, FastEthernet1/0
O 1.100.1.3 [110/2] via 10.100.2.2, 00:00:59, FastEthernet1/0
145.100.0.0/24 is subnetted, 1 subnets
C 145.100.0.0 is directly connected, FastEthernet0/0
160.100.0.0/24 is subnetted, 1 subnets
O 160.100.0.0 [110/3] via 10.100.2.2, 00:00:59, FastEthernet1/0
10.0.0.0/24 is subnetted, 3 subnets
C 10.100.2.0 is directly connected, FastEthernet1/0
C 10.100.0.0 is directly connected, Serial2/0
O 10.100.1.0 [110/2] via 10.100.2.2, 00:00:59, FastEthernet1/0

Label table MPLS

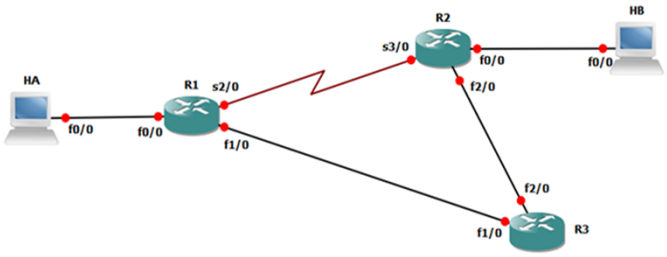


R1#show mpls forwarding-table

Local tag	Outgoing tag or VC	Prefix or Tunnel Id	Bytes tag switched	Outgoing interface	Next Hop
16	17	1.100.1.2/32	0	Fa1/0	10.100.2.2
17	Pop tag	10.100.1.0/24	0	Fa1/0	10.100.2.2
18	20	160.100.0.0/24	0	Fa1/0	10.100.2.2
19	Pop tag	1.100.1.3/32	0	Fa1/0	10.100.2.2

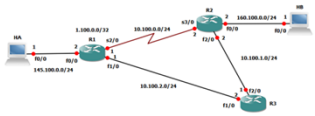
MPLS TE example

Tunel1 entre R1 y R2 a 50 Kbps dynamic. esto quiere decir que pasará por donde le diga OSPf. Seguramente R1-R3-R2. Si quiero que pase por R1-R2 deberé ponerlo explícito.
Tunel 2 entre R2-R3-R1 a 100 Kbps explícito por lo que pasará siempre R2-R3-R1
Se puede ver con "show ip route" en cada router



51

Programació Routers

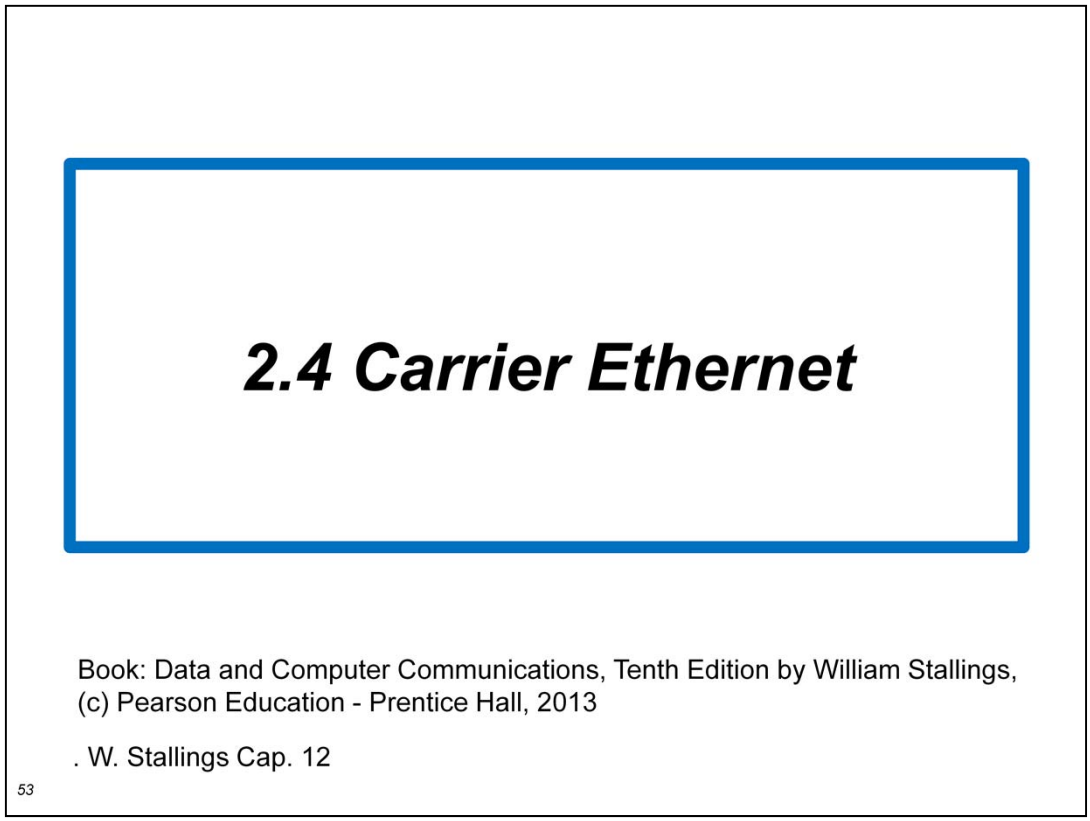


R1

```
ip cef
mpls label protocol ldp
mpls traffic-eng tunnels
interface Loopback0
 ip address 1.100.1.1 255.255.255.255
interface Tunnel1
 ip unnumbered Loopback0
 tunnel destination 1.100.1.2
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng autoroute
announce
  tunnel mpls traffic-eng bandwidth  50
 tunnel mpls traffic-eng path-option 1
dynamic
interface FastEthernet0/0
 ip address 145.100.0.2 255.255.255.0
interface FastEthernet1/0
 ip address 10.100.2.1 255.255.255.0
 mpls ip
 mpls traffic-eng tunnels
 ip rsvp bandwidth 1000
interface Serial2/0
 ip address 10.100.0.1 255.255.255.0
 mpls ip
 mpls traffic-eng tunnels
 clock rate 56000
 ip rsvp bandwidth 100
router ospf 1
 mpls traffic-eng router-id Loopback0
 mpls traffic-eng area 0
passive-interface FastEthernet0/0
 network 1.100.1.1 0.0.0.0 area 0
 network 10.100.0.0 0.0.0.255 area 0
 network 10.100.2.0 0.0.0.255 area 0
 network 145.100.0.0 0.0.0.255 area 0
```

52

T



Aprofitem la definició de la trama ethernet i el seu sincronisme per transmetre dades a nivell 2 fora de les LANs. Necesitarem infrastructura WAN amb commutació Ethernet a nivell 2.

No fa servir canals de 64 Kbps. Pot tenir sistema de transmissió físic a nivell 1 propi (Eth), però com a paquet també es pot utilitzar SDH (igual que ATM).

Ethernet MAC Frame format

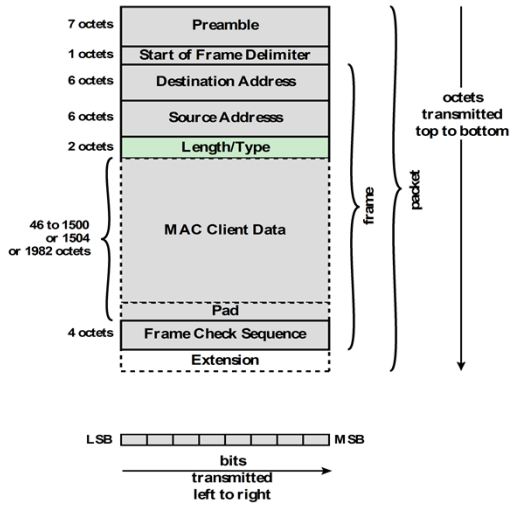


Figure 12.4 IEEE 802.3 MAC Frame Format

54

IEEE 802.3 defines three types of MAC frames. The basic frame is the original frame format. In addition, to support data link layer protocol encapsulation within the data portion of the frame, two additional frame types have

been added. A Q-tagged frame supports 802.1Q VLAN capability, as described in Section 12.3. An envelope frame is intended to allow inclusion of additional prefixes and suffixes to the data field required by higher-layer encapsulation protocols such as those defined by the IEEE 802.1 working group (such as Provider Bridges and MAC Security), ITU-T, or IETF (such as MPLS). Figure 12.4 depicts the frame format for all three types of frames; the differences are contained in the MAC Client Data field. Several additional fields encapsulate the frame to form an 802.3 packet.

Full Duplex Operation

- Traditional Ethernet half duplex
- Using full-duplex, station can transmit and receive simultaneously
- 100-Mbps Ethernet in full-duplex mode, giving a theoretical transfer rate of 200 Mbps
- Stations must have full-duplex adapter cards
- And must use switching hub
  - Each station constitutes separate collision domain
  - CSMA/CD algorithm no longer needed
  - 802.3 MAC frame format used

55

A traditional Ethernet is half duplex: a station can either transmit or receive a frame, but it cannot do both simultaneously. With full-duplex operation, a station can transmit and receive simultaneously. If a 100-Mbps Ethernet ran in full-duplex mode, the theoretical transfer rate becomes 200 Mbps. Several changes are needed to operate in full-duplex mode. The attached stations must have full-duplex rather than half-duplex adapter cards. The central point in the star wire cannot be a simple multiport repeater but rather must be a switching hub. In this case each station constitutes a separate collision domain. In fact, there are no collisions and the CSMA/CD algorithm is no longer needed. However, the same 802.3 MAC frame format is used and the attached stations can continue to execute the CSMA/CD algorithm, even though no collisions can ever be detected.

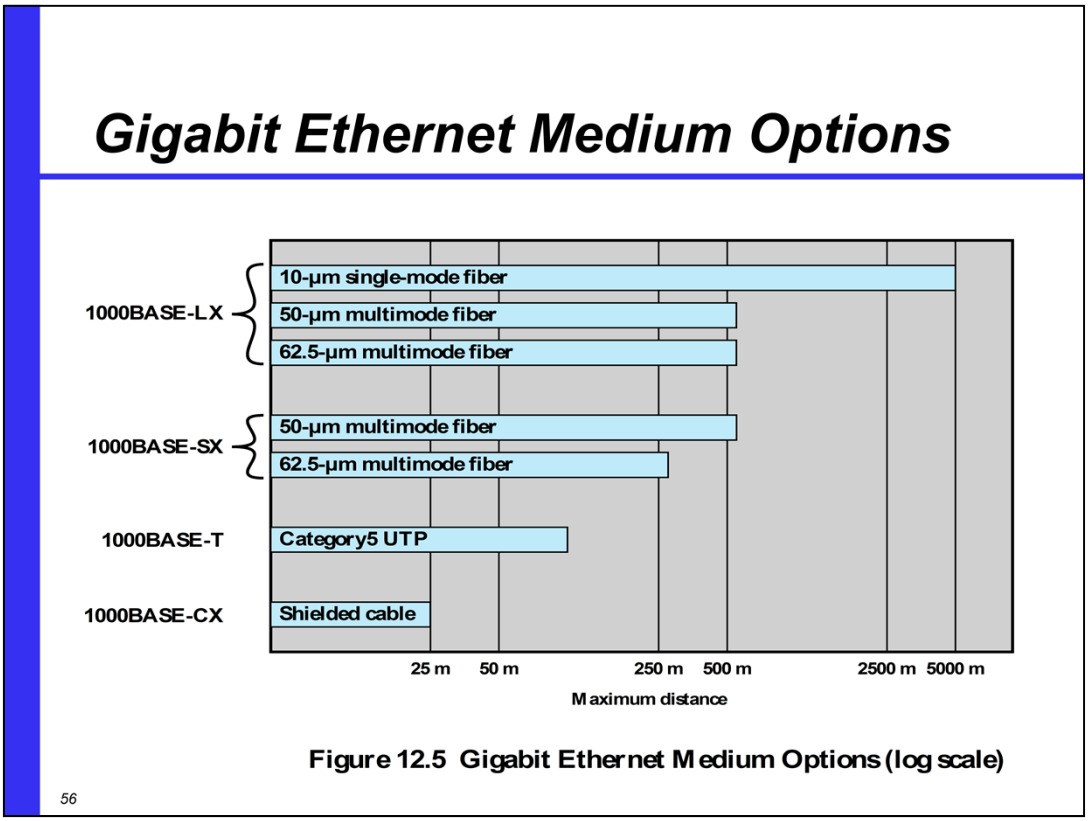


Figure 12.5 Gigabit Ethernet Medium Options (log scale)

The current 1-Gbps specification for IEEE 802.3 includes the following physical layer alternatives (Figure 12.5):

10Gbps Ethernet

- Growing interest in 10Gbps Ethernet
  - High-speed backbone use
  - Future wider deployment
- Alternative to ATM and other WAN technologies
- Uniform technology for LAN, MAN, or WAN
- Advantages of 10Gbps Ethernet
  - No expensive, bandwidth-consuming conversion between Ethernet packets and ATM cells
  - IP and Ethernet together offers QoS and traffic policing approach ATM
  - Have a variety of standard optical interfaces

57

With gigabit products still fairly new, attention has turned in the past several years to a 10-Gbps Ethernet capability. The principle driving requirement for 10 Gigabit Ethernet is the increase in Internet and intranet traffic.. The technology also allows the construction of metropolitan area networks (MANs) and WANs that connect geographically dispersed LANs between campuses or points of presence (PoPs). Thus, Ethernet begins to compete with ATM and other wide area transmission and networking technologies. In most cases where the customer requirement is data and TCP/IP transport, 10-Gbps Ethernet provides substantial value over ATM transport for both network end users and service providers: The combination of IP and Ethernet offers quality of service and traffic policing capabilities that approach those provided by ATM, so that advanced traffic engineering technologies are available to users and providers. A wide variety of standard optical interfaces (wavelengths and link distances) have been specified for 10-Gbps Ethernet, optimizing its operation and cost for LAN, MAN, or WAN applications.

10Gbps Ethernet exemple

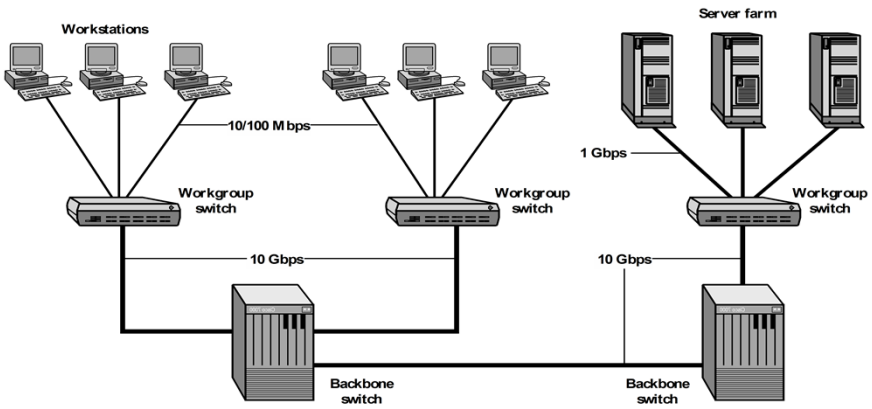


Figure 12.6 Example 10 Gigabit Ethernet Configuration

58

Figure 12.6 illustrates potential uses of 10-Gbps Ethernet. Higher-capacity backbone pipes will help relieve congestion for workgroup switches, where Gigabit Ethernet uplinks can easily become overloaded, and for server farms, where 1-Gbps network interface cards are already in widespread use. The goal for maximum link distances cover a range of applications: from 300 m to 40 km. The links operate in full-duplex mode only, using a variety of optical fiber physical media.

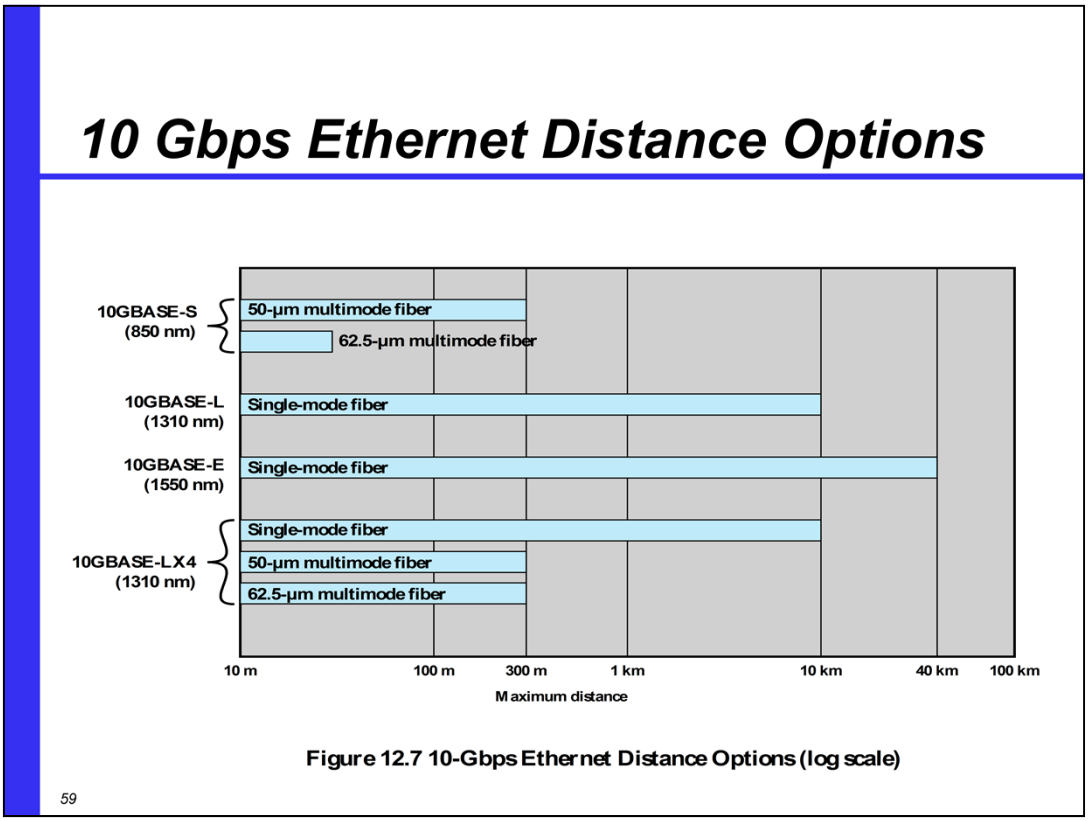


Figure 12.7 10-Gbps Ethernet Distance Options (log scale)

Four physical layer options are defined for 10-Gbps Ethernet (Figure 12.7). The first three of these have two suboptions: an "R" suboption and a "W" suboption. The R designation refers to a family of physical layer implementations that use a signal encoding technique known as 64B/66B, described in Appendix 12A. The R implementations are designed for use over *dark fiber*, meaning a fiber optic cable that is not in use and that is not connected to any other equipment. The W designation refers to a family of physical layer implementations that also use 64B/66B signaling but that are then encapsulated to connect to SONET equipment.

.

100-Gbps Ethernet

- Preferred technology for wired LAN
- Preferred carrier for bridging wireless technologies into local Ethernet networks
- Cost-effective, reliable and interoperable
- Popularity of Ethernet technology:
  - Availability of cost-effective products
  - Reliable and interoperable network products
  - Variety of vendors

60

Ethernet is widely deployed and is the preferred technology for wired local area networking. Ethernet dominates enterprise LANs, broadband access, data center networking, and has also become popular for communication across metropolitan and even wide area networks. Further, it is now the preferred carrier wire line vehicle for bridging wireless technologies, such as WiFi and WiMAX, into local Ethernet networks.

This popularity of Ethernet technology is due to the availability of cost-effective, reliable, and interoperable networking products from a variety of vendors. The development of converged and unified communications, the evolution of massive server farms, and the continuing expansion of VoIP, TVoIP, and Web 2.0 applications have driven the need for ever faster Ethernet switches. The following are market drivers for 100-Gbps Ethernet:

Data center/Internet media providers: To support the growth of Internet multimedia content and Web applications, content providers have been expanding data centers, pushing 10-Gbps Ethernet to its limits. Likely to be high-volume early adopters of 100-Gbps Ethernet.

Metro-Video/Service Providers: Video on demand has been driving a new generation of 10-Gbps Ethernet metropolitan/core network buildouts. Likely to be high-volume adopters in the medium term.

Enterprise LANs: Continuing growth in convergence of voice/video/data and in unified communications is driving up network switch demands. However, most enterprises still rely on 1-Gbps or a mix of 1-Gbps and 10-Gbps Ethernet, and adoption of 100-Gbps Ethernet is likely to be slow.

Internet exchanges/ISP (Internet Service Provider) core routing: With the massive amount of traffic flowing through these nodes, these installations are likely to be early adopters of 100-Gbps Ethernet.

Example 100 Gbps Ethernet

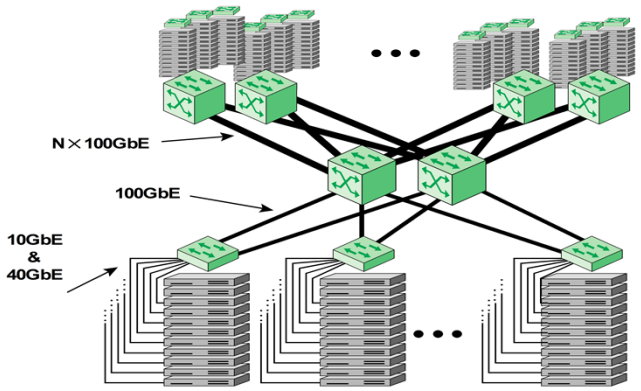


Figure 12.8 Example 100-Gbps Ethernet Configuration for Massive Blade Server Site

61

An example of the application of 100-Gbps Ethernet is shown in Figure 12.8, taken from [NOWE07]. The trend at large data centers, with substantial banks of blade servers, is the deployment of 10-Gbps ports on individual servers to handle the massive multimedia traffic provided by these servers. Such arrangements are stressing the on-site switches needed to interconnect large numbers of servers. A 100GbE rate was proposed to provide the bandwidth required to handle the increased traffic load. It is expected that 100GbE will be deployed in switch uplinks inside the data center as well as providing interbuilding, intercampus, MAN, and WAN connections for enterprise networks. The success of Fast Ethernet, Gigabit Ethernet, and 10-Gbps Ethernet highlights the importance of network management concerns in choosing a network technology. The 40-Gbps and 100-Gbps Ethernet specifications offer compatibility with existing installed LANs, network management software, and applications. This compatibility has accounted for the survival of 30-year-old technology in today's fast-evolving network environment.

Media Options for 40-Gbps and 100-Gbps Ethernet

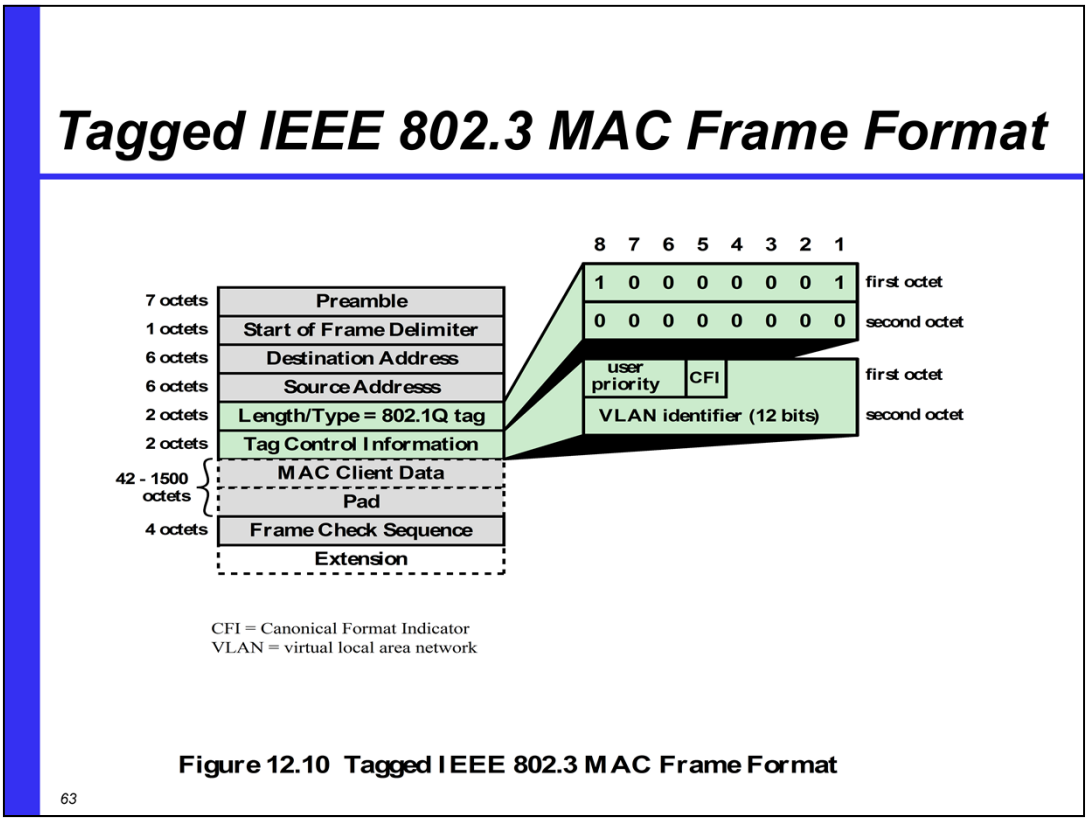
	40 Gbps	100 Gbps
1m backplane	40GBASE-KR4	
10 m copper	40GBASE-CR4	1000GBASE-CR10
100 m multimode fiber	40GBASE-SR4	1000GBASE-SR10
10 km single mode fiber	40GBASE-LR4	1000GBASE-LR4
40 km single mode fiber		1000GBASE-ER4

Naming nomenclature:

Copper: K = backplane; C = cable assembly
Optical: S = short reach (100m); L - long reach (10 km); E = extended long reach (40 km)
Coding scheme: R = 64B/66B block coding
Final number: number of lanes (copper wires or fiber wavelengths)

62

IEEE 802.3ba specifies three types of transmission media (Table 12.3): copper backplane, twisted pair, and optical fiber. For copper media, four separate physical lanes are specified. For optical fiber, either 4 or 10 wavelength lanes are specified, depending on data rate and distance.



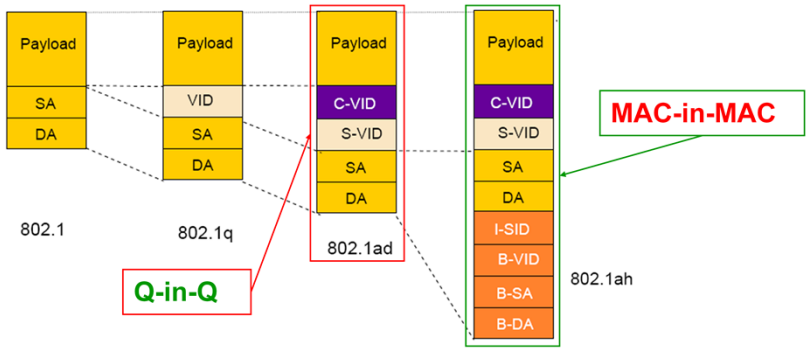
The IEEE 802.1Q standard, last updated in 2005, defines the operation of VLAN bridges and switches that permits the definition, operation and administration of VLAN topologies within a bridged/switched LAN infrastructure. In this section, we will concentrate on the application of this standard to 802.3 LANs. Recall from Chapter 11 that a VLAN is an administratively configured broadcast domain, consisting of a subset of end stations attached to a LAN. A VLAN is not limited to one switch but can span multiple interconnected switches. In that case traffic between switches must indicate VLAN membership. This is accomplished in 802.1Q by inserting a tag with a VLAN identifier (VID) with a value in the range from 1 to 4094. Each VLAN in a LAN configuration is assigned a globally unique VID. By assigning the same VID to end systems on many switches, one or more VLAN broadcast domains can be extended across a large network. Figure 12.10 shows the position and content of the 802.1 tag, referred to as Tag Control Information (TCI). The presence of the 2-octet TCI field is indicated by set the Length/Type field in the 802.3 MAC from to a value of 8100 hex. The TCI consists of three subfields:

User priority (3 bits): The priority level for this frame. **Canonical format indicator (1 bit):** s always set to zero for Ethernet switches. CFI is used for compatibility reason between Ethernet type network and Token Ring type network. If a frame received at an Ethernet port has a CFI set to 1, then that frame should not be forwarded as it is to an untagged port.**VLAN identifier (12 bits):** the identification of the VLAN. Of the 4096 possible VIDs, a VID of 0 is used to identify that the TCI contains only a priority value, and 4095 (FFF) is reserved, so the maximum possible number of VLAN configurations is 4094.

T

Provider Backbone Bridge Traffic Engineering (PBB-TE)

- **IEEE has developed a number of standards providing enhancements to the original Ethernet standards. PBB-TE adapts Ethernet technology to carrier class transport networks**
  - **802.1Q**: Virtual LAN
  - **802.1ad**: Provider Bridging
  - **802.1ah**: Provider Backbone Bridging
  - **802.1ag**: Connectivity Fault Management (OAM)

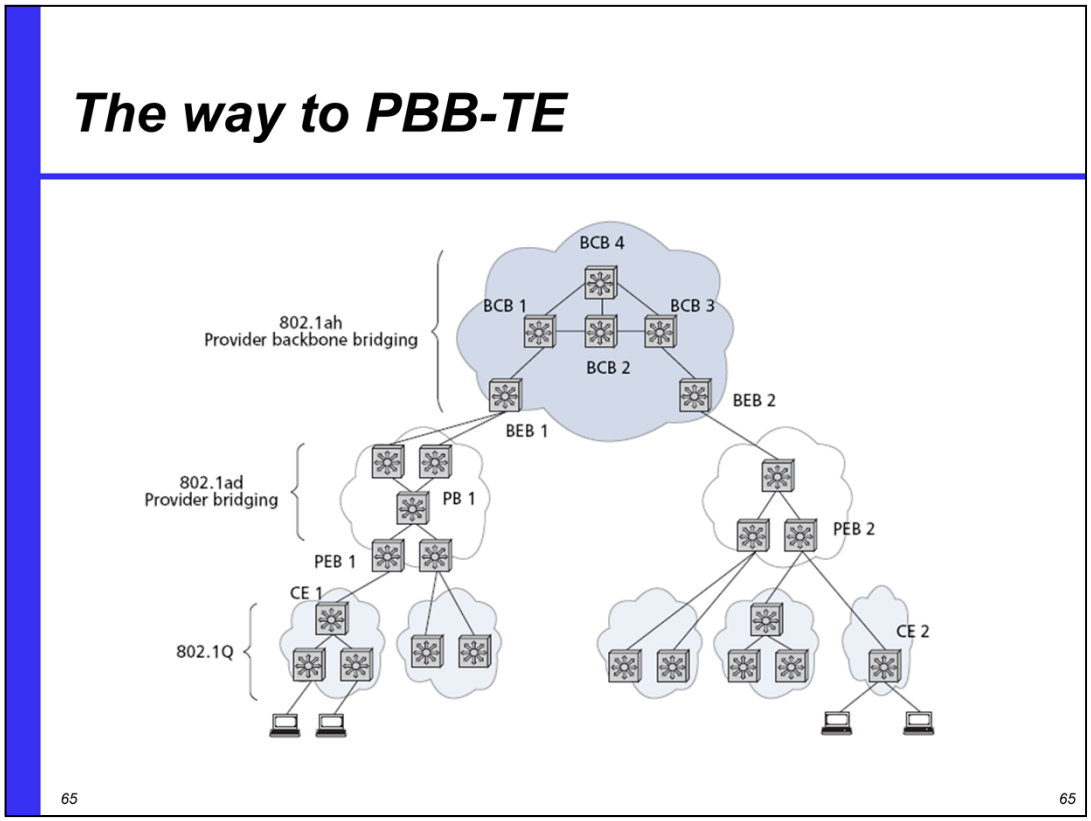


64 64

QinQ permet ampliar el nombre de circuits virtuals i crear aniuaments (un VLAN dins un altra).

Mac in Mac permet crear nivells de commutació a nivells diferents (com la xarxa telefònica de veu). Es tracta d'un tunneling.

T



Mac in Mac permet crear nivells jeràrquics de commutació. Això reduexi la complicació de les taules d'enrutament.

T

Tècnica aplicable a qualsevol tecnologia de nivell 2 (o 3).

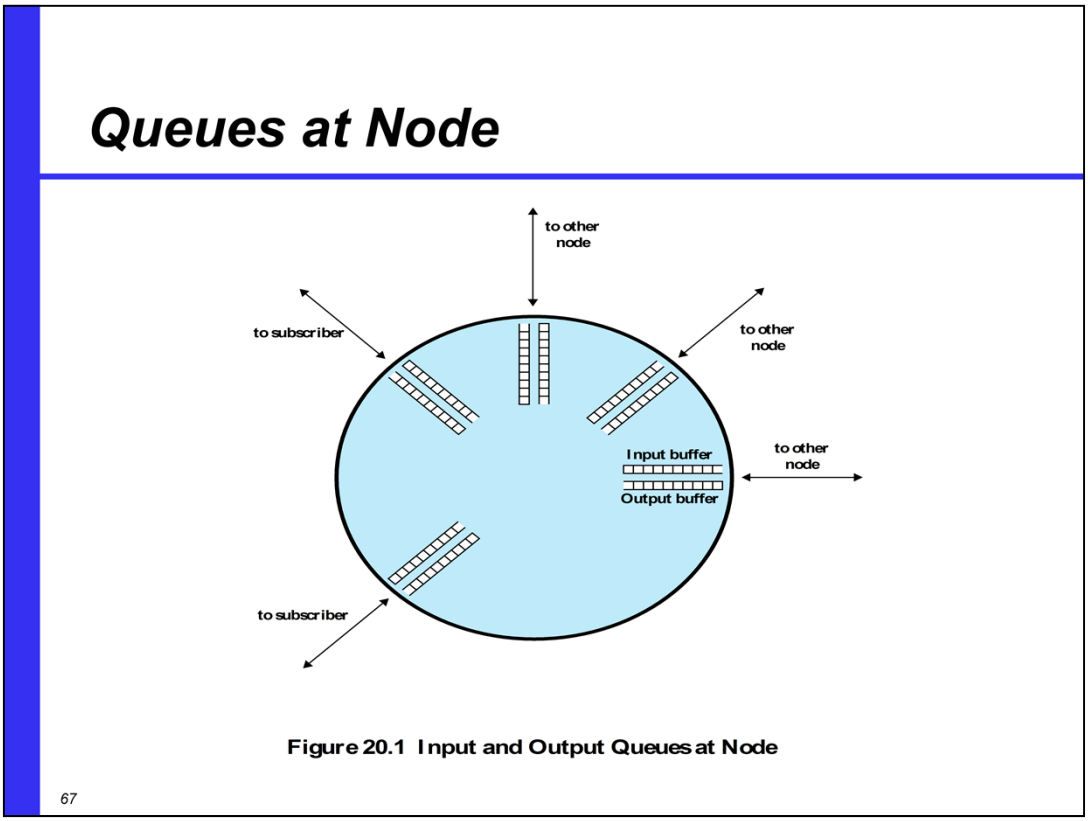


Figure 20.1 Input and Output Queues at Node

Consider the queuing situation at a single packet switch or router, such as is illustrated in Figure 20.1. Any given node has a number of I/O ports attached to it: one or more to other nodes, and zero or more to end systems. On each port, packets arrive and depart. We can consider that there are two buffers, or queues, at each port, one to accept arriving packets, and one to hold packets that are waiting to depart. In practice, there might be two fixed-size buffers associated with each port, or there might be a pool of memory available for all buffering activities. In the latter case, we can think of each port having two variable-size buffers associated with it, subject to the constraint that the sum of all buffer sizes is a constant. In any case, as packets arrive, they are stored in the input buffer of the corresponding port. The node examines each incoming packet, makes a routing decision, and then moves the packet to the appropriate output buffer. Packets queued for output are transmitted as rapidly as possible; this is, in effect, statistical time division multiplexing. If packets arrive too fast for the node to process them (make routing decisions) or faster than packets can be cleared from the outgoing buffers, then eventually packets will arrive for which no memory is available.

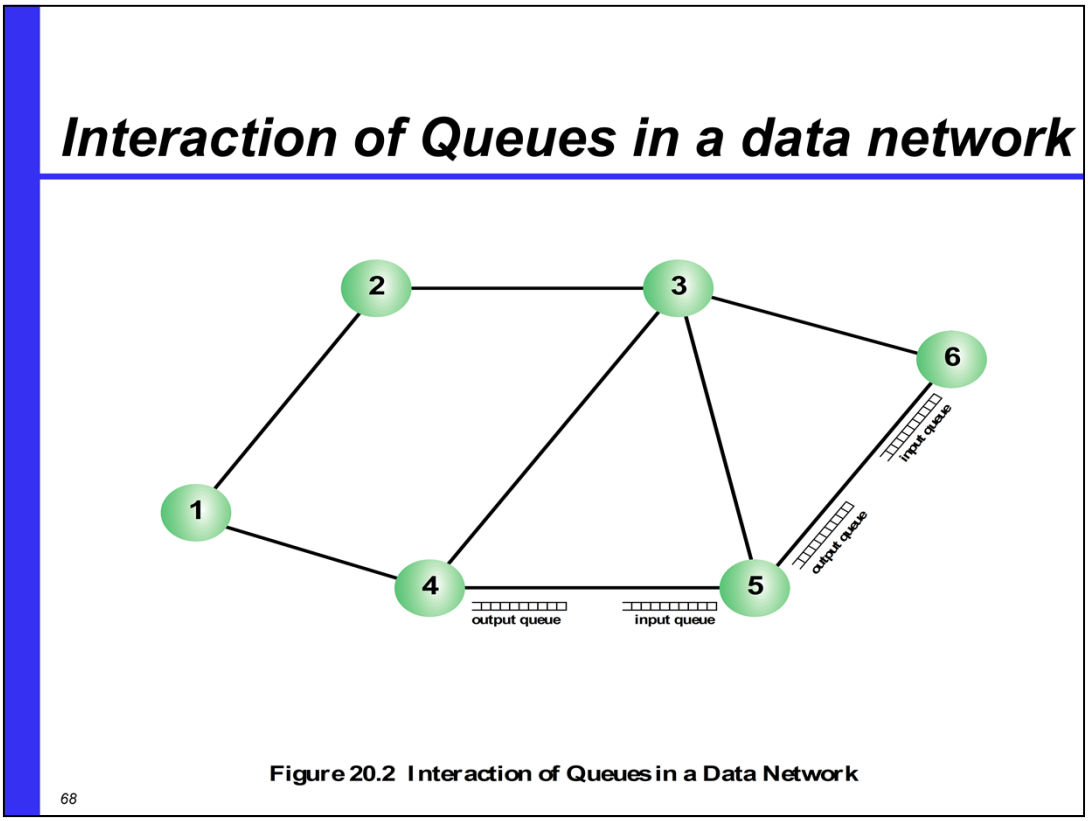
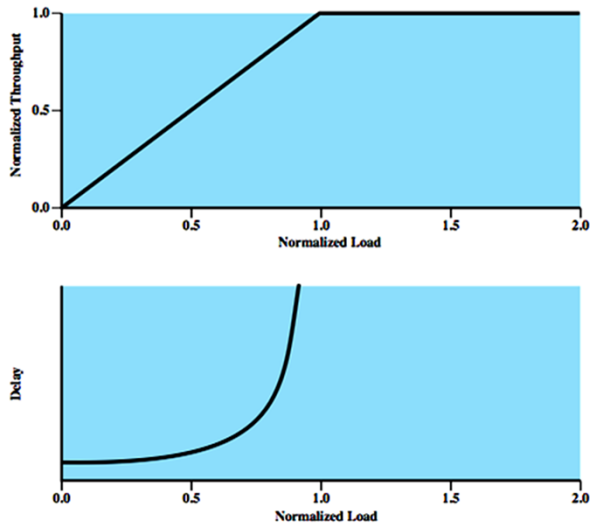


Figure 20.2 Interaction of Queues in a Data Network

When such a saturation point is reached, one of two general strategies can be adopted. The first such strategy is to discard any incoming packet for which there is no available buffer space. The alternative is for the node that is experiencing these problems to exercise some sort of flow control over its neighbors so that the traffic flow remains manageable. But, as Figure 20.2 illustrates, each of a node's neighbors is also managing a number of queues. If node 6 restrains the flow of packets from node 5, this causes the output buffer in node 5 for the port to node 6 to fill up. Thus, congestion at one point in the network can quickly propagate throughout a region or the entire network. While flow control is indeed a powerful tool, we need to use it in such a way as to manage the traffic on the entire network.

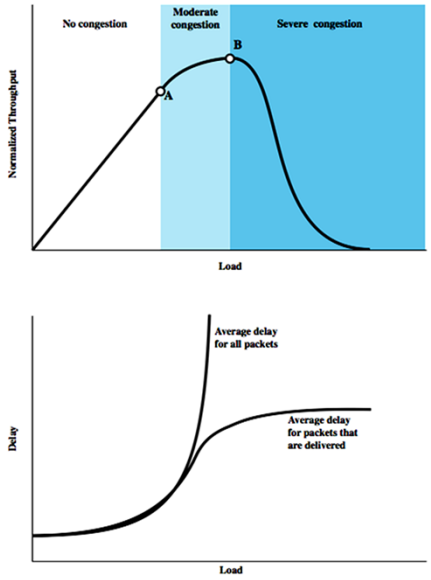
Ideal Network Utilization



69

Figure suggests the ideal goal for network utilization. The top graph plots the steady-state total throughput (number of packets delivered to destination end systems) through the network as a function of the offered load (number of packets transmitted by source end systems), both normalized to the maximum theoretical throughput of the network. For example, if a network consists of a single node with two full-duplex 1-Mbps links, then the theoretical capacity of the network is 2 Mbps, consisting of a 1-Mbps flow in each direction. In the ideal case, the throughput of the network increases to accommodate load up to an offered load equal to the full capacity of the network; then normalized throughput remains at 1.0 at higher input loads. Note, however, what happens to the end-to-end delay experienced by the average packet even with this assumption of ideal performance. At negligible load, there is some small constant amount of delay that consists of the propagation delay through the network from source to destination plus processing delay at each node. As the load on the network increases, queuing delays at each node are added to this fixed amount of delay. When the load exceeds the network capacity, delays increase without bound.

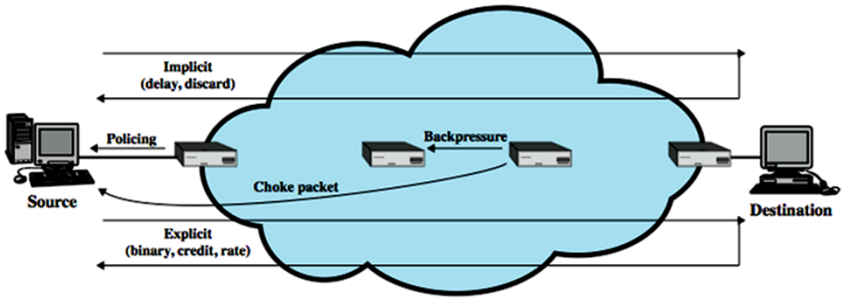
Effects of Congestion - No Control



70

The ideal case reflected Figure assumes infinite buffers and no overhead related to congestion control. In practice, buffers are finite, leading to buffer overflow, and attempts to control congestion consume network capacity in the exchange of control signals. Let us consider what happens in a network with finite buffers if no attempt is made to control congestion or to restrain input from end systems. The details will, of course, differ depending on network configuration and on the statistics of the presented traffic.

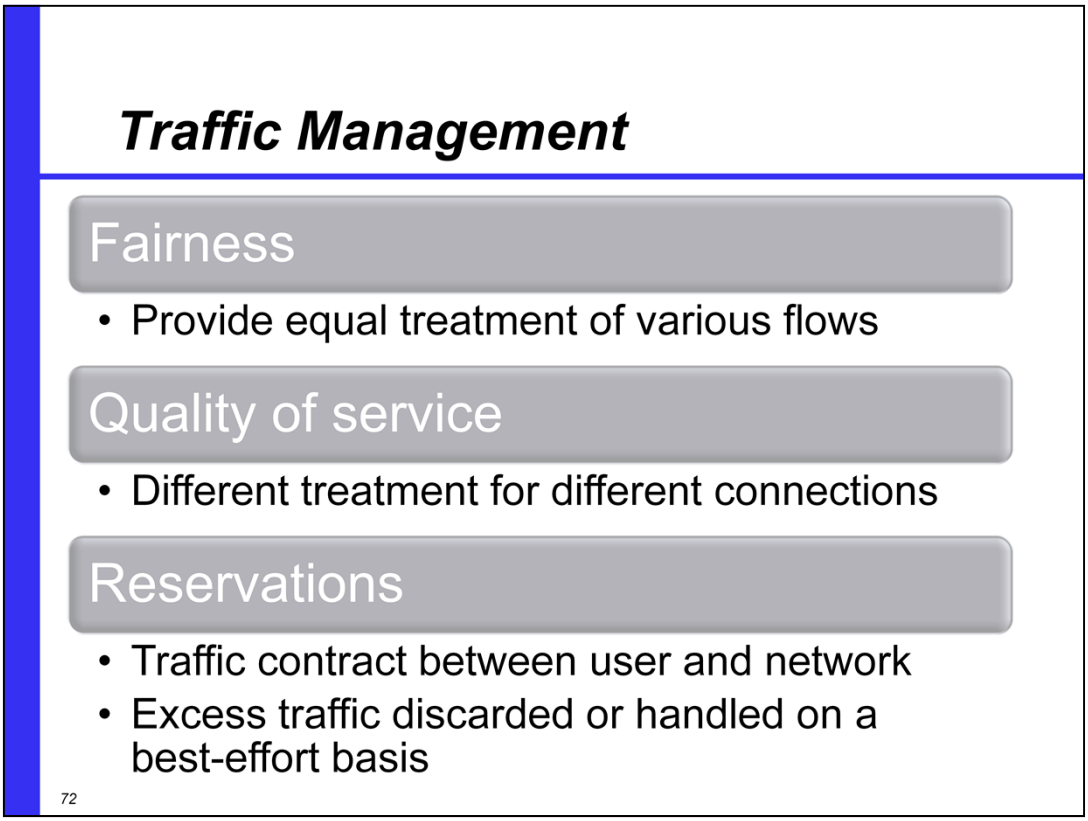
Mechanisms for Congestion Control



71

In this book, we discuss various techniques for controlling congestion in packet-switching, frame relay, and ATM networks, and in IP-based internets. To give context to this discussion, Stallings DCC9e Figure 13.5 provides a general depiction of important congestion control techniques, which include:

- backpressure
- choke packets
- implicit congestion signaling
- explicit congestion signaling



There are a number of issues related to congestion control that might be included under the general category of traffic management. In its simplest form, congestion control is concerned with efficient use of a network at high load. The various mechanisms discussed in the previous section can be applied as the situation arises, without regard to the particular source or destination affected. When a node is saturated and must discard packets, it can apply some simple rule, such as discard the most recent arrival. However, other considerations can be used to refine the application of congestion control techniques and discard policy.

Traffic Shaping/Traffic Policing

- Two important tools in network management:
  - Traffic shaping
    - *Concerned with traffic leaving the switch*
    - *Reduces packet clumping*
    - *Produces an output packet stream that is less burst and with a more regular flow of packets*
  - Traffic policing
    - *Concerned with traffic entering the switch*
    - *Packets that don't conform may be treated in one of the following ways:*
      - *Give the packet lower priority compared to packets in other output queues*
      - *Label the packet as nonconforming by setting the appropriate bits in a header*
      - *Discard the packet*

Two important tools in managing network are traffic shaping and traffic policing. Traffic shaping is aimed at smoothing out traffic flow by reducing packet clumping that leads to fluctuations in buffer occupancy. In essence, if the input to a switch on a certain channel or logical connection or flow is bursty, traffic shaping produces an output packet stream that is less bursty and with a more regular flow of packets.

Traffic policing discriminates between incoming packets that conform to quality of service (QoS) agreement and those that don't. Packets that don't conform may be treated in one of the following ways:

1. Give the packet lower priority compared to packets in other output queues. 2. Label the packet as nonconforming by setting the appropriate bits in a header. Downstream switches may treat nonconforming packets less favorably if congestion occurs. 3. Discard the packet. In essence, traffic shaping is concerned with traffic leaving the switch and traffic policing is concerned with traffic entering the switch. Two important techniques that can be used for traffic shaping or traffic policing are token bucket and leaky bucket.

Token Bucket / Leaky Bucket

- Widely used traffic management tool
- Advantages:
  - Many traffic sources can be defined easily and accurately
  - Provides a concise description of the load to be imposed by a flow, enabling the service to determine easily the resource requirement
  - Provides the input parameters to a policing function

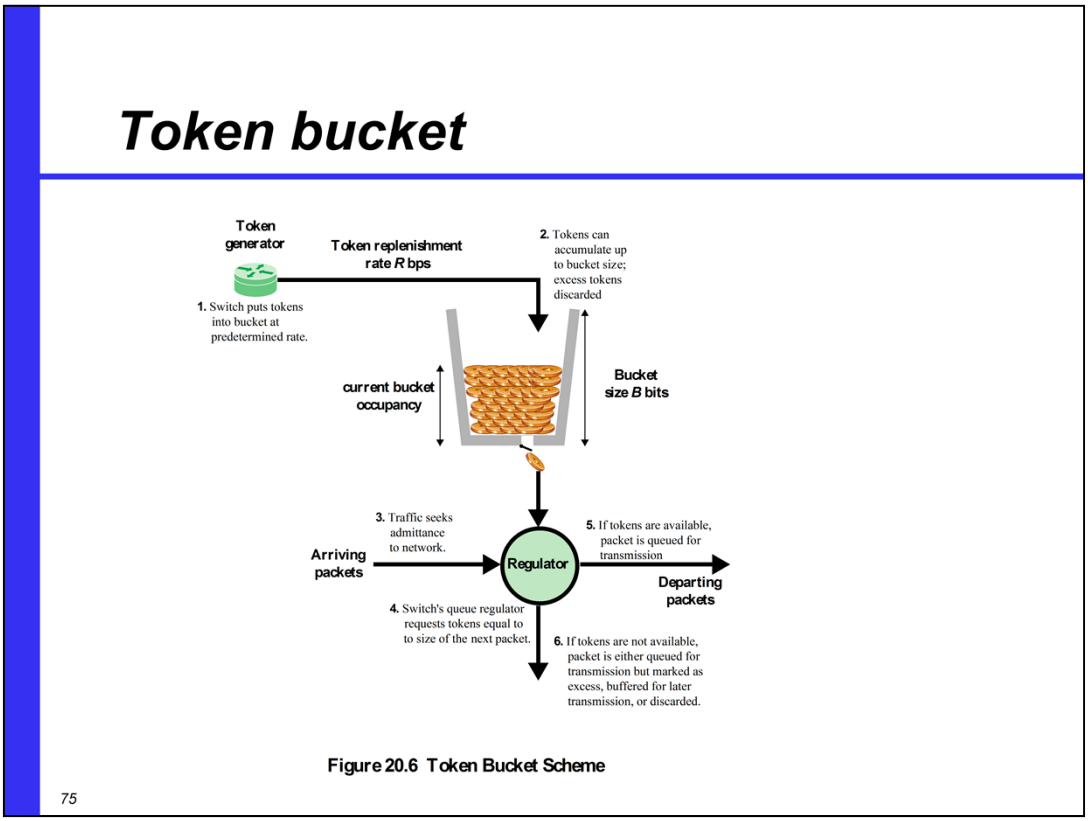
74

A widely used traffic management tool is token bucket. This is a way of characterizing and managing traffic that has three advantages:

1. Many traffic sources can be defined easily and accurately by a token bucket scheme.

2. The token bucket scheme provides a concise description of the load to be imposed by a flow, enabling the service to determine easily the resource requirement.

3. The token bucket scheme provides the input parameters to a policing function. This scheme provides a concise description of the peak and average traffic load the recipient can expect and it also provides a convenient mechanism by which the sender can implement a traffic flow policy. Token bucket is used in the Bluetooth specification and in differentiated services.



A token bucket traffic specification consists of two parameters: a token replenishment rate *R* and a bucket size *B*. The token rate *R* specifies the continually sustainable data rate; that is, over a relatively long period of time, the

average data rate to be supported for this flow is *R*. The bucket size *B* specifies the amount by which the data rate can exceed *R* for short periods of time. The exact condition is as follows: during any time period *T*, the amount of data sent cannot exceed $RT + B$. Figure illustrates this scheme and explains the use of the term *bucket*. The bucket represents a counter that indicates the allowable number of bytes of data that can be sent at any time. The bucket fills with byte tokens at the rate of *R* (i.e., the counter is incremented R times per second), up to the bucket capacity (up to the maximum counter value). Data arrive from the user and are assembled into packets, which are queued for transmission. A packet may be transmitted if there are sufficient tokens to match the packet size.

Leaky bucket

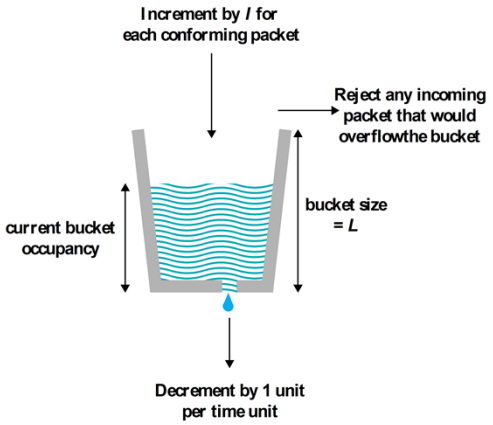


Figure 20.7 Leaky Bucket Algorithm

76

Another scheme, similar to token bucket, is leaky bucket. Leaky bucket is used in the asynchronous transfer mode (ATM) specification and in the ITU-T H.261 standard for digital video coding and transmission. The basic principle

of leaky bucket is depicted in Figure 20.7. The algorithm maintains a running count of the cumulative amount of data sent in a counter X . The counter is decremented at a constant rate of one unit per time unit to a minimum value of zero; this is equivalent to a bucket that leaks at a rate of 1. The counter is incremented by I for each arriving packet, where I is the size of the packet, subject to the restriction that the maximum counter value is L . Any arriving cell that would cause the counter to exceed its maximum is defined as nonconforming; this is equivalent to a bucket with a capacity of L .

T

Example: FR congestion control

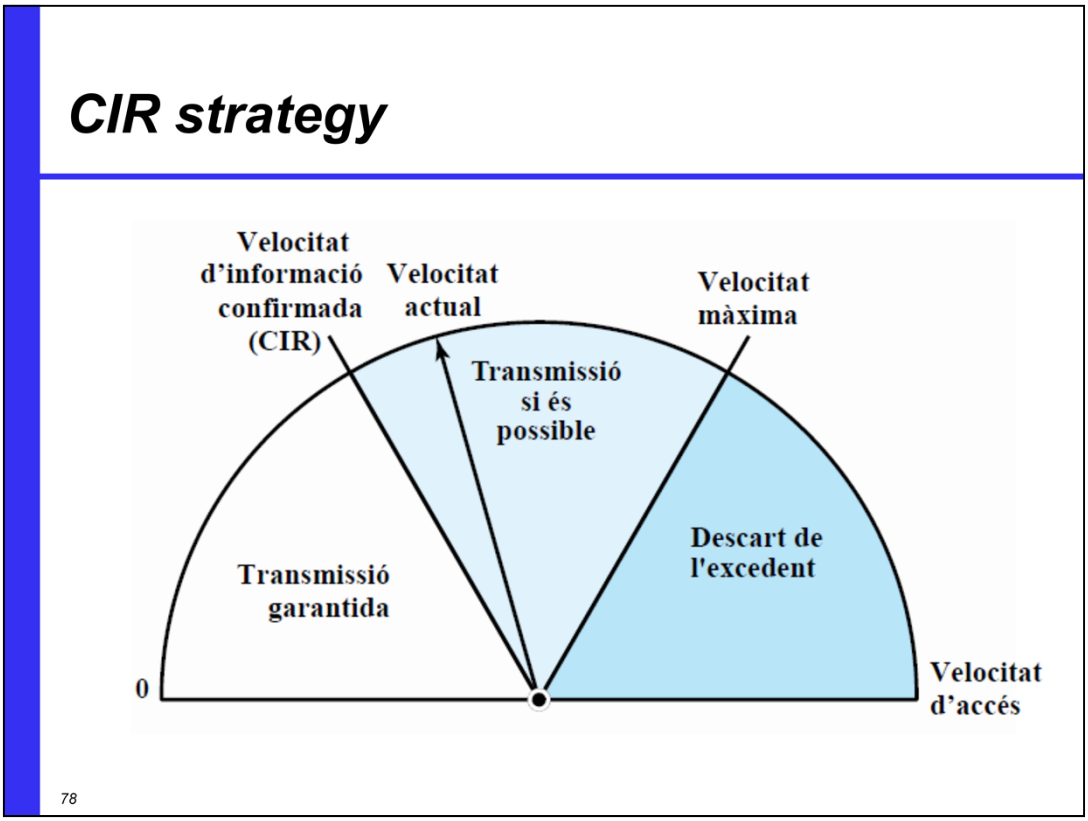
Tècnica	Tipus	Funció	Elements clau
Control de descart	Estratègia de descart	Proporciona una guia per a la xarxa sobre les trames que cal descartar	Bit DE
Notificació de la congestió explícita cap enrere	Elusió de la congestió	Proporciona una guia per als sistemes finals sobre la congestió a la xarxa	Bit BECN o missatge CLLM
Notificació de la congestió explícita cap endavant	Elusió de la congestió	Proporciona una guia per als sistemes finals sobre la congestió a la xarxa	Bit FECN
Notificació de la congestió implícita	Recuperació de la congestió	El sistema final infereix la congestió de la pèrdua de trames	Números de seqüència a PDU de capa superior

77

D'aquestes tècniques de control d ela congestió a FR ens quedem amb la tècnica de Discard Strategy amb el bit de descartar (equival a prioritat).

Si es viola alguna norma d'entrada (en base a algoritmes) la unitat de dades, sigui trama o paquet, es marca amb prioritat baixa. En cas de congestió, aquelles trames marcades seran eliminades.

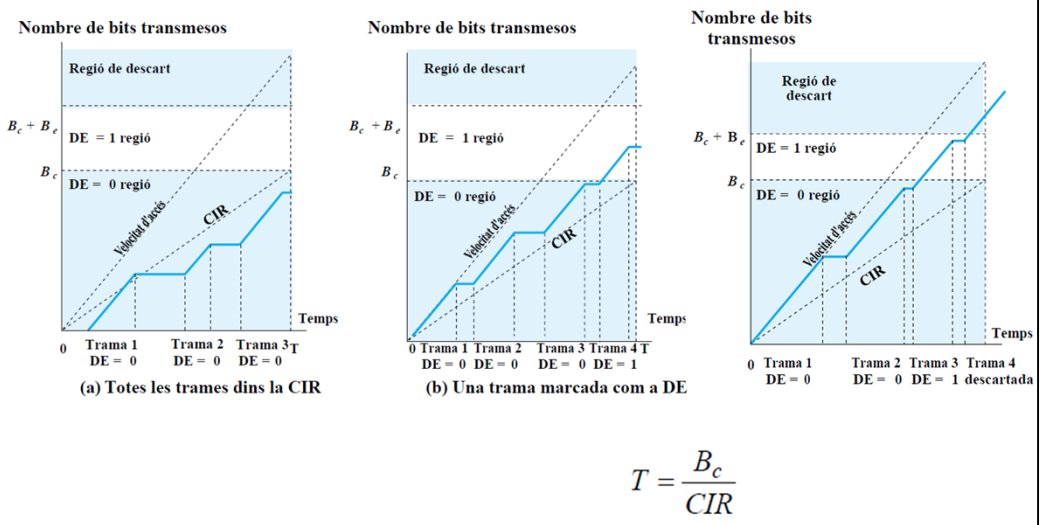
T



El CIR indica el límit de transmissió garantida (Bc). A partir d'aquí hi ha un altra límit (Be) en el que les trames es marquen amb baixa prioritat, i superat aquest límit (maximum rate) no es deixen entrar més dades.

T

Congestion control



Es fixa un temps de mesura T. Es compten els bits que s'envien al llarg de T. Si el número supera Bc es marca la unitat de dades que tingui un mínim d'un bit superat. Si supera Be un sol bit no es deixa entrar la trama. Veiem tres casos diferents en funció del ritme d'enviament de trames.