**MỤC LỤC**

[**CHƯƠNG 1: GIỚI THIỆU VỀ QOS** 3](#_Toc182228280)

[**1.1.** **Tại sao lại chọn QoS** 3](#_Toc182228281)

[**1.2.** **Ta sẽ học được những gì từ module này** 3](#_Toc182228282)

[**CHƯƠNG 2: CHẤT LƯỢNG TRUYỀN DẪN CỦA MẠNG** 3](#_Toc182228283)

[**2.1.**  **Ưu tiên chất lượng truy cập** 3](#_Toc182228284)

[**2.2.**  **Bandwith, congestion, delay và jitter** 4](#_Toc182228285)

[**2.3.** **Paclet loss** 5](#_Toc182228286)

[**CHƯƠNG 3: ĐẶC ĐIỂM CỦA LƯU LƯỢNG** 6](#_Toc182228287)

[**3.2.**  **Network traffic trends** 6](#_Toc182228288)

[**3.3.**  **Voice** 7](#_Toc182228289)

[**3.4.**  **Video** 7](#_Toc182228290)

[**3.5.**  **Data** 8](#_Toc182228291)

[**CHƯƠNG 4: THUẬT TOÁN XẾP HÀNG** 9](#_Toc182228292)

[**4.2.**  **Tổng quan về hàng đợi** 9](#_Toc182228293)

[**4.3.**  **First in firsr out** 9](#_Toc182228294)

[**4.4.**  **Weighted Fair Queuing (WFQ)** 10](#_Toc182228295)

[**4.5.** **Class-Bassed Weighted Fair Queuing (CBWFQ)** 10](#_Toc182228296)

[**4.6.**  **Low latency queuing (LLQ)** 11](#_Toc182228297)

[**CHƯƠNG 5: MÔ HÌNH QOS** 12](#_Toc182228298)

[**5.2.**  **Lựa chọn mô hình QoS phù hợp** 12](#_Toc182228299)

[**5.3.**  **Best effort** 12](#_Toc182228300)

[**5.4.**  **Integrated Services** 13](#_Toc182228301)

[**5.5.**  **Differentiated Services** 14](#_Toc182228302)

[**CHƯƠNG 6: CÁC KỸ THUẬT ĐẢM BẢO QOS** 15](#_Toc182228303)

[**6.2.**  **Tránh mất gói tin** 15](#_Toc182228304)

[**6.3.**  **Các công cụ QoS** 15](#_Toc182228305)

[**6.4.**  **Phân loại và đánh dấu** 15](#_Toc182228306)

[**6.5.**  **Đánh dấu ở Layer 2** 16](#_Toc182228307)

[**6.6.** **Đánh dấu ở lớp 3** 16](#_Toc182228308)

[**6.7.** **Loại dịch vụ và trường lớp lưu lượng truy cập** 17](#_Toc182228309)

[**6.8.** **Giá trị DSCP** 18](#_Toc182228310)

[**6.9.** **Bit chọn lớp** 18](#_Toc182228311)

[**6.10.** **Ranh giới tin cậy** 19](#_Toc182228312)

[**6.11.** **Tránh tắc nghẽn** 19](#_Toc182228313)

[**6.12.**  **Định hình và chính sách** 20](#_Toc182228314)

[**6.13.** **Chính sách QoS** 20](#_Toc182228315)

**Module 9: QoS**

# **CHƯƠNG 1: GIỚI THIỆU VỀ QOS**

## **Tại sao lại chọn QoS**

- Tưởng tượng rằng bạn đang trên một con đường tắc nghẽn và bạn vội để đến nhà hàng ăn tối với người thân. Bạn nghe thấy tiếng còi hú và đèn nhấp nháy từ chiếc xe cứu thương phía sau lưng. Việc cần làm là phải nhường đường cho chiếc cứu thương này đi qua. Chiếc cứu thương này cần đến bệnh viện đúng giờ hơn là việc bạn đến nhà hàng đúng giờ.

- Việc chiếc xe cứu thương này được ưu tiên lưu thông trên con đường hơn là những chiếc xe còn lại. Đây là cách QoS hoạt động.

## **Ta sẽ học được những gì từ module này**

- Chủ đề: Khái niệm QoS

- Mục tiêu: Giải thích được, ứng dụng của QoS với thiết bị mạng

|  |  |
| --- | --- |
| Thuật ngữ | Giải thích |
| Network Transmission | Giải thích được thuộc tính truyền dẫn của mạng ảnh hưởng đến chất lượng như nào |
| Traffic Characteristics | Mô tả được yêu cầu ít nhất của mạng dnafh cho gói tin voice, video và dữ liệu |
| Queuing Algorithms | Mô tả được thuật toán hàng đợi được dùng trong các thiết bị mạng |
| QoS Models | Mô tả sự khác nhau các mô hình QoS |
| QoS Implementation Techniques | Giải thích QoS dùng cơ chế nào để đảm bảo chất lượng truyền dẫn |

# **CHƯƠNG 2: CHẤT LƯỢNG TRUYỀN DẪN CỦA MẠNG**

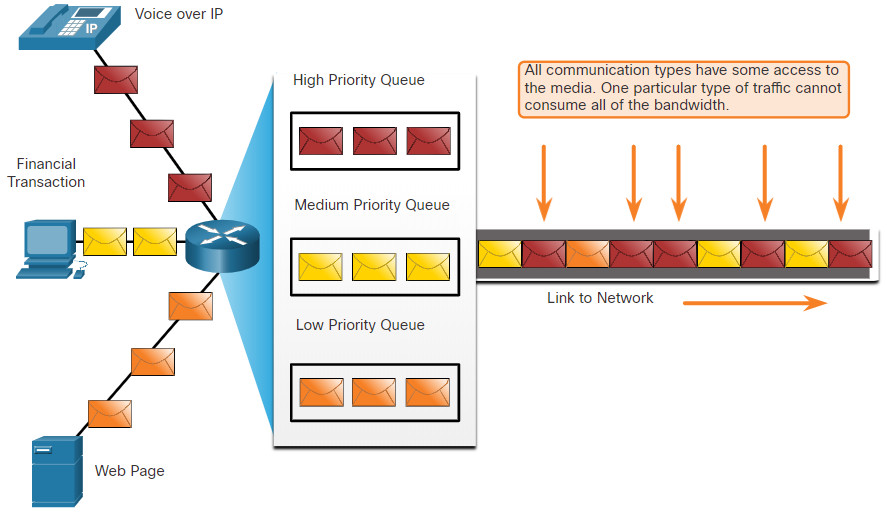
## **2.1. Ưu tiên chất lượng truy cập**

- QoS là một yêu cầu ngày càng tăng của các mạng hiện nay. Các ứng dụng mới, như thoại và truyền video trực tiếp, tạo ra kỳ vọng cao hơn về chất lượng cung cấp giữa người dùng.

- Hiện tượng tắc nghẽn xảy ra khi nhiều đường truyền giao tiếp kết hợp vào một thiết bị duy nhất như bộ định tuyến, và sau đó phần lớn dữ liệu đó được đặt trên chỉ một vài cổng đầu ra hoặc trên một cổng chậm hơn. Tắc nghẽn cũng có thể xảy ra khi các gói dữ liệu lớn ngăn chặn các gói dữ liệu nhỏ hơn được truyền một cách kịp thời.

- Khi khối lượng lưu lượng truy cập lớn hơn so với khả năng vận chuyển qua mạng, các thiết bị xếp hàng (giữ) các gói dữ liệu trong bộ nhớ cho đến khi có đủ tài nguyên để truyền chúng. Xếp hàng các gói dữ liệu gây ra độ trễ vì các gói dữ liệu mới không thể được truyền cho đến khi các gói dữ liệu trước đó đã được xử lý. Nếu số lượng gói dữ liệu cần xếp hàng tiếp tục tăng, bộ nhớ trong thiết bị sẽ đầy và các gói dữ liệu sẽ bị bỏ rơi. Một kỹ thuật QoS có thể giúp giải quyết vấn đề này là phân loại dữ liệu thành nhiều hàng đợi, như được hiển thị trong hình.

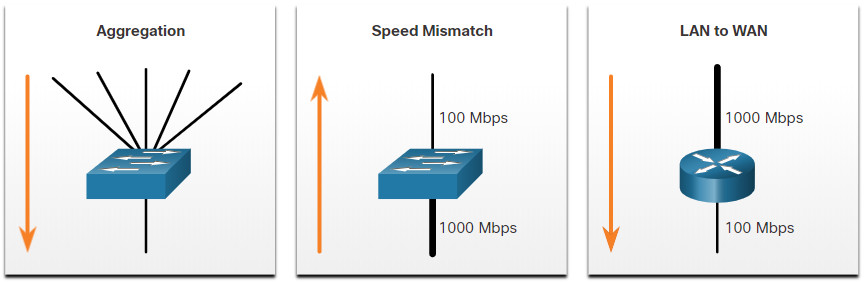
**Lưu ý**: Một thiết bị chỉ áp dụng QoS khi nó đang gặp phải một số loại tắc nghẽn.



## **2.2. Bandwith, congestion, delay và jitter**

- Băng thông mạng được đo bằng số bit có thể truyền trong một giây, hoặc bit mỗi giây (bps). Ví dụ, một thiết bị mạng có thể được mô tả là có khả năng hoạt động ở tốc độ 10 gigabit mỗi giây (Gbps).

- Tắc nghẽn mạng gây ra độ trễ. Một cổng gặp tắc nghẽn khi nó phải xử lý lưu lượng truy cập nhiều hơn khả năng của nó. Các điểm tắc nghẽn mạng là những ứng viên lý tưởng cho các cơ chế QoS. Độ trễ hoặc thời gian chờ phản hồi đề cập đến thời gian cần thiết để một gói dữ liệu di chuyển từ nguồn đến đích.



- Hai loại độ trễ là cố định và biến đổi. Độ trễ cố định là một khoảng thời gian cụ thể mà một quá trình cụ thể mất, chẳng hạn như thời gian cần thiết để đặt một bit trên phương tiện truyền dẫn. Độ trễ biến đổi mất một khoảng thời gian không xác định và bị ảnh hưởng bởi các yếu tố như lượng lưu lượng đang được xử lý.

|  |  |
| --- | --- |
| **Loại** | **Mô tả** |
| Code delay | Khoảng thời gian cố định để nén dữ liệu tại nguồn tước khi truyền đến thiết bị mạng |
| Packetization delay | Khoảng thời gian cố định để phân giải gói tin với các thông tin header cần thiết |
| Queuing delay | Khoảng thời gian có thể thay đổi của frame hoặc gói tin đợi để được truyền đi |
| Serialization delay | Khoảng thời gian cố định để truyền 1 frame lên dây |
| Propagation delay | Khoảng thời gian có thể thay đổi mà cần để frame di chuyển từ nguồn tới đích |
| De-jitter delay | Khoảng thời gian cố định cần thiết để đệm một luồng gói dữ liệu và sau đó gửi chúng ra theo khoảng cách khác nhau |

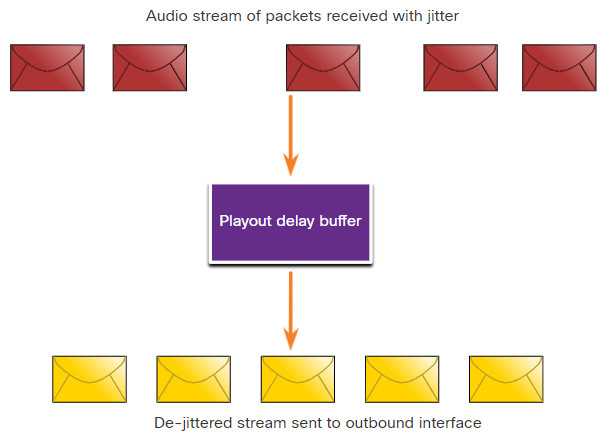
- Jitter là sự thay đổi trong độ trễ của các gói dữ liệu nhận được. Ở phía gửi, các gói dữ liệu được gửi theo một luồng liên tục với các gói dữ liệu cách đều nhau. Do tắc nghẽn mạng, xếp hàng không đúng hoặc lỗi cấu hình, độ trễ giữa mỗi gói dữ liệu có thể thay đổi thay vì giữ nguyên. Cả độ trễ và jitter đều cần được kiểm soát và giảm thiểu để hỗ trợ lưu lượng thời gian thực và tương tác.

## **2.3. Paclet loss**

- Không có kĩ thuật QoS nào được áp dụng, các gói dữ liệu được xử lý theo thứ tự chúng được nhận. Khi xảy ra tắc nghẽn, các thiết bị mạng như bộ định tuyến và bộ chuyển mạch có thể bỏ gói dữ liệu. Điều này có nghĩa là các gói nhạy cảm với thời gian, chẳng hạn như video và thoại thời gian thực, sẽ bị bỏ rơi với tần suất tương tự như dữ liệu không nhạy cảm với thời gian, chẳng hạn như email và duyệt web.

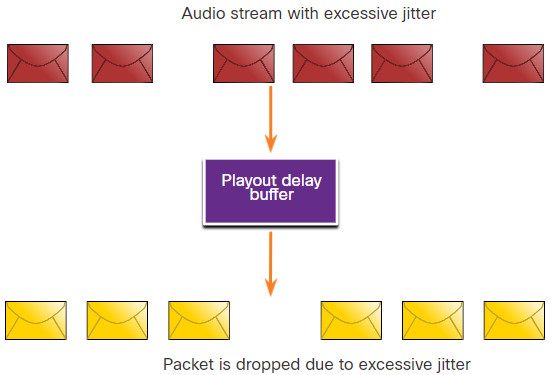
- Khi một bộ định tuyến nhận được một luồng âm thanh kỹ thuật số giao thức thời gian thực (RTP) cho Voice IP (VoIP), nó phải bù đắp cho jitter gặp phải. Cơ chế xử lý chức năng này là bộ đệm độ trễ phát lại. Bộ đệm độ trễ phát lại phải đệm các gói dữ liệu này và sau đó phát lại chúng theo một luồng ổn định, như được hiển thị trong hình. Các gói dữ liệu kỹ thuật số sau đó được chuyển đổi lại thành luồng âm thanh analog.

**Playout Delay Buffer Compensates for Jitter**



- Nếu jitter quá lớn đến mức khiến các gói dữ liệu được nhận ra ngoài phạm vi của bộ đệm này, các gói dữ liệu ngoài phạm vi sẽ bị loại bỏ và âm thanh bị gián đoạn, như được hiển thị trong hình.

**Packet Dropped Due to Excessive Jitter**



- Đối với tổn thất nhỏ như một gói dữ liệu, bộ xử lý tín hiệu số (DSP) nội suy những gì nó nghĩ rằng âm thanh nên là và không có vấn đề gì khi nghe đối với người dùng. Tuy nhiên, khi jitter vượt quá khả năng của DSP để bù đắp cho các gói dữ liệu bị thiếu, các vấn đề về âm thanh sẽ được nghe thấy.

- Mất gói dữ liệu là một nguyên nhân rất phổ biến của các vấn đề về chất lượng giọng nói trên mạng IP. Trong một mạng được thiết kế đúng cách, mất gói dữ liệu nên gần bằng không. Các codec giọng nói được sử dụng bởi DSP có thể chịu được một số mức độ mất gói dữ liệu mà không ảnh hưởng đáng kể đến chất lượng giọng nói.

- Các kỹ sư mạng sử dụng cơ chế QoS để phân loại các gói thoại cho mất gói dữ liệu bằng không. Băng thông được đảm bảo cho các cuộc gọi thoại bằng cách ưu tiên cho lưu lượng thoại hơn lưu lượng không nhạy cảm với độ trễ.

# **CHƯƠNG 3: ĐẶC ĐIỂM CỦA LƯU LƯỢNG**

## **3.2. Network traffic trends**

- Trong một chủ đề trước, bạn đã học về chất lượng truyền tải mạng. Trong chủ đề này, bạn sẽ học về đặc điểm lưu lượng (giọng nói, video và dữ liệu). Vào đầu những năm 2000, các loại lưu lượng IP phổ biến là giọng nói và dữ liệu. Lưu lượng thoại có nhu cầu băng thông dự đoán được và thời gian đến gói dữ liệu đã biết. Lưu lượng dữ liệu không phải là thời gian thực và có nhu cầu băng thông không thể đoán trước được.

- Lưu lượng dữ liệu có thể đột biến tạm thời, như khi một tệp lớn đang được tải xuống. Sự đột biến này có thể tiêu thụ toàn bộ băng thông của một liên kết. Gần đây hơn, lưu lượng video đã trở nên ngày càng quan trọng đối với giao tiếp và hoạt động kinh doanh. Theo Cisco Visual Networking Index (VNI), lưu lượng video chiếm 70% tổng lưu lượng vào năm 2017. Đến năm 2022, video sẽ chiếm 82% tổng lưu lượng. Ngoài ra, lưu lượng video di động sẽ đạt 60,9 exabyte mỗi tháng vào năm 2022, tăng từ 6,8 exabyte mỗi tháng vào năm 2017. Loại yêu cầu mà lưu lượng thoại, video và dữ liệu đặt ra cho mạng là rất khác nhau.

## **3.3. Voice**

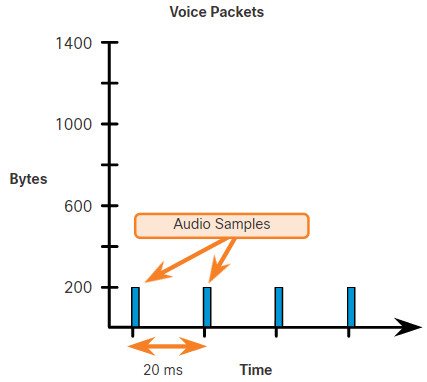
- Lưu lượng thoại có thể dự đoán được và mượt mà, như được hiển thị trong hình. Tuy nhiên, giọng nói rất nhạy cảm với độ trễ và gói dữ liệu bị bỏ qua. Không có ý nghĩa gì khi truyền lại giọng nói nếu các gói dữ liệu bị mất; do đó, các gói thoại phải nhận được ưu tiên cao hơn các loại lưu lượng khác. Ví dụ, các sản phẩm của Cisco sử dụng phạm vi cổng RTP từ 16384 đến 32767 để ưu tiên lưu lượng thoại. Giọng nói có thể chịu được một mức độ nhất định của độ trễ, jitter và mất mà không có bất kỳ ảnh hưởng đáng chú ý nào. Độ trễ không nên vượt quá 150 mili giây (ms). Jitter không nên vượt quá 30 ms và mất gói thoại không nên vượt quá 1%. Lưu lượng thoại yêu cầu ít nhất 30 Kbps băng thông. Bảng dưới đây tóm tắt các đặc điểm và yêu cầu của lưu lượng thoại.

|  |  |
| --- | --- |
| **Thuộc tính của lưu lượng giọng nói** | **Yêu cầu 1 chiều** |
| Mượt Nhỏ Nhạy cảm việc rơi Nhạy cảm việc độ trễ Ưu tiên UDP | Độ trễ =< 150ms jitter =< 30ms mất =< 1% băng thông (30-128kbps) |

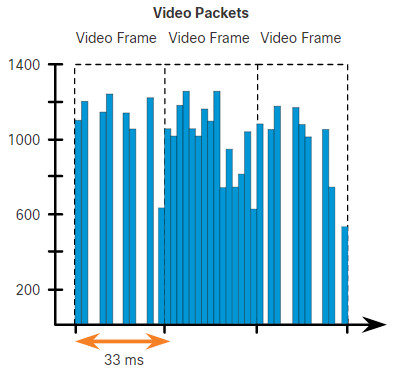
## **3.4. Video**

- Không có QoS và một lượng lớn dung lượng băng thông bổ sung, chất lượng video thường bị suy giảm. Hình ảnh xuất hiện mờ, răng cưa hoặc chuyển động chậm. Phần âm thanh của nguồn cấp có thể trở nên không đồng bộ với video.

- Lưu lượng video có xu hướng không thể đoán trước, không nhất quán và đột biến so với lưu lượng thoại. So với giọng nói, video kém đàn hồi hơn đối với tổn thất và có khối lượng dữ liệu lớn hơn mỗi gói. Lưu ý trong hình cách các gói thoại đến mỗi 20 ms và mỗi gói có kích thước 200 byte có thể dự đoán được.



- Ngược lại, số lượng và kích thước của các gói video thay đổi mỗi 33 ms dựa trên nội dung của video, như được hiển thị trong hình. Ví dụ, nếu luồng video bao gồm nội dung không thay đổi nhiều từ khung này sang khung khác, thì các gói video sẽ nhỏ và ít gói hơn là cần thiết để duy trì trải nghiệm người dùng chấp nhận được. Tuy nhiên, nếu luồng video bao gồm nội dung thay đổi nhanh chóng, chẳng hạn như một chuỗi hành động trong một bộ phim, thì các gói video sẽ lớn hơn. Cần nhiều hơn mỗi khoảng thời gian 33 ms để duy trì trải nghiệm người dùng chấp nhận được.



- Các cổng UDP như 554 được sử dụng cho Giao thức phát trực tuyến thời gian thực (RSTP) và nên được ưu tiên hơn các lưu lượng mạng khác ít nhạy cảm với độ trễ hơn. Tương tự như giọng nói, video có thể chịu được một mức độ nhất định của độ trễ, jitter và mất mà không có bất kỳ ảnh hưởng đáng chú ý nào. Độ trễ không nên vượt quá 400 mili giây (ms). Jitter không nên vượt quá 50 ms và mất gói video không nên vượt quá 1%. Lưu lượng video yêu cầu ít nhất 384 Kbps băng thông. Bảng dưới đây tóm tắt các đặc điểm và yêu cầu của lưu lượng video.

## **3.5. Data**

- Hầu hết các ứng dụng sử dụng TCP hoặc UDP. Không giống như UDP, TCP có thể sửa lỗi. Các ứng dụng dữ liệu không dung nạp với mất dữ liệu, chẳng hạn như email và trang web, sử dụng TCP để đảm bảo rằng, nếu các gói dữ liệu bị mất trong quá trình truyền, chúng sẽ được gửi lại. Lưu lượng dữ liệu có thể mượt mà hoặc đột biến. Lưu lượng điều khiển mạng thường mượt mà và có thể dự đoán được. Khi có thay đổi cấu trúc, lưu lượng điều khiển mạng có thể đột biến trong vài giây. Nhưng khả năng của các mạng hiện nay có thể dễ dàng xử lý sự gia tăng lưu lượng điều khiển mạng khi mạng hội tụ.

- Tuy nhiên, một số ứng dụng TCP có thể tiêu thụ một phần lớn dung lượng mạng. FTP sẽ tiêu thụ càng nhiều băng thông càng tốt khi bạn tải xuống một tệp lớn, chẳng hạn như phim hoặc trò chơi. Bảng dưới đây tóm tắt các đặc điểm của lưu lượng dữ liệu.

Thuộc tính của lưu lượng dữ liệu:

* Mượt/đột biến
* ít/nhiều
* không nhạy cảm với việc bỏ rớt
* không nhạy cảm với độ trễ
* TCP gửi lại

# **CHƯƠNG 4: THUẬT TOÁN XẾP HÀNG**

## **4.2. Tổng quan về hàng đợi**

- Chủ đề trước đã đề cập đến đặc điểm lưu lượng. Chủ đề này sẽ giải thích các thuật toán xếp hàng được sử dụng để thực hiện QoS. Chính sách QoS được thực hiện bởi người quản trị mạng trở nên hoạt động khi xảy ra tắc nghẽn trên liên kết.

- Xếp hàng là một công cụ quản lý tắc nghẽn có thể đệm, ưu tiên và, nếu cần, sắp xếp lại các gói dữ liệu trước khi được truyền đến đích. Có một số thuật toán xếp hàng có sẵn.

- Để phục vụ mục đích của khóa học này, chúng ta sẽ tập trung vào các thuật toán sau:

* First-In, First-Out (FIFO)
* Weighted Fair Queuing (WFQ)
* Class-Based Weighted Fair Queuing (CBWFQ)
* Low Latency Queuing (LLQ)

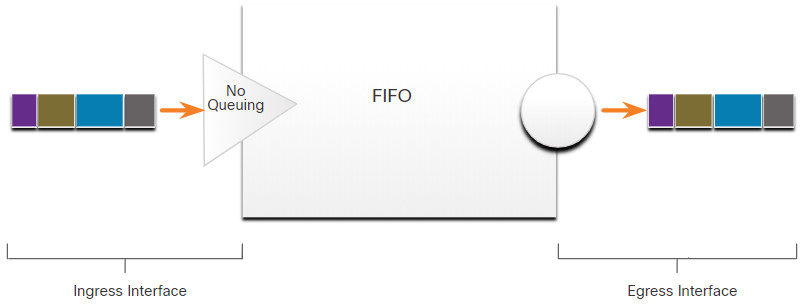
## **4.3. First in firsr out**

- Ở dạng đơn giản nhất, xếp hàng Đầu vào Đầu ra (FIFO), còn được gọi là xếp hàng theo thứ tự đến trước, phục vụ trước, đệm và chuyển tiếp các gói theo thứ tự chúng đến.

- FIFO không có khái niệm về ưu tiên hoặc lớp lưu lượng và do đó, không đưa ra quyết định nào về ưu tiên gói. Chỉ có một hàng đợi và tất cả các gói được xử lý như nhau. Các gói được gửi ra khỏi một cổng theo thứ tự chúng đến, như được hiển thị trong hình. Mặc dù một số lưu lượng có thể quan trọng hơn hoặc nhạy cảm với thời gian dựa trên phân loại ưu tiên, nhưng lưu ý rằng lưu lượng được gửi ra theo thứ tự nhận được.

- Khi sử dụng FIFO, lưu lượng quan trọng hoặc nhạy cảm với thời gian có thể bị bỏ rơi khi có tắc nghẽn trên cổng bộ định tuyến hoặc bộ chuyển mạch. Khi không cấu hình các chiến lược xếp hàng khác, tất cả các cổng, ngoại trừ các cổng nối tiếp ở E1 (2,048 Mbps) trở xuống, đều sử dụng FIFO theo mặc định. (Các cổng nối tiếp ở E1 trở xuống sử dụng WFQ theo mặc định.)

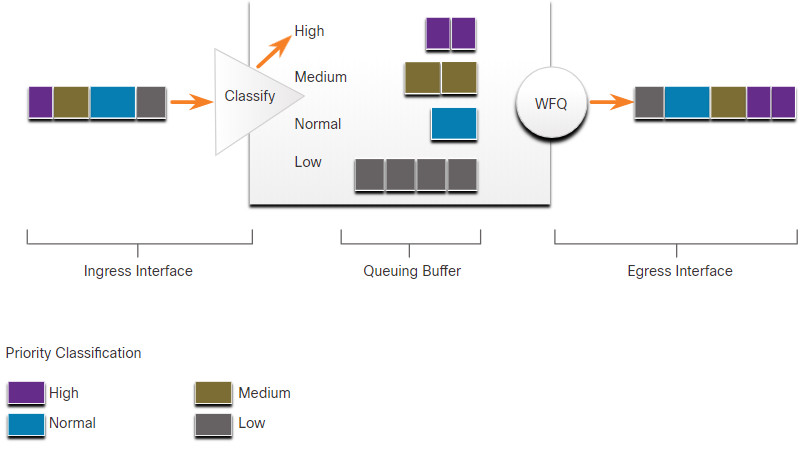
- FIFO, là phương pháp xếp hàng nhanh nhất, hiệu quả cho các liên kết lớn có độ trễ thấp và tắc nghẽn tối thiểu. Nếu liên kết của bạn có rất ít tắc nghẽn, xếp hàng FIFO có thể là phương pháp xếp hàng duy nhất bạn cần sử dụng.



## **4.4. Weighted Fair Queuing (WFQ)**

- Xếp hàng công bằng có trọng số (WFQ) là một phương pháp lập lịch tự động cung cấp phân bổ băng thông công bằng cho tất cả lưu lượng mạng. WFQ không cho phép cấu hình các tùy chọn phân loại. WFQ áp dụng ưu tiên hoặc trọng số cho lưu lượng được xác định và phân loại nó thành các cuộc trò chuyện hoặc luồng, như được hiển thị trong hình.

**Weighted Fair Queuing Example**



- Sau đó, WFQ xác định lượng băng thông mỗi luồng được phép tương đối với các luồng khác. Thuật toán dựa trên luồng được sử dụng bởi WFQ đồng thời lập lịch cho lưu lượng tương tác lên đầu hàng đợi để giảm thời gian phản hồi. Sau đó, nó chia sẻ công bằng băng thông còn lại giữa các luồng băng thông cao. WFQ cho phép bạn ưu tiên lưu lượng tương tác, dung lượng thấp, chẳng hạn như phiên Telnet và thoại, hơn lưu lượng dung lượng cao, chẳng hạn như phiên FTP. Khi nhiều luồng chuyển tệp xảy ra đồng thời, các luồng chuyển tệp được cung cấp băng thông tương đương.

- WFQ phân loại lưu lượng thành các luồng khác nhau dựa trên địa chỉ tiêu đề gói, bao gồm các đặc điểm như địa chỉ IP nguồn và đích, địa chỉ MAC, số cổng, giao thức và giá trị Loại dịch vụ (ToS). Giá trị ToS trong tiêu đề IP có thể được sử dụng để phân loại lưu lượng.

- Các luồng lưu lượng băng thông thấp, chiếm phần lớn lưu lượng, nhận được dịch vụ ưu tiên cho phép gửi toàn bộ tải trọng được cung cấp của chúng một cách kịp thời. Các luồng lưu lượng dung lượng cao chia sẻ công suất còn lại một cách tương đương với nhau.

Sự giới hạn

- WFQ không được hỗ trợ với tunnel và mã hóa vì các tính năng này sửa đổi thông tin nội dung gói được yêu cầu bởi WFQ để phân loại.

- Mặc dù WFQ tự động thích ứng với các điều kiện lưu lượng mạng thay đổi, nhưng nó không cung cấp mức độ kiểm soát chính xác về phân bổ băng thông mà CBWFQ cung cấp.

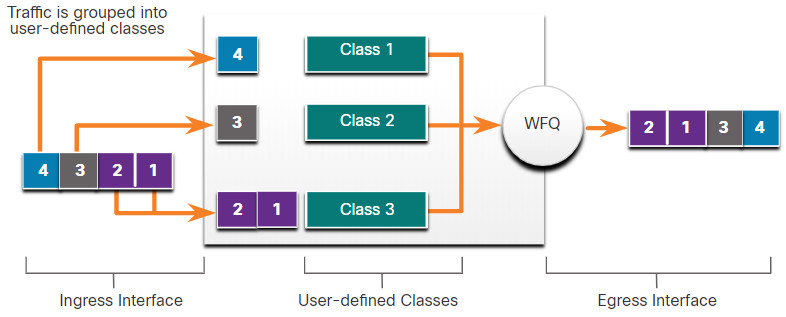
## **4.5. Class-Bassed Weighted Fair Queuing (CBWFQ)**

- Xếp hàng công bằng có trọng số dựa trên lớp (CBWFQ) mở rộng chức năng WFQ tiêu chuẩn để cung cấp hỗ trợ cho các lớp lưu lượng do người dùng xác định. Với CBWFQ, bạn xác định các lớp lưu lượng dựa trên các tiêu chí khớp bao gồm giao thức, danh sách kiểm soát truy cập (ACL) và cổng đầu vào. - - Các gói thỏa mãn các tiêu chí khớp cho một lớp tạo thành lưu lượng cho lớp đó. Một hàng đợi FIFO được dành riêng cho mỗi lớp và lưu lượng thuộc về một lớp được chuyển hướng đến hàng đợi cho lớp đó, như được hiển thị trong hình.

- Khi một lớp đã được xác định theo các tiêu chí khớp của nó, bạn có thể gán cho nó các đặc điểm. Để đặc trưng hóa một lớp, bạn gán cho nó băng thông, trọng số và giới hạn gói tối đa. Băng thông được gán cho một lớp là băng thông được đảm bảo cung cấp cho lớp trong thời gian tắc nghẽn.

- Để đặc trưng hóa một lớp, bạn cũng chỉ định giới hạn hàng đợi cho lớp đó, đó là số lượng tối đa các gói được phép tích lũy trong hàng đợi cho lớp. Các gói thuộc về một lớp tuân theo các giới hạn băng thông và hàng đợi đặc trưng cho lớp.

**CBWFQ Example**



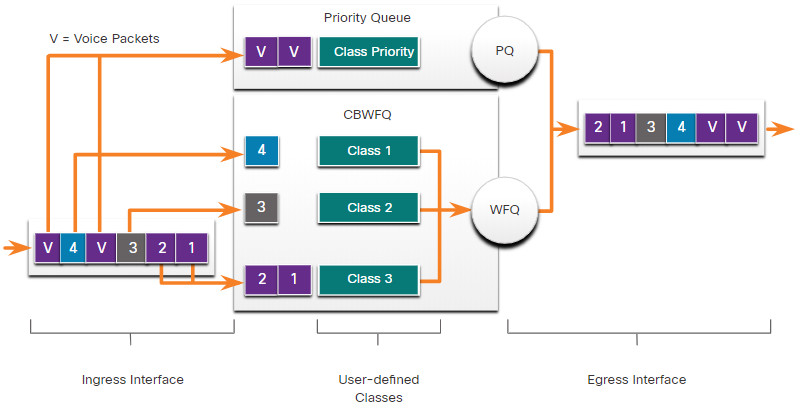
- Sau khi một hàng đợi đạt đến giới hạn hàng đợi được cấu hình, việc thêm nhiều gói vào lớp sẽ khiến việc thả đuôi hoặc bỏ gói có hiệu lực, tùy thuộc vào cách cấu hình chính sách lớp. Thả đuôi có nghĩa là một bộ định tuyến chỉ đơn giản là bỏ bất kỳ gói nào đến cuối hàng đợi đã sử dụng hết tài nguyên giữ gói của nó. Đây là phản hồi xếp hàng mặc định đối với tắc nghẽn. Thả đuôi xử lý tất cả lưu lượng như nhau và không phân biệt giữa các lớp dịch vụ.

## **4.6. Low latency queuing (LLQ)**

- Tính năng Xếp hàng độ trễ thấp (LLQ) mang đến xếp hàng ưu tiên nghiêm ngặt (PQ) cho CBWFQ. PQ nghiêm ngặt cho phép các gói nhạy cảm với độ trễ như giọng nói được gửi trước các gói trong các hàng đợi khác. LLQ cung cấp xếp hàng ưu tiên nghiêm ngặt cho CBWFQ, giảm jitter trong các cuộc trò chuyện thoại, như được hiển thị trong hình.

- Không có LLQ, CBWFQ cung cấp WFQ dựa trên các lớp được xác định mà không có hàng đợi ưu tiên nghiêm ngặt cho lưu lượng thời gian thực. Trọng số cho một gói thuộc về một lớp cụ thể được lấy từ băng thông bạn đã gán cho lớp khi bạn cấu hình nó. Do đó, băng thông được gán cho các gói của một lớp xác định thứ tự gửi các gói. Tất cả các gói đều được phục vụ công bằng dựa trên trọng số; không có lớp gói nào có thể được cấp ưu tiên nghiêm ngặt. Sơ đồ này đặt ra vấn đề cho lưu lượng thoại khó chịu với độ trễ, đặc biệt là sự thay đổi về độ trễ. Đối với lưu lượng thoại, sự thay đổi về độ trễ dẫn đến những bất thường trong truyền dẫn thể hiện dưới dạng jitter trong cuộc trò chuyện được nghe.

- LLQ cho phép các gói nhạy cảm với độ trễ như giọng nói được gửi trước (trước các gói trong các hàng đợi khác), ưu tiên các gói nhạy cảm với độ trễ hơn các lưu lượng khác. Mặc dù có thể phân loại các loại lưu lượng thời gian thực khác nhau vào hàng đợi ưu tiên nghiêm ngặt, Cisco khuyến nghị chỉ nên chuyển lưu lượng thoại đến hàng đợi ưu tiên.



# **CHƯƠNG 5: MÔ HÌNH QOS**

## **5.2. Lựa chọn mô hình QoS phù hợp**

* -Best-effort Model
* -Integrated service (IntServ)
* -Differentiated services (DiffServ)

|  |  |
| --- | --- |
| **Mô hình** | **Mô tả** |
| Best-effort model | Mô hình này không thật sự ứng dụng QoS vì không cấu hình QoS  Mô hình này không cần thiết phải sử dung QoS |
| Integrated services | IntServ cung cấp QoS rất cao cho các gói IP với việc giao hàng được đảm bảo  Nó định nghĩa một quá trình báo hiệu cho các ứng dụng để báo hiệu cho mạng rằng chúng yêu cầu QoS đặc biệt trong một thời gian và băng thông nên được dành riêng  IntServ có thể hạn chế đáng kể khả năng mở rộng của một mạng |
| Differentiated services | DiffServ cung cấp khả năng mở rộng và linh hoạt cao trong việc triển khai QoS  Các thiết bị mạng nhận ra các lớp lưu lượng và cung cấp các mức QoS khác nhau cho các lớp lưu lượng khác nhau |

## **5.3. Best effort**

- Thiết kế cơ bản của internet là phân phối gói dữ liệu tối ưu và không cung cấp bất kỳ đảm bảo nào. Cách tiếp cận này vẫn chiếm ưu thế trên internet ngày nay và vẫn phù hợp cho hầu hết các mục đích. Mô hình tối ưu xử lý tất cả các gói mạng theo cùng một cách, vì vậy một tin nhắn thoại khẩn cấp được xử lý theo cùng một cách như một bức ảnh kỹ thuật số được đính kèm với một email. Không có QoS, mạng không thể phân biệt giữa các gói và do đó, không thể xử lý các gói ưu tiên.

- Mô hình tối ưu tương tự như khái niệm gửi thư bằng thư bưu điện tiêu chuẩn. Thư của bạn được xử lý giống hệt như mọi bức thư khác. Với mô hình tối ưu, thư có thể không bao giờ đến, và trừ khi bạn có một thỏa thuận thông báo riêng với người nhận thư, bạn có thể không bao giờ biết rằng thư đã không đến.

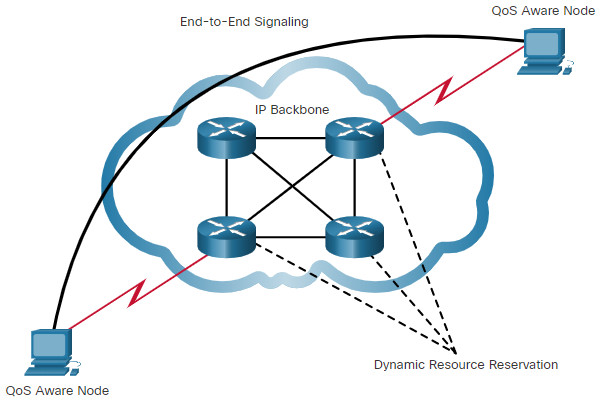
|  |  |
| --- | --- |
| **Lợi ích** | **Hạn chế** |
| Mô hình mở rộng cao nhất | Không có |
| Khả năng mở rộng bị hạn chế bởi băng thông có sẵn, trong trường hợp đó tất cả lưu lượng đều bị ảnh hưởng như nhau | Các gói dữ liệu sẽ đến bất cứ khi nào chúng có thể và theo bất kỳ thứ tự nào có thể khi chúng đến |
| Không yêu cầu kĩ thuật QoS đặt biệt | Không có gói dữ liệu nào được ưu tiên |
| Mô hình triển khai nhanh và dễ nhất | Dữ liệu quan trọng được xử lý giống như dữ liệu thông thường. |

## **5.4. Integrated Services**

- Mô hình kiến trúc IntServ (RFC 1633, 2211 và 2212) được phát triển vào năm 1994 để đáp ứng nhu cầu của các ứng dụng thời gian thực, chẳng hạn như video từ xa, hội nghị đa phương tiện, ứng dụng trực quan hóa dữ liệu và thực tế ảo. IntServ là một mô hình đa dịch vụ có thể đáp ứng nhiều yêu cầu QoS.IntServ cung cấp QoS đầu cuối mà các ứng dụng thời gian thực yêu cầu.

- IntServ quản lý rõ ràng các tài nguyên mạng để cung cấp QoS cho các luồng hoặc luồng riêng lẻ, đôi khi được gọi là vi luồng. Nó sử dụng cơ chế đặt trước tài nguyên và kiểm soát nhập học làm khối xây dựng để thiết lập và duy trì QoS. Điều này tương tự như một khái niệm được gọi là “QoS cứng”. QoS cứng đảm bảo các đặc điểm lưu lượng, chẳng hạn như băng thông, độ trễ và tỷ lệ mất gói, từ đầu đến cuối. QoS cứng đảm bảo các mức dịch vụ dự đoán được và được đảm bảo cho các ứng dụng quan trọng đối với nhiệm vụ.

**Lợi íchs and Drawbacks of IntServ Model**



|  |  |
| --- | --- |
| **Lợi ích** | **Hạn chế** |
| Kiểm soát nhập học tài nguyên đầu cuối rõ ràng  Kiểm soát nhập học chính sách theo yêu cầu  Báo hiệu số cổng động | Tốn nhiều tài nguyên do yêu cầu kiến trúc có trạng thái để báo hiệu liên tục.  Phương pháp dựa trên luồng không thể mở rộng cho các triển khai lớn như internet. |

## **5.5. Differentiated Services**

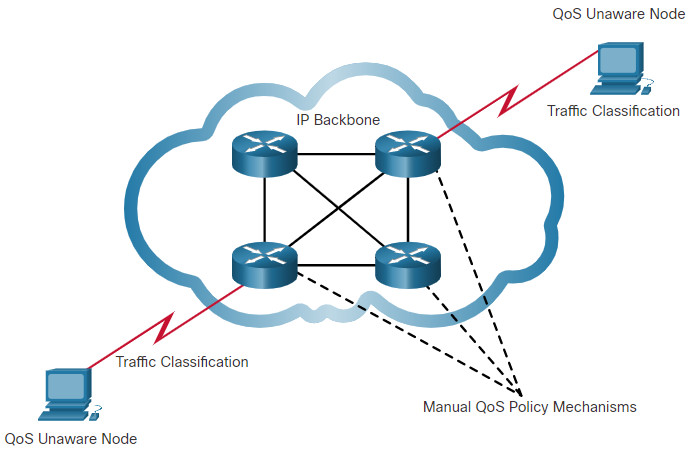
- Mô hình QoS dịch vụ phân biệt (DiffServ) chỉ định một cơ chế đơn giản và có thể mở rộng để phân loại và quản lý lưu lượng mạng. Ví dụ, DiffServ có thể cung cấp dịch vụ đảm bảo độ trễ thấp cho lưu lượng mạng quan trọng như thoại hoặc video, đồng thời cung cấp đảm bảo lưu lượng tối ưu đơn giản cho các dịch vụ không quan trọng như lưu lượng web hoặc chuyển tệp.

- Thiết kế DiffServ khắc phục được những hạn chế của cả mô hình tối ưu và IntServ. Mô hình DiffServ được mô tả trong RFC 2474, 2597, 2598, 3246, 4594. DiffServ có thể cung cấp một QoS “gần như được đảm bảo” trong khi vẫn hiệu quả về chi phí và có thể mở rộng.

- Mô hình DiffServ tương tự như khái niệm gửi một gói hàng bằng dịch vụ chuyển phát. Bạn yêu cầu (và trả tiền) cho một mức độ dịch vụ khi bạn gửi một gói hàng. Trong suốt mạng gói hàng, mức độ dịch vụ bạn đã trả tiền được công nhận và gói hàng của bạn được cung cấp dịch vụ ưu tiên hoặc bình thường, tùy thuộc vào những gì bạn yêu cầu.

- DiffServ không phải là một chiến lược QoS đầu cuối vì nó không thể thực thi các đảm bảo đầu cuối. Tuy nhiên, DiffServ QoS là một cách tiếp cận có thể mở rộng hơn để triển khai QoS. Không giống như IntServ và QoS cứng, trong đó các máy chủ cuối báo hiệu nhu cầu QoS của chúng cho mạng, DiffServ không sử dụng báo hiệu. Thay vào đó, DiffServ sử dụng phương pháp “QoS mềm”. Nó hoạt động trên mô hình QoS được cung cấp, trong đó các yếu tố mạng được thiết lập để phục vụ nhiều lớp lưu lượng khác nhau, mỗi lớp có yêu cầu QoS khác nhau.

**Benefits and Drawbacks of DiffServ Model**



|  |  |
| --- | --- |
| **Lợi ích** | **Hạn chế** |
| Có thể mở rộng cao  Cung cấp nhiều mức chất lượng khác nhau | Không có đảm bảo tuyệt đối về chất lượng dịch vụ  Yêu cầu một bộ cơ chế phức tạp để hoạt động cùng nhau trong toàn bộ mạng |

# **CHƯƠNG 6: CÁC KỸ THUẬT ĐẢM BẢO QOS**

## **6.2. Tránh mất gói tin**

- Bây giờ bạn đã học về đặc điểm lưu lượng, thuật toán xếp hàng và mô hình QoS, đã đến lúc tìm hiểu về các kỹ thuật triển khai QoS.

- Hãy bắt đầu với mất gói dữ liệu. Mất gói dữ liệu thường là kết quả của tắc nghẽn trên một cổng. Hầu hết các ứng dụng sử dụng TCP đều gặp chậm vì TCP tự động điều chỉnh theo tắc nghẽn mạng. Các phân đoạn TCP bị bỏ rơi khiến các phiên TCP giảm kích thước cửa sổ của chúng. Một số ứng dụng không sử dụng TCP và không thể xử lý các lượt bỏ (luồng dễ vỡ).

- Các cách tiếp cận sau đây có thể ngăn chặn việc bỏ rơi trong các ứng dụng nhạy cảm:

* Tăng dung lượng liên kết để giảm bớt hoặc ngăn chặn tắc nghẽn.
* Đảm bảo đủ băng thông và tăng không gian bộ đệm để chứa các luồng đột biến lưu lượng từ các luồng dễ vỡ. WFQ, CBWFQ và LLQ có thể đảm bảo băng thông và cung cấp chuyển tiếp ưu tiên cho các ứng dụng nhạy cảm với việc bỏ rơi.
* Bỏ các gói ưu tiên thấp hơn trước khi xảy ra tắc nghẽn. Cisco IOS QoS cung cấp các cơ chế xếp hàng, chẳng hạn như phát hiện sớm ngẫu nhiên có trọng số (WRED), bắt đầu bỏ các gói ưu tiên thấp hơn trước khi xảy ra tắc nghẽn.

## **6.3. Các công cụ QoS**

- Ba phân loại của công cụ QoS:

* Phân loại và đánh dấu
* Phòng tránh tắt nghẽn
* Kiểm soát tắt nghẽn

## **6.4. Phân loại và đánh dấu**

- Trước khi một gói dữ liệu có thể áp dụng chính sách QoS cho nó, gói dữ liệu phải được phân loại. Phân loại và đánh dấu cho phép chúng ta xác định hoặc “đánh dấu” các loại gói dữ liệu. Phân loại xác định lớp lưu lượng mà các gói hoặc khung thuộc về. Chỉ sau khi lưu lượng được đánh dấu, các chính sách mới có thể được áp dụng cho nó.

- Cách một gói dữ liệu được phân loại phụ thuộc vào việc triển khai QoS. Các phương pháp phân loại luồng lưu lượng ở Lớp 2 và 3 bao gồm sử dụng cổng, ACL và bản đồ lớp. Lưu lượng cũng có thể được phân loại ở Lớp 4 đến 7 bằng cách sử dụng Nhận dạng ứng dụng dựa trên mạng (NBAR).

**Lưu ý**: NBAR là một tính năng phát hiện giao thức và phân loại của phần mềm Cisco IOS hoạt động với các tính năng QoS. NBAR nằm ngoài phạm vi của khóa học này.

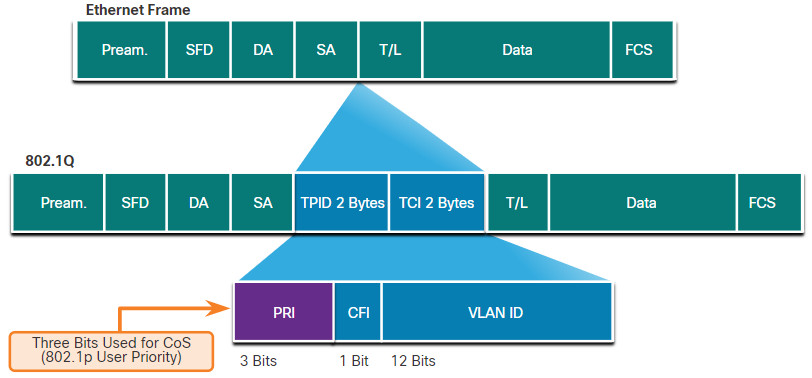
- Đánh dấu có nghĩa là chúng tôi đang thêm một giá trị vào tiêu đề gói. Các thiết bị nhận gói dữ liệu sẽ xem trường này để xem nó có khớp với một chính sách đã định hay không. Đánh dấu nên được thực hiện càng gần thiết bị nguồn càng tốt. Điều này thiết lập ranh giới tin cậy.

- Cách đánh dấu lưu lượng thường phụ thuộc vào công nghệ. Bảng trong hình mô tả một số trường đánh dấu được sử dụng trong các công nghệ khác nhau. Quyết định có nên đánh dấu lưu lượng ở Lớp 2 hoặc 3 (hoặc cả hai) không phải là đơn giản và nên được đưa ra sau khi xem xét các điểm sau:

* Đánh dấu khung Lớp 2 có thể được thực hiện cho lưu lượng không phải IP.
* Đánh dấu khung Lớp 2 là tùy chọn QoS duy nhất có sẵn cho các bộ chuyển mạch không “nhận biết IP”.
* Đánh dấu Lớp 3 sẽ mang thông tin QoS từ đầu đến cuối.

## **6.5. Đánh dấu ở Layer 2**

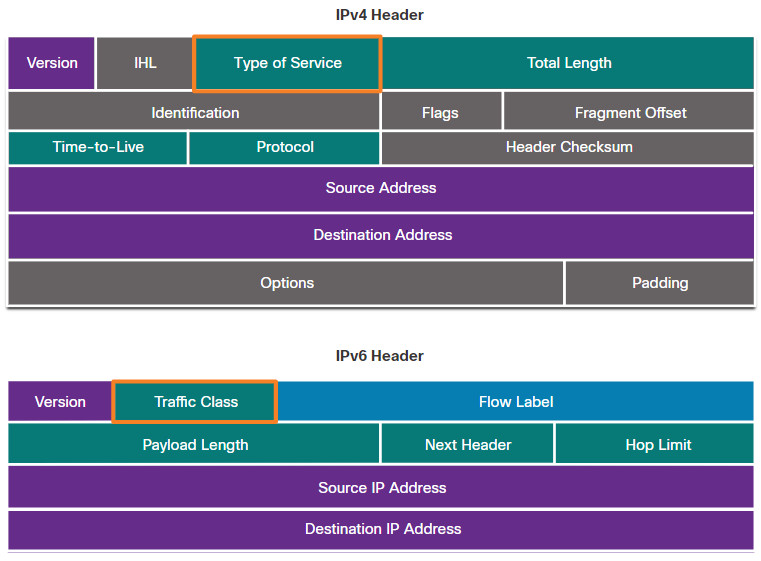
- 802.1Q là tiêu chuẩn IEEE hỗ trợ gắn thẻ VLAN ở Lớp 2 trên mạng Ethernet. Khi 802.1Q được triển khai, hai trường được thêm vào Ethernet Frame. Như được hiển thị trong hình, hai trường này được chèn vào frame Ethernet sau trường địa chỉ MAC nguồn.



- Tiêu chuẩn 802.1Q cũng bao gồm sơ đồ ưu tiên QoS được gọi là IEEE 802.1p. Tiêu chuẩn 802.1p sử dụng ba bit đầu tiên trong trường Tag Control Information (TCI). Được biết đến là trường Ưu tiên (PRI), trường 3 bit này xác định đánh dấu Class of Service (CoS). Ba bit có nghĩa là một khung Ethernet Lớp 2 có thể được đánh dấu bằng một trong tám mức ưu tiên (giá trị từ 0 đến 7) như được hiển thị trong hình.

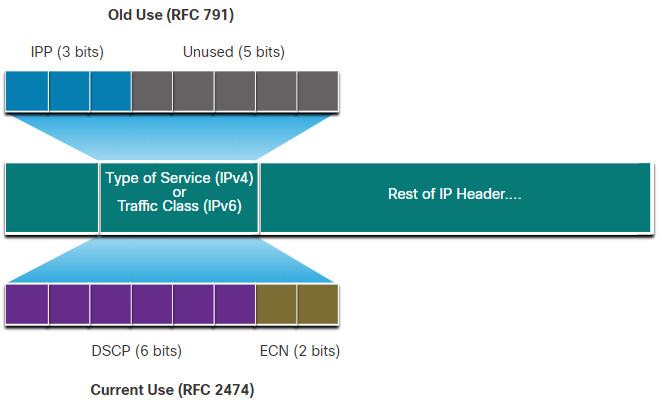
## **6.6. Đánh dấu ở lớp 3**

- IPv4 và IPv6 chỉ định một trường 8 bit trong tiêu đề gói của chúng để đánh dấu các gói. Cả IPv4 và IPv6 đều hỗ trợ một trường 8 bit để đánh dấu: trường loại dịch vụ (ToS) cho IPv4 và trường lớp lưu lượng cho IPv6 (Traffic Class).



## **6.7. Loại dịch vụ và trường lớp lưu lượng truy cập**

- Type of Service (IPv4) và Traffic Class (IPv6) mang theo đánh dấu gói như được chỉ định bởi các công cụ phân loại QoS. Trường này sau đó được tham chiếu bởi các thiết bị nhận chuyển tiếp các gói dựa trên chính sách QoS được chỉ định phù hợp.



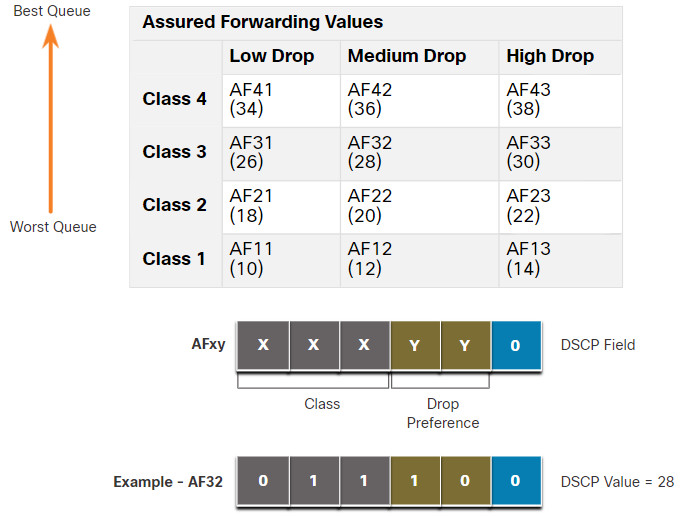
- RFC 2474 thay thế RFC 791 và định nghĩa lại trường ToS bằng cách đổi tên và mở rộng trường IPP. Trường mới, như được hiển thị trong hình, có 6 bit được phân bổ cho QoS. Được gọi là Differentiated Services Code Point (DSCP), sáu bit này cung cấp tối đa 64 lớp dịch vụ có thể có. Hai bit IP Extended Congestion Notification (ECN) còn lại có thể được sử dụng bởi các bộ định tuyến nhận biết ECN để đánh dấu các gói thay vì bỏ chúng. Đánh dấu ECN thông báo cho các bộ định tuyến hạ nguồn rằng có tắc nghẽn trong luồng gói.

## **6.8. Giá trị DSCP**

- Best-Effort (BE) – Đây là mặc định cho tất cả các gói IP. Giá trị DSCP là 0. Hành vi mỗi bước nhảy là định tuyến bình thường. Khi một bộ định tuyến gặp tắc nghẽn, các gói này sẽ bị bỏ rơi. Không có kế hoạch QoS nào được triển khai.

- Expedited Forwarding (EF) – RFC 3246 định nghĩa EF là giá trị thập phân DSCP 46 (nhị phân 101110). Ba bit đầu tiên (101) ánh xạ trực tiếp vào giá trị CoS Lớp 2 là 5 được sử dụng cho lưu lượng thoại. Ở Lớp 3, Cisco khuyến nghị chỉ sử dụng EF để đánh dấu các gói thoại.

- Assured Forwarding (AF) – RFC 2597 định nghĩa AF để sử dụng 5 bit DSCP quan trọng nhất để chỉ định hàng đợi và ưu tiên bỏ. Định nghĩa của AF được minh họa trong hình

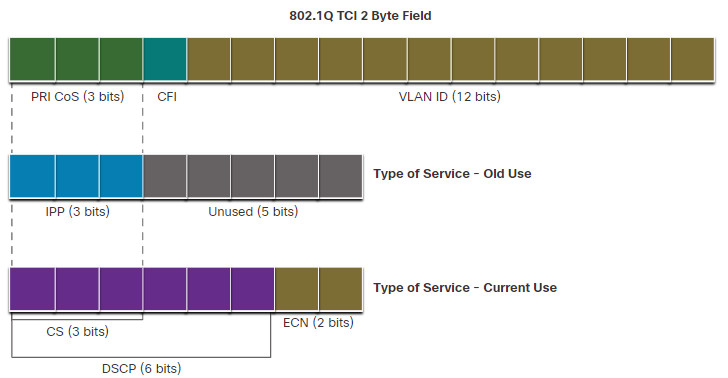


- Công thức AFxy được quy định như sau:

* Ba bit quan trọng nhất đầu tiên được sử dụng để chỉ định lớp.
* Lớp 4 là hàng đợi tốt nhất và Lớp 1 là hàng đợi tệ nhất.
* Hai bit quan trọng nhất thứ 4 và thứ 5 được sử dụng để chỉ định ưu tiên bỏ.
* Bit quan trọng nhất thứ 6 được đặt thành 0.

## **6.9. Bit chọn lớp**

- Vì 3 bit quan trọng nhất đầu tiên của trường DSCP chỉ định lớp, nên các bit này cũng được gọi là các bit Class Selector (CS). 3 bit này ánh xạ trực tiếp vào 3 bit của trường CoS và trường IPP để duy trì khả năng tương thích với 802.1p và RFC 791, như được hiển thị trong hình.Bảng trong hình cho thấy cách các giá trị CoS ánh xạ vào các Class Selectors và giá trị 6 bit DSCP tương ứng.



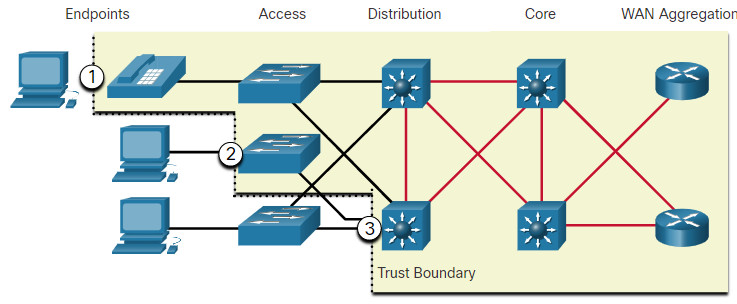
## **6.10. Ranh giới tin cậy**

- Đánh dấu nên xảy ra ở đâu? Lưu lượng truy cập nên được phân loại và đánh dấu càng gần nguồn của nó càng khả thi về mặt kỹ thuật và hành chính. Điều này xác định ranh giới tin cậy, như được hiển thị trong hình.

- Các điểm cuối đáng tin cậy có khả năng và thông minh để đánh dấu lưu lượng ứng dụng thành các giá trị CoS Lớp 2 và / hoặc DSCP Lớp 3 phù hợp.

- Các ví dụ về các điểm cuối đáng tin cậy bao gồm điện thoại IP, điểm truy cập không dây, cổng và hệ thống hội nghị truyền hình, trạm hội nghị IP và hơn thế nữa. Các điểm cuối bảo mật có thể có lưu lượng được đánh dấu ở bộ chuyển mạch Lớp 2.

- Lưu lượng cũng có thể được đánh dấu ở các bộ chuyển mạch / bộ định tuyến Lớp 3.



- Đánh dấu lại lưu lượng, ví dụ, đánh dấu lại các giá trị CoS thành các giá trị IP Precedent hoặc DSCP, thường là cần thiết.

## **6.11. Tránh tắc nghẽn**

- Quản lý tắc nghẽn bao gồm các phương pháp xếp hàng và lập lịch nơi lưu lượng dư thừa được đệm hoặc xếp hàng trong khi chờ được gửi ra khỏi giao diện thoát. Chúng giám sát tải trọng lưu lượng mạng để cố gắng dự đoán và tránh tắc nghẽn tại các điểm nghẽn mạng và internet thông thường trước khi tắc nghẽn trở thành vấn đề. Các công cụ này có thể giám sát độ sâu trung bình của hàng đợi. Khi hàng đợi thấp hơn ngưỡng tối thiểu, sẽ không có lượt bỏ. Khi hàng đợi được lấp đầy đến ngưỡng tối đa, một tỷ lệ nhỏ các gói sẽ bị bỏ qua. Khi vượt quá ngưỡng tối đa, tất cả các gói sẽ bị bỏ qua.

- Một số kỹ thuật tránh tắc nghẽn cung cấp ưu tiên cho các gói nào sẽ bị bỏ rơi. Ví dụ, Cisco IOS QoS bao gồm phát hiện sớm ngẫu nhiên có trọng số (WRED) như một giải pháp tránh tắc nghẽn có thể. Thuật toán WRED cho phép tránh tắc nghẽn trên các giao diện mạng bằng cách cung cấp quản lý bộ đệm và cho phép lưu lượng TCP giảm hoặc giảm tốc trước khi bộ đệm cạn kiệt. Sử dụng WRED giúp tránh thả đuôi và tối đa hóa việc sử dụng mạng và hiệu suất ứng dụng dựa trên TCP. Không có tránh tắc nghẽn cho lưu lượng dựa trên Giao thức Datagram Người dùng (UDP), chẳng hạn như lưu lượng thoại. Trong trường hợp lưu lượng dựa trên UDP, các phương pháp như xếp hàng và kỹ thuật nén giúp giảm và thậm chí ngăn chặn mất gói UDP.

## **6.12. Định hình và chính sách**

- Traffic shaping và traffic policing là hai cơ chế được cung cấp bởi phần mềm Cisco IOS QoS để ngăn ngừa tắc nghẽn. Traffic shaping giữ lại các gói tin dư thừa trong hàng đợi và sau đó lên lịch truyền chúng sau một khoảng thời gian nhất định.

- Kết quả của việc định hình lưu lượng là tốc độ gói tin đầu ra được làm mịn, như thể hiện trong hình.

- Shaping (định hình) ngụ ý sự tồn tại của hàng đợi và đủ bộ nhớ để lưu trữ các gói tin bị trì hoãn, trong khi policing (kiểm soát) thì không. Hãy đảm bảo rằng bạn có đủ bộ nhớ khi bật chức năng shaping. Ngoài ra, shaping yêu cầu một chức năng lập lịch để truyền các gói tin bị trì hoãn sau đó. Chức năng lập lịch này cho phép bạn sắp xếp hàng đợi shaping thành các hàng đợi khác nhau.



- Ví dụ về các chức năng lập lịch là CBWFQ và LLQ:

- Shaping là một khái niệm hướng ra ngoài; các gói tin đi ra khỏi một giao diện sẽ được xếp hàng đợi và có thể được định hình. Ngược lại, policing được áp dụng cho lưu lượng truy cập hướng vào trong trên một giao diện. Khi tốc độ lưu lượng đạt đến tốc độ tối đa đã cấu hình, lưu lượng vượt quá sẽ bị loại bỏ.

- Policing thường được các nhà cung cấp dịch vụ triển khai để thực thi tốc độ thông tin khách hàng theo hợp đồng (CIR).

- Tuy nhiên, nhà cung cấp dịch vụ cũng có thể cho phép vượt quá CIR nếu mạng của họ hiện không bị tắc nghẽn.

## **6.13. Chính sách QoS**

- Chính sách QoS của bạn phải xem xét toàn bộ đường dẫn từ nguồn đến đích. Nếu một thiết bị trên đường dẫn sử dụng chính sách khác với mong muốn, thì toàn bộ chính sách QoS sẽ bị ảnh hưởng.

- Ví dụ, hiện tượng giật hình trong khi phát video có thể là do một switch trên đường dẫn không được đặt giá trị CoS phù hợp.

- Một số hướng dẫn giúp đảm bảo trải nghiệm tốt nhất cho người dùng cuối bao gồm:

* Bật chức năng xếp hàng đợi ở mọi thiết bị trên đường dẫn giữa nguồn và đích.
* Phân loại và đánh dấu lưu lượng càng gần nguồn càng tốt.
* Định hình và kiểm soát luồng lưu lượng càng gần nguồn càng tốt.