QoS

QoS (Quality of Services), network trafiklerinde gecikmeye karşı hassasiyeti bulunan belirli servislerin paketlerine öncelik verilmesini sağlayan teknolojidir.

Ses ve Video Trafiği Karakteristiği

Bir ses haberleşmesinin sağlıklı gerçekleştirilebilmesi için ses paketlerinin maksimum gecikme süresi (delay/latency) 150 ms , gecikme süresindeki değişimin (**Jitter**) maksimim 30 ms, bant genişliğinin minimum 30-128 Kbit arasında olması gerekiyor.

Bir video haberleşmesinin sağlıklı gerçekleştirilebilmesi için video paketlerinin maksimum gecikme süresi 200-240 ms arasında, gecikme süresindeki değişimin maksimim 30-50 ms arasında, bant genişliğinin minimum 384 Kbit – 20 Mbit arasında olması gerekiyor.

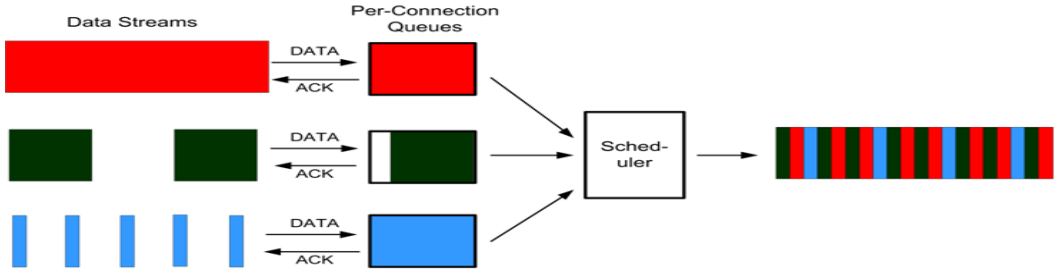
Veri Trafiği Tipleri

- **Interactive / Mission Critical**, borsa işlemleri gibi çok sık güncellenen/değişim gösteren ve gecikmelerin önemli olduğu uygulamaların trafikleridir. Bu nedenle yüksek öncelik verilmesi gereken trafiklerdir.
- **Interactive / No Mission Critical**, çok sık değişim gösteren ama değişiminin çalışma akışına etkisinin çok yüksek olmadığı uygulamalardır. Yine de yüksek veya ortalama öncelik verilmesi gereken trafiklerdir.
- **Non Interactive / Mission Critical**, çok sık değişim gösetermeyen ama çalışma akışı için önemli olan uygulamaların trafikleridir. Bu trafikler için yüksek bant genişliği tahsis edilebilir.
- **Non Interactive / No Mission Critical**, çok sık değişim göstermeyen ve çalışma akışına etkisi olmayan uygulamaların trafikleridir. Öncelik verilmesine gerek yoktur.

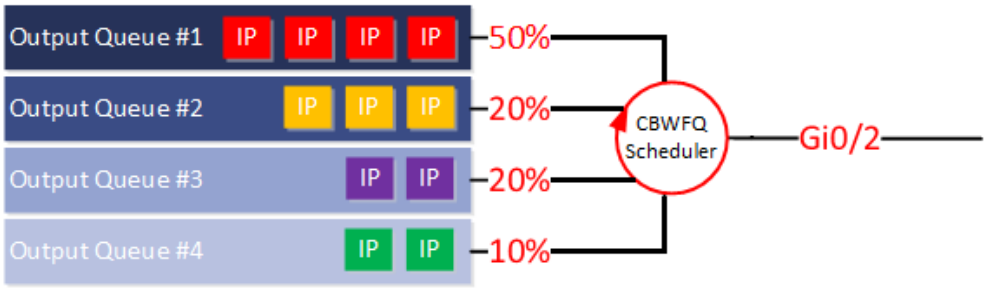
Queuing Strategy

Paketler hedef router veya switche ulaştığında işleme alınmadan önce bir kuyruğa eklenir. Bu kuyruk yapılarında farklı mekanizmalar kullanılabilmektedir. Bu mekanizmalar;

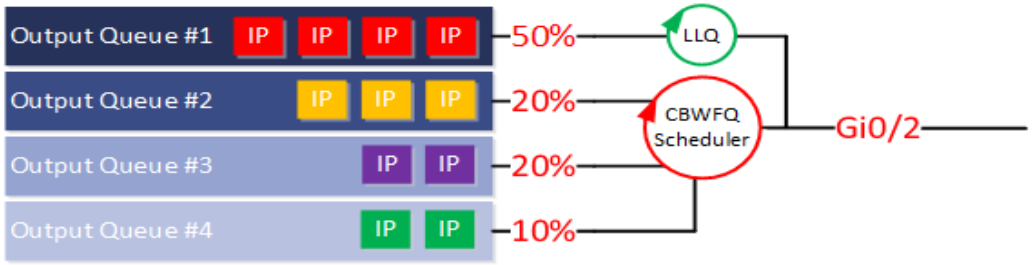
- **FIFO Queuing (First In First Out)**, tek bir kuyrukta bütün paketlerin önceliğinin eşit olduğu ve paketler geldiği sırada işleme alındığı çalışma mekanizmasıdır.
- **WFQ (Weighted Fair Queuing)**, tek bir kuyruk yerine farklı kuyruklar oluşturulmaktadır. Oluşturulan bu kuyruklara eklenecek paket tipleri belirli özelliklere göre belirlenir ve her işlemde oluşturulan kuyruklardan birer tane paket alınır. Bu sayede hiçbir kuyruk trafiğine öncelik verilmeden paketler geldiği sırada işlenmiş olur. Bu kuyruklama stratejisi daha çok Seri interface gibi yavaş hatlarda kulanılmıştır. Günümüzde pek tercih edilmemektedir.



- **CBWFQ (Class Based Weighed Fair Queuing)**, birçok kuyruk/class oluşturulur ve ACL'ler kullanılarak bu kuyruklara eklenecek paket tipleri belirtilir. Oluşturulan kuyruklara ise belirli bant genişliği atanarak kuyruklara (önem seviyesine bağlı olarak) eklenen paketlere öncelik verilmesi sağlanır. Günümüzde yaygın kullanılan kuyruklama stratejisidir.



- **LLQ (Low Latency Queuing)**, CBWFQ mekanizmasının ses trafiğine öncelik verilen versiyonudur. Ses trafiğine LLQ adı verilen ayrı bir kuyruk yapısı kullanarak gerçekleştiriyor (**bu kuyruğa ses dışında farklı trafiklerin koyulması da sağlanabiliyor**). Eğer ki ses paketlerinin bulunduğu kuyrukta bekleyen paket varsa, diğer kuyruklar bekletilerek ses paketlerine öncelik verilmesi sağlanıyor. Ses paketlerinin bulunduğu kuyrukta paket yoksa diğer kuyruktaki paketler kendilerine ayrılan bant genişliğince paketleri iletmeye devam ediyor. Yani LLQ kuyruğuna gelen paketler bekletilmeden doğrudan işleme alınıyor.



QoS Modelleri

- **Best Effort**, herhangi bir QoS mekanizması kullanılmıyor.
- **Intefrated Services (IntServ)**, bu model için **RSVP** (Resource Reservation Protocol) protokolü kullanılıyor. RSVP protokolü uçtan uça veri haberleşmesi yapılamdan önce ihtiyaç duyulacak bant genişliği rezerve ediliyor. Her bağlantı için ayrıca hat rezerve edilmesi gerekiyor. Bu nedenle çok ölçeklenebilir bir model olmadığı için günümüzde pek tercih edilmemektedir.
- **Differentiated Services (DiffServ)**, her cihaz üzerinde istenilen QoS konfigürasyonu yapılabilmektedir /her cihaz birbirinden bağımsız çalışmaktadır). Olumsuz yanı ise her cihazda ayrıca QoS konfigüürasyonu yapılması gerekiyor. Günümüzde yaygın olarak kullanılmaktadır.

QoS Uygulama Teknikleri

Bir QoS mekanizması uygulanmak istendiğinde öncelikle paketlerin etiketlenmesi gerekiyor. Bu etiketler sayesinde paketler routerlarda sınıflandırılabiliyor ve belirli paketlere öncelik verilebiliyor.

Paketlere L2 veya L3 cihazlarda etiket bilgisi eklenebiliyor. Bu işlem router veya switch üzerinde ACL'ler kullanarak yapabilirken aynı zamanda paketin çıkış yaptığı kaynak cihazdan (örneğin bir FTP sunucudan paket çıkarken veya ip telefonlardan paket çıkarken) etiketli çıkması da sağlanabiliyor. Ayrıca paketler **NBAR (Network Based Application Recognition)** mekanizması kullanılarak paketlerin uygulama trafiğine göre ayırt edilip etiketlenmesi de sağlanabiliyor.

L2'de Ethernet, Wifi veya MPLS teknolojilerinde kullanılan paketler için öncelik verilebilirken, L3'de IPv4 veya IPv6 protokolleri için 3 bit veya 6 bitlik öncelik değerleri verilebiliyor.

Etiketleme işlemi için;

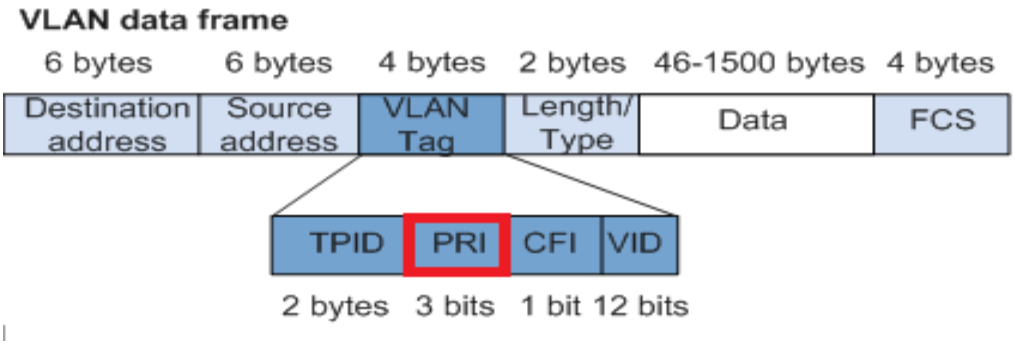
- Ethernet Teknolojisinde -> L2 -> **CoS (Class of Service)** -> 3 bitlik etiket ekleniyor
- Wifi Teknolojisinde –> L2 -> **TID (Wi-fi Traffic Identifier)** -> 3 bitlik etilet ekleniyor
- MPLS -> L2 -> **EXP (Experimental)** -> 3 bitlik etilet ekleniyor
- Ipv4 and IPv6 -> L3 -> **IPP (IP Predence)** -> 3 bitlik etilet ekleniyor
- Ipv4 and IPv6 -> L3 -> **DSCP (Differentialed Services Code Point)** -> 6 bitlik etilet ekleniyor

L2 QoS

Normalde Ethernet başlığında QoS hizmeti için kullanılan herhangi bir alan bulunmuyor. Bu nedenle L2'de paketlere öncelik verebilmek için VLAN teknolojisi kullanılıyor. VLAN teknolojisinde switchler arası bağlantılarda kullanılan portlar Trunk moduna alınıyordu. Trunk moduna alınan portlarda, paketlerin switchler arası anahtarlanma sürecinde ait oldukları VLAN'ları ayırt edebilmek için Ethernet başlığına ek olarak 802.1q etiketi ekleniyordu. Bu etiket içerisinde bulunan **PRI** bitleriyle (CoS bits) belirli VLAN trafiğine (Voice VLAN gibi) öncelik verilmesi sağlanabiliyor.

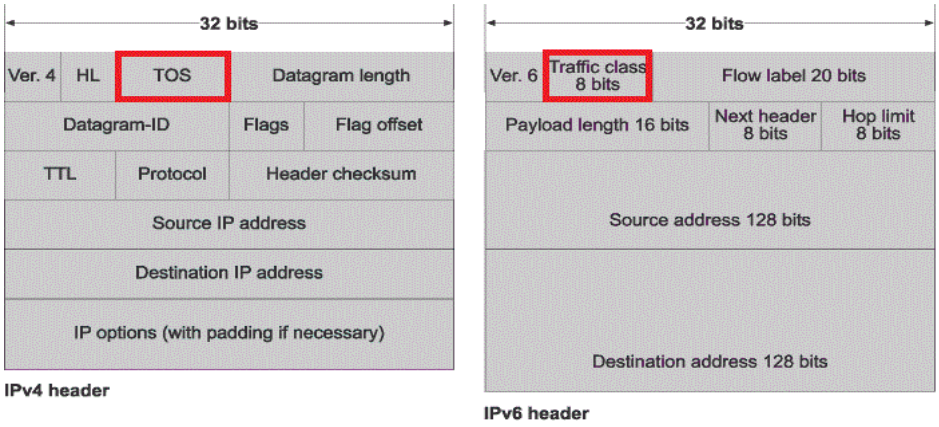
VLAN'lara öncelik verebilmek için 0-7 arasında bir değer tanımlanıyor. Bu değerler;

- 0 -> **Önceliksiz veri trafiği için kullanılıyor**
- 1 -> **Orta öncelikli veriler için**
- 2 -> **Yüksek öncelikli veriler için**
- 3 -> Çağrı sinyalleşmesi için
- 4 -> Video konferansı için
- 5 -> Ses trafiği için kullanılıyor
- 6 ve 7 -> Rezerve



L3 QoS

IPv4 başlık bilgisi için **ToS (Type of Service)** adında bir alan kullanılırken IPv6 başlık bilgisinde **Traffic Class** adında alan bulunuyor. Bu alanlar sayesinde, pakete eklenen etiket bilgisi iletimi boyunca (uçtan uca) korunuyor. Yani L2 QoS konfigürasyonunda paketlere öncelik bilgisi switch üzerinde tanımlanan VLAN konfigürasyonuna göre belirleniyordu (bu aynı zamanda bir VLAN'in her switch üzerinde farklı öncelik değerlerine sahip olabilmesi demektir). Bu durum paketin, switchler arası her geçişte etiket bilgisinin switch üzerinde tanımlanan konfigürasyonlar doğrultusunda yenilenmesi anlamına geliyor. L3 QoS uygulamasında ise etiket bilgisi başlık üzerinde olduğu için paket, kaynaktan hedefe iletimi boyunca her router üzerinde aynı etiket bilgisiyle taşınıyor.



|→ **TOS ve Traffic Class alanlardaki 8 bitin tamamı kullanılmıyor. Kullanılan teknolojiye göre sadece ilk 3 veya ilk 6 biti kullanılabiliyor.** Bunun nedeni L2'de 802.1q ile belirlenen öncelik bilgisi (CoS) ile L3'de bulunan öncelik bilgisi (**IP Precedence**) arasında değerlerin aktarılabilmesidir (CoS için 0-7 arası değer alabiliyor –> 3 bit). Bu sayede paket L3'den L2'ye veya L2'den L3'e geçiş yapsa da öncelik bilgisi değişmeden paketin uçtan uca iletilmesi sağlanabiliyor (Sonuç olarak paket kaynaktan çıktığında LAN içerisinde switchler üzerinde yönlendirilecek. Ardından routerlar arasında yönlendirildkten sonra yine hedef istemciye erişebilmek için LAN içerisinde switchler üzerinden iletilmesi gerekecek).

|→ L3'den L2'te geçiş için 3 bitin yeterli olmadığı görülünce başlıklarda bulunan 8 bitin ilk 6 biti kullanılmaya başlanıyor (**Differentiated Services Code Point**). Bu mekanizmada L3 öncelik değerinin L2'ye de uygun olabilmesi için (L2-L3 arasında aktarılabilmesi);

- **AF (Assured Forwarding)** ve **EF (Expedited Forwarding)** olmak üzere trafik ikiye ayrılıyor. Burada EF sadece ses trafiğini temsil etmek için kullanılıyor ve her zaman öncelik veriliyor.
- Kullanılan 6 bitin öncelikle ilk 3 bitine bakılıyor. L2'de olduğu gibi ilk 3 bite göre trafik **sınıflandırılıyor**. (Örnek olarak; ilk 3 bit : 5 ise -> Ses trafiğidir, 4 ise -> video trafiğidir gibi).
    - AF1 -> 001
    - AF2 -> 010
    - AF3 -> 011
    - AF4 -> 100
    - EF -> 101
- İlk 3 bitten sonraki 3 bitin sadece ilk 2 biti kullanılarak, oluşturulan sınıflar için alt sınıflar tanımlanıyor ve trafikler bu şekilde drop edilme durumlarına göre **önceliklendiriliyor**. Yani her AF sınıfı için 3 tane alt sınıf oluşturuluyor (Burada son bit daima 0 oluyor ve kullanılmıyor).
    - 01 -> 1 - Yüksek öncelik
    - 10 -> 2 -Orta öncelik
    - 11 -> 3 - Düşük öncelikli
- **Sadece EF için alt sınıfı oluşturulmuyor (EF -> 010 110).**
- Örnek olarak oluşan AF1 sınıfının alt sınıflarını listeyecek olursak;
    - AF11 -> 001 010 (AF1 sınıfı için, geçişte önceliği 1. sırada, drop edilme önceliğinde 3. sırada)
    - AF12 -> 001 100 (AF1 sınıfı için, geçişte önceliği 2. sırada, drop edilme önceliğinde 2. sırada)
    - AF13 -> 001 110 (AF1 sınıfı için, geçişte önceliği 3. sırada, drop edilme önceliğinde 1. sırada) şeklinde olacaktır. Benzer şekilde AF2 ,AF3, AF4 içinde geçerlidir.

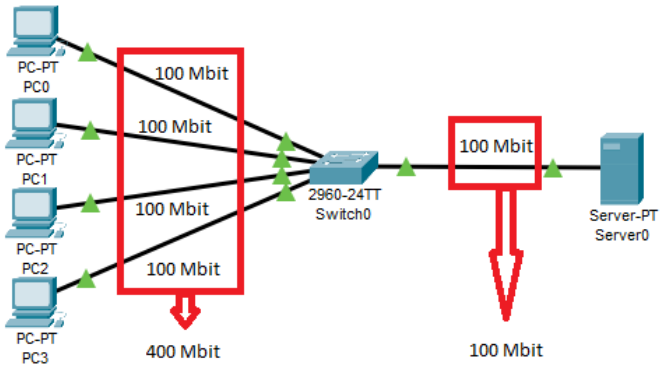
- Oluşturulan bu sınıfların toplamında oluşan değerlere ise **DSCP** değeri denilmektedir. DSCP değeri ile AF değerleri birbiri yerine kullanılabiliyor. Bu nedenle dönüşümü önemlidir. Örnek;
    - AF11 -> 001010 -> 10 (DSCP)
    - EF –> 101110 -> 46 (DSCP)

Özetle L3 QoS yapabilmek için gerçekleştirilebilecek üç seçenek bulunuyor. İlk seçenek Best-Effort (BE) ile hiçbir trafiğe öncelik verilmeden veri iletimi sağlamak (DSCP değeri 0 oluyor) . İkinci seçenek ise AF kullanılarak paketler farklı sınıflarda farklı öncelik değerlerinde iletilebilir. Üçüncü seçenek ise sadece EF kullanılarak ses paketlerien öncelik verilebiliyor.

Congestion ve Aggregation

Congestion, iki kaynağın maksimum kapasitesindeki farklılıklardan dolayı diğer donanımların potansiyelini sınırlamasına verilen isimdir. Congestion;

Aggregation, bir kaynağa/bağlantıya sınırlarından fazla işlem yüklendiğinde oluşabiliyor.



- Speed Mismatched, karşılıklı portların bant genişlikleri aynı olmadığı zamanlarda oluşabilir.
- LAN to WAN, LAN'da kullanılan bant genişliğii ile ISP'den satın alınan internet bağlantısının bant genişliği farklı olduğu zaman oluşabiliyor.

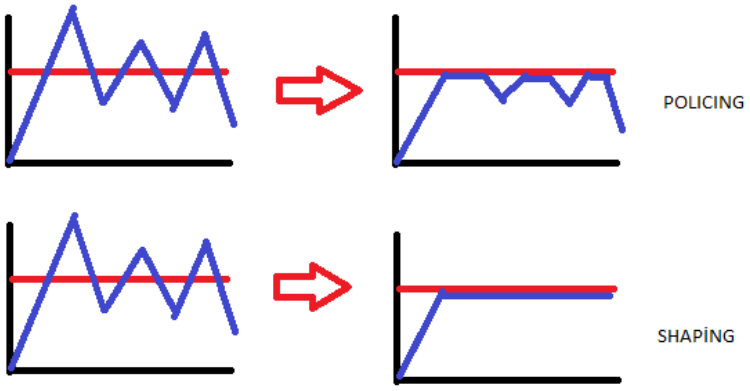
Congestion için alınabilecek iki ölem bulunuyor;

- **Düşük öncelikli paketler drop edilir**.
- **WRED** (Weighted Rondım Early Detection), kuyruk dolmaya başladığında (dolmadan önce) kuyrukta bekleyen paketler arasında random paketler seçilerek drop edilir. Bu sayede iletimde TCP kullanıldığı için kaynak cihaza kaybolan paketler yeniden talep edilir (bu sayed aralardaki paketlerin seçilip drop edildiği için her drop edilen paket yeniden taklep edilir yani bağlantı koparılmadığı için yeniden oturum kurulmasına gerek kalmaz). Paketler yeniden talep edildiğinde kaynak cihaz, paketlerin drop edilmeye başladığını anlayarak paketlerin iletim hızını düşürür (Bu hızı kaybolan paket oranına göre belirliyor). Bu sayede hedef cihazda **Tail Drop** durumu yaşanmaz.

Shaping and Policing

Policing, ISP'nin kullanıcıya verdiği bant genişliğinden yüksek trafik oluşturduğunda bant genişliğin aşan trafiğin doğrudan drop edilmesine deniliyor.

Shaping, kullanıcının internet trafiğini düzensiz göndermek yerine bekletilerek ISP tarafından kullanıcıya verilen bant genişliğini dolmadığı zamanlarda göndermesine deniliyor.

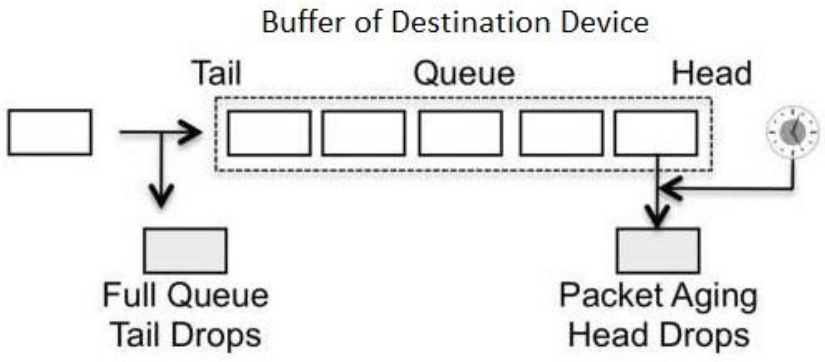


NOT

- Switch veya routerlarda kullnaılan kuyrukalma stratejilerini görebilmek için "**sh ip interface <Interface ID>**" komutu kullanılabilir.
- Shaping çok tercih edilmiyor çünkü ses paketleri bekletidiğinde Jitter oluşuyor.
- Konfigürasyonlar ve daha fazlası için "**CCNP – 09 – QoS**" notlarına göz atabilirsin.

Terminolojiler

- Tail Drop, TCP mekanizmasında birim zamanda iletilen paket sayısının hedef cihazın işleyebildiğinden daha yüksek olduğu durumlarda hedef cihazın TCP paketlerini tuttuğu kuyruk dolacaktır. Bu durumdan sonra gönderilen paketler kuyruğa eklenemeyeceği için paketler drop edilecektir. Buna **Tail Drop** denilmektedir. Kurulan TCP bağlantısı ise RST bitleriyle aniden sonlandırılacaktır.



- Delay,/Latency, bir bağlantıda paketin kaynak cihazdan hedef cihaza ulaşana kadar geçen süre olarak tanımlanabilir.
- RTP (Real-Time Trandfer Protocol) , internet üzeirnde ses iletimi için kullanılan paket formatıdır.
- RTPS (Real-Time Streaming Protocol), internet üzeirnde video iletimi için kullanılan paket formatıdır. UDP 554 gibi portlar kullanılmaktadır.

- Jitter, ses paketlerinin hedefe ulaşırken gösterdikleri gecikme sürelerindeki farklılıklara verilen isimdir.
- De-Jitter, farklı gecikme sürelerinde gelen ses paketleri sabit bir gecikme süresinde sabitleyerek işliyor.
    - Örnek olarak 100 ms, 70 ms, 110 ms, 130 ms gecikmelerle gelen paketler için paketler arasında 60 ms'lik periyotlar belirliyor. 100 ms'de gelen paketi 30 ms daha geciktiriyor, 70 ms gecikmeyle gelen paketi 60 ms geciktiriyor ve bütün paketler arasındaki gecikme süresini örnek olarak 130 ms'ye çekebiliyor. Bu sayede jitter oluşumunu önlüyor.
    - De-jitter kullanımı paket kayıplarına neden olabiliyor. UDP protokolünde kaybolan paketler yeniden talep edilmediği için hedef cihazda **DSP** (Digital Signal Processor) kullanılarak hedefe ulaşan bir önceki ses paketi yinelenerek boşluk dolduruluyor. Bu sayede kullanıcı yaşanan kesintileri hissetmeden iletişimine devam ediyor.