Traffic Shaping

https://campus.barracuda.com/doc/46209125/

Limited network resources make bandwidth prioritization a necessity. The Barracuda NextGen Firewall F-Series enables traffic shaping to prioritize network resources according to a number of factors such as time of day, application type and user identity. Traffic shaping supports the following features:

- **Data Traffic Classification –** Classify traffic into different bandwidth allocation priorities.
- **Traffic Prioritization –** Increase the bandwidth and lower the latency of important traffic.
- **Bandwidth Partition –** Specify bandwidth limits for certain traffic types.
- **Network Overflow Protection –** Prevent protocols without flow control mechanisms from congesting the network.
- **Dynamically Adjusted Shaping –** Adjust traffic to dynamic factors such as time of day or download volume.
- **Shaping of VPN Transports** – Adjust VPN tunnel settings to make sure remote locations are assigned enough bandwidth for business critical web applications.

Traffic Classification

In addition to security classification, you can use the access rule set to classify network traffic for traffic shaping. Classification by the access rule set is *static* - it does not change after the session is initiated unless you select the session in the rule set and change the QoS band. For classification according to *dynamic* factors such as the time of day or download volume, the Barracuda NextGen Firewall F-Series provides the QoS profile. To connect the rule-based *static* classification (session) and traffic shaping, the QoS band is used.

Network data can be shaped in the following ways:

- **Outbound shaping –** The traffic is shaped before it is delivered to a network interface.
- **Inbound shaping –** The traffic is shaped after it is received by a network interface.

QoS Profiles

When configuring the QoS profile for traffic shaping, an expandable “tree” of *virtual interfaces* is added to the network interfaces where traffic must be shaped. A virtual tree consists of a ***root*** *virtual interface* that can be attached to a *real* network interface, and a number of subnodes. When assigning a virtual tree to a physical network interface, you can enable and specify the rates for inbound and/or

outbound traffic shaping. The outbound and inbound rate of a virtual interface is ignored when the **QoS Band** policy in the corresponding access rule is set to **No-Shaping**.

For more information on configuring virtual trees, see How to Create a QoS Profile.

Virtual Interfaces

The main purpose of a virtual interface is to shape and reduce traffic throughput from different sources to an available bandwidth according to priorities. Data is transmitted over the virtual interface and then forwarded inbound or outbound according to the traffic shaping settings. The most important characteristics of a virtual interface are:

- **A limiting bandwidth –** This limit specifies the maximum data rate that is available for the virtual interface itself.
- **A priority weighting** (*high, medium* or *low*) – This priority determines how the available bandwidth is partitioned if **more** data arrives than the bandwidth limit allows.

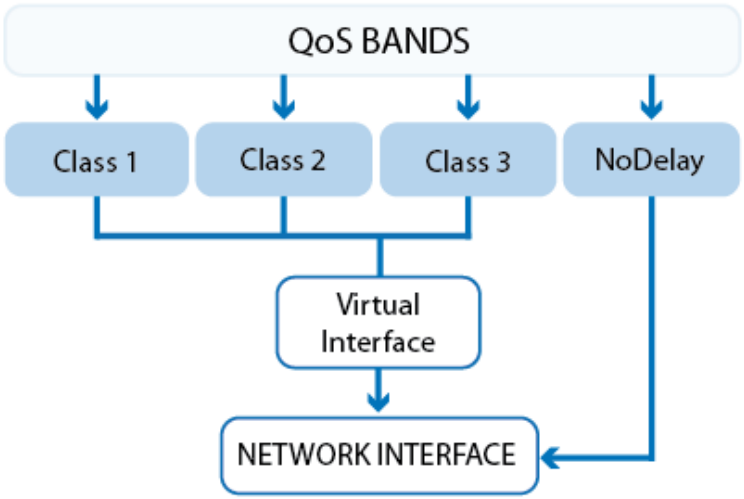
> Partitioning is never static. For example, if all available traffic has a *low* priority, it is assigned the whole bandwidth available on the subinterface. The Weighted Random Early Drop (WRED) queue management algorithm is used for prioritization.

To specify the bandwidth ratio of the traffic being propagated by a virtual node, you can select three priorities: **class1**, **class2**, and **class3**. For high priority traffic that should not be restricted to a bandwidth limit, you can assign the **NoDelay** priority. The **NoDelay** priority should only be used in well defined circumstances, to avoid crowding out the other traffic.

The bandwidth ratio is enforced in two steps:

1. **Virtual Interfaces:** Depending on the source, traffic is assigned and processed by the assigned virtual interface. Traffic is shaped according to the bandwidth ratios set for each interface. This does not apply to **NoDelay** priority.
2. **Virtual Root Interface:** The virtual root interface is used to ensure that the combined traffic from all virtual interfaces do not exceed the global limits.

Example Setup



Traffic Prioritization

The QoS band evaluates and prioritizes traffic (**high**, **medium**, or **low**). It continually evaluates an IP packet's ToS (type of service), current data volume, and the absolute time domain. With QoS bands, you can construct routing-dependent traffic shaping schemes. For example, you can configure an Internet connection in normal and fallback (ISDN) operation. For more information, see the Traffic Shaping examples.

QoS bands prioritize traffic flow in the shaping tree (together with the virtual interface). The connection between the traffic shaping engine and the firewall is done by the **shaping connectors**. There are eight connectors available for out-of-the-box traffic management: **Interactive**, **VoIP**, **Business**, **Internet**, **Background**, **LowPrio**, **LowestPrio**, and **Choke**.

- **VoIP** will always be given first priority. The same applies for **Interactive** which is limited to 90% of the overall available bandwidth, thus always leaving at least 10% for VoIP traffic. The bandwidth which is *not* used by **VoIP** or **Interactive**.
- The bandwidth ratio of **Business : Internet : Background** is **10:2:1** for residual bandwidth which is *not* used by **VoIP** or **Interactive**. In addition, **Internet** has a built-in size limit of 10 MB after which a session is downgraded to **Background**, thus receiving a smaller bandwidth ratio after the limit is exceeded.
- The **LowPrio** virtual interface is limited to 5% of the overall available bandwidth. The bandwidth ratio of the **LowPrio : LowestPrio : Choke** shaping connectors is **10:2:1**.
- The **Choke** virtual interface is limited to 0.1% of the overall bandwidth. These shaping connectors are ideally used to slow down somewhat unnecessary traffic and applications which cannot be completely blocked.

For more information, see How to Create and Apply QoS Bands.

Access Rules

In order to use a shaping connector, you must refer to it in a access rule. In the rule configuration, you can select between *forward* and *reverse*:

- **forward –** This direction is defined by traffic that is generated by the session initiator (client).
- **reverse –** This direction is defined by traffic that is generated by the responder (server).

For each traffic type, shaping may be configured differently. For instructions on how to create a QoS Band and apply traffic shaping to a access rule, see How to Create and Apply QoS Bands.

TCP Flow Control

Because traffic shaping affects packet delivery, it also affects the TCP flow control mechanism. Ideally, the TCP flow control reduces its flow rate to an amount where the shaping mechanism is no longer forced to discard packets. This is only possible if the traffic shaping mechanism can delay packets long enough for the TCP flow control to detect a smaller bandwidth by measuring longer RTTs (round trip times). A longer delay involves larger queue sizes that should be considered when configuring virtual interfaces. Long delays also result in larger latency values, which might be unwanted for other protocols. Therefore, in the case of mixed TCP and other protocol traffic, consider using separate traffic shaping nodes for TCP with different queue size settings.

It is also the TCP flow control mechanism which makes the priority weights approximate values. For example, there are 20 TCP sessions that are all trying to receive the maximum bandwidth possible—where 10 are classified as high and 10 are classified as medium priority. If you configure a ratio of 1:2 for the two priorities, you will observe this ratio when measuring the output for the two priorities. But if you change to setup to 1 high priority TCP session and 39 medium TCP sessions, the results change. The single TCP session gets less bandwidth than expected, because the flow control of the 39 TCP sessions generates more traffic while trying to find an optimum rate than the single high priority session. So to favor a small number of TCP sessions over a large number of unprivileged TCP sessions, you should anticipate a larger ratio in order to get the desired output ratio.

Figures