SANDVINE

ActiveLogic Traffic Shaping Techniques

BEST PRACTICES

During solution design, consider the following:

1. Group applications with the same latency requirements together. It will help to ensure that proper latency goal values are used, and QoE remains at high level.
2. Select applications for traffic management grouped by Sandvine standard Signature categories for simplified policy design.
3. Define in advance:
   a. Applications that need to be identified and blocked/shaped all the times due to business needs.
   b. Minimum guaranteed bandwidth for each application set even when congestion exists in the network.
   c. WFQ weights for each application, guaranteeing bandwidth reservation while borrowing for application set or while competing for the bandwidth within ShapingObject responsible for bandwidth management for application set.
   d. Latency goals for each application set, guaranteeing latency-sensitive data won't spend too much time in queue waiting to be de-queued.
   e. Fair Factors for different tiers.

INTRODUCTION

Sandvine's ActiveLogic, a hyperscale data plane and policy enforcer, is a key network element for operators to optimize network resources. In order to manage the ever-increasing band-width demands, operators need to rethink the "dumb data pipes" approach, and adopt an application-centric view of the network. By taking this approach, operators can bring surgical-level optimization, translating into better quality of experience and resource allocation, versus using strict policies for policing or blocking traffic.

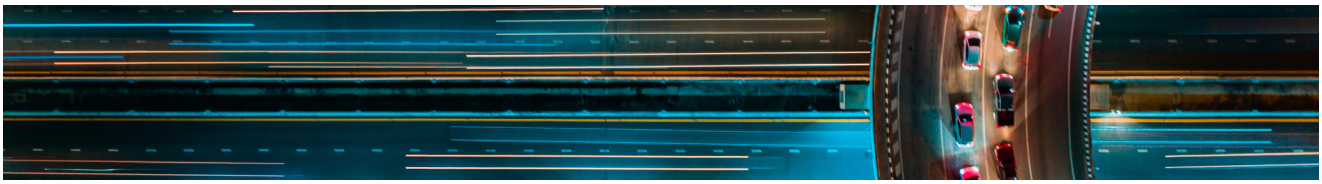
Typically, network operators want to manage/optimize network and service resources by using enhanced QoS and shaping techniques and adjusting subscriber behavior as per the service level agreement. Enhanced QoS mechanisms and traffic shaping improves packet delivery using deeper queues, forwarding and dropping regulations, traffic classification and prioritization, minimum guaranteed and peak bandwidth limits. This is the most suitable solution for bandwidth management in a dynamic network environment, ensuring good performance for higher priority subscribers and applications.

ACTIVELOGIC TRAFFIC SHAPING

Sandvine's ActiveLogic offers the most advanced and flexible shaping functions available in the market today. Shaping tools perform their duties by delaying some or all of the packets in selected traffic streams in order to bring those streams into compliance with traffic profiles. Typically, a shaper has a finite-sized buffer, and packets are discarded only when there is insufficient buffer space to hold the delayed packets. The main advantage of shaping over policing is a reduced need for retransmissions, since packets are buffered instead of immediately dropped the moment the defined rate is exceeded.

For shaping to yield the desired results, traffic must be precisely selected for shaping first. The flexibility of the selection mechanism plays an essential role on the effectiveness of the shaping solution. Traffic selection in ActiveLogic takes advantage of following:

- Sandvine's best-in-class DPI engine, enabling traffic selection (and thus control) based on all of the information extracted from the flows. This includes but is not limited to IP addresses, prefixes and ranges, transport protocols, ports and port ranges, VLAN tags, VLAN priorities, MPLS labels, DSCP markings, services and applications detected by Sandvine's DPI engine, connection state information, and L7 protocol headers extracted by ActiveLogic.
- Sandvine's BGP integration enables visibility of BGP paths and communities, which can equally be used for traffic selection.
- Sandvine Intelligent Feeds adds content classification information to the flows, also selectable for traffic matching.
- Traffic selection can also observe the time of day, allowing for different policies to be active during different parts of the day (e.g., peak hours and off-peak hours)



- When deployed in combination with Maestro Policy Engine, in addition to all of the above, traffic can also be selected and controlled based on locations, access network types, network nodes, geographic regions, subscriber identity, subscriber tiers, or any other attributes that can be mapped back to the subscriber.

Sandvine Use Cases and Shaping

Sandvine offers operators many use cases with traffic optimization strategies, including shaping, to complement their capacity investments but also extend the lifetime of existing infrastructure while enhancing subscriber QoE at a reduced cost. These strategies respond directly to many converging network trends, including explosive growth in per-subscriber traffic, uneven subscriber and application usage, and increasing frequency and duration of congestion. Following table shows how ActiveLogic shaping techniques help operators across different use cases.

Fair Usage and Congestion Management	Using various shaping techniques and flexibility of configurations, it offers most precise fair usage and congestion management solution, enabling operators to define the applications or application categories, and/or subscriber tiers to be prioritized during times of congestion.
Video Streaming Management	This use case offers consistent delivery of high-quality video experience to all users while dramatically reducing bandwidth and protecting other services from disruptive video traffic pattern. Sandvine enforces intelligent shaping on a per-stream basis, ensuring fairness and bandwidth efficiency without compromising quality. More advanced approaches are available: real-time congestion awareness, service plans, device types, and other factors for extremely precise optimization.
Heavy User Management	While using a combination of fair usage and usage management capabilities, this use case allows precise management of heavy users, which is crucial for preserving QoE for majority of the users and optimizing overall current serving capacity of the network.
Wholesale and Peering Link Management	In order to manage cost, this use case allows small network operators to enforce peak bandwidth levels on a per-peering link basis while maximizing the value of the traffic carried over each link. Incorporating state-of-the-art shaping features, this use case offers smoothed packet output rate which uses link capacity efficiently while ensuring good customer QoE.

SHAPING CONCEPTS

ShapingObjects

ShapingObjects represent the actual traffic queues within ActiveLogic, and have various criteria and limits that are defined in shaping rules. .

Shaping Rule

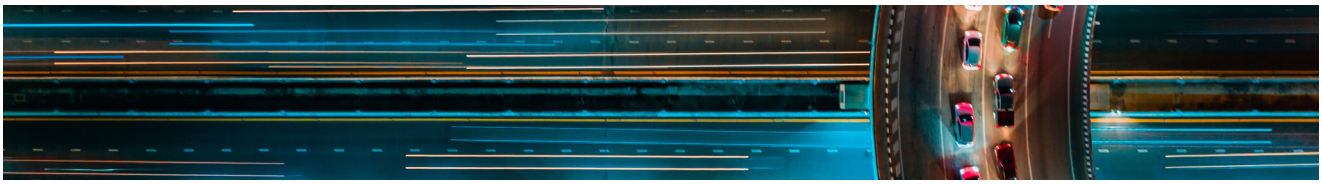
Shaping rule connects a set of objects for traffic identification (matching conditions) to a ShapingObject, limiting the connections matching the conditions according to the ShapingObject. Additionally, a shaping rule can use multiple ShapingObjects.

Borrowing

Borrowing allows shaping rules to borrow bandwidth from shaping objects when their "primary" shaping object is full (provided there is bandwidth left in the object to borrow from).

Parallel Queuing

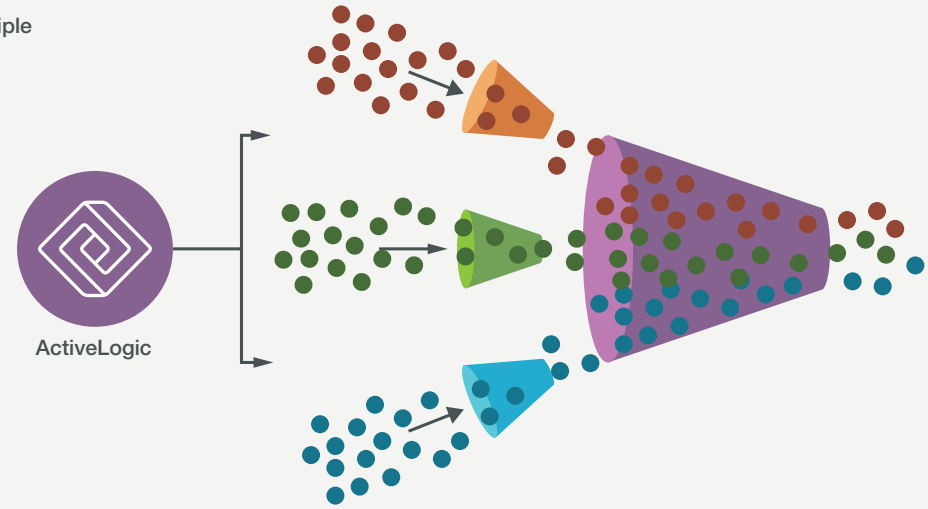
A key aspect ActiveLogic is the way packets are queued for shaping. Since a connection can concurrently match multiple shaping rules, packets from a given connection are queued into multiple queues in parallel as shown in **Figure 1**. In ActiveLogic, a packet can be queued to any number of queues in parallel and each queue is assigned a queue length and a bandwidth. Only when the packet is de-queued from all queues is it actually transmitted on the wire.



Parallel Queueing allows you to designate a certain bandwidth for a user (e.g., a given subscriber gets 50 Mbps from its subscription) and a certain bandwidth for a distribution site (e.g., the bandwidth between distribution and core is 10 Gbps), and then a certain bandwidth for a whole access site (e.g. the bandwidth between access to distribution is 1 Gbps).

Figure 1

A packet can be queued to multiple queues in parallel



Active Queue Management

Active Queue Management (AQM) is an industry standard algorithm for how packets are queued, and which packets are dropped (when drops are necessary). AQM can accomplish early drops (or mark packets with an explicit congestion notification), curtailing application flow rates before hitting congestion.

In ActiveLogic, a ShapingObject is essentially a queue and there is a queue for each priority level in each ShapingObject. When the packet has been evaluated against all matching ShapingObjects, it ends up being a forwarded packet, a tail dropped packet, or a potential early drop. The drop decision for a potential early drop is made by the AQM chosen. These preemptive drops will cause an adjustment in the TCP window by well-behaving TCP stacks at the end points.

The currently available AQM options are BLUE, per-connection BLUE, and CoDel. It is also possible to disable using AQM for a ShapingObject, in which case the queue works in simple “tail-drop” mode, similar to a policer.

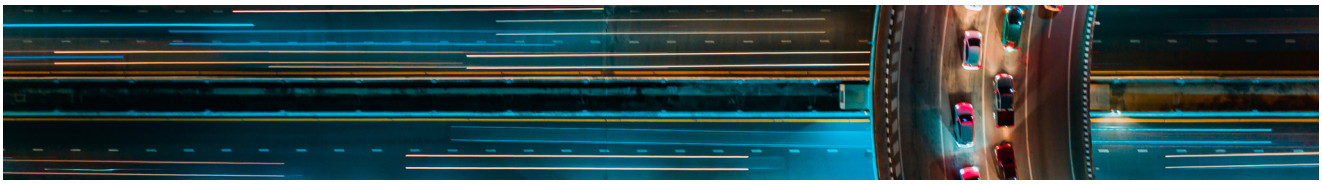
Priority Dequeuing

A shaping rule assigns a priority number to a connection (i.e., VoIP =1, HTTP = 2, P2P = 3 etc.). Since the packets in the ShapingObject queue are dequeued and transmitted in order, the priority can be enforced while maintaining the bandwidth limits set by the ShapingObject.

Priority dequeuing is accomplished by allowing all priority levels to dequeue a certain amount (queue goal) in a burst. ActiveLogic creates predictable behavior by assigning only one priority to a single connection – regardless of how many rules it matches.

Weighted Fair Queuing

With WFQ, a ratio can be set for each priority (from 4-9) for how much of the ShapingObject capacity the priority is allowed to consume – strict priority is used for priorities from 1 through 3 strict. By default, priority is handled by allowing bursts up to the queue goal.



WFQ is used to avoid starving connections that receive lower (higher number) priority. It configures a percentage for various priority levels and the percentage is the ratio of the burst capacity reserved for this priority level. In the case that a priority level does not consume its reserved burst capacity, it is consumed by other priority levels requesting it in strict priority order (i.e., if priority 5 does not burst and priority 3 and 4 both want all of it, priority 3 gets all of it).

WFQ allows constructing a rule set where a certain application from an application set (e.g., YouTube) is given a priority level and then entitled to a defined ratio of the capacity. It will ensure that traffic of YouTube can have more bandwidth reservation compare to Netflix within application set "Video OTT".

Virtual Queueing

As stated earlier, shapers perform their duties by delaying some or all of the packets in selected traffic streams in order to bring those streams into compliance with traffic profiles. For delay sensitive traffic, a ShapingObject with the virtual queueing option enabled can be used to eliminate the queuing delay, preserving QoE, while ensuring throughput is kept under the set limit. When this setting is enabled, packets are not actually queued, but the drops by the AQM are still performed and the ShapingObject behaves similar to a policer.

FAIR USAGE AND SHAPING

Sandvine's Network Optimization use cases focus on fairly sharing network resources among customers, ensuring access to some level of bandwidth during times of congestion and during peak hours. This fair usage approach helps operators achieve a desired level of user experience and throughput efficiency. Sandvine Network Optimization use cases take advantage of best-in-class traffic shaping with advanced queue management, industry-leading application intelligence, and broad awareness of various subscriber attributes in providing contextual, real-time traffic management.

ActiveLogic can use the following shaping mechanisms when doing traffic management in the network:

- Subscriber Fair-Split – split bandwidth equally in real time between active subscribers
- Subscriber Fair-Split Weighted by Tier/Group (Fair Factor) – split bandwidth unfairly or unequally between active subscribers belonged to different tiers/groups in real time
- Subscriber Fair-Split with Application Awareness – split bandwidth fairly or unfairly between application groups
- Strict Priority Shaping – a priority assigned by the shaping rule to a connection impacting dequeuing and transmission order while maintaining the bandwidth limits
- Weight Fair Queue Shaping – enable setting a ratio for each priority from 4-9 through for how much of the ShapingObject capacity the priority is allowed to consume

Subscriber Fair-Split

Fair-Split shaping is designed to maximize network utilization while ensuring congestion management is transparent if there is no congestion. With this approach, each active subscriber (or any other element like subscriber tier) is given an equal portion of the available bandwidth on the network at any instance. Unused bandwidth is made available to the rest of the subscribers on that network segment, ensuring maximum efficiency from the existing network infrastructure. During times of congestion, Fair-Split prevents heavy users from negatively impacting the quality of experience for other subscribers. **Figure 2** shows how the Fair-Split and borrowing concepts interwork to achieve fairness.

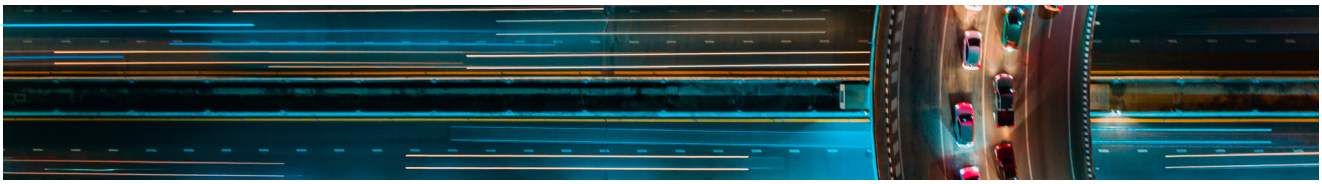
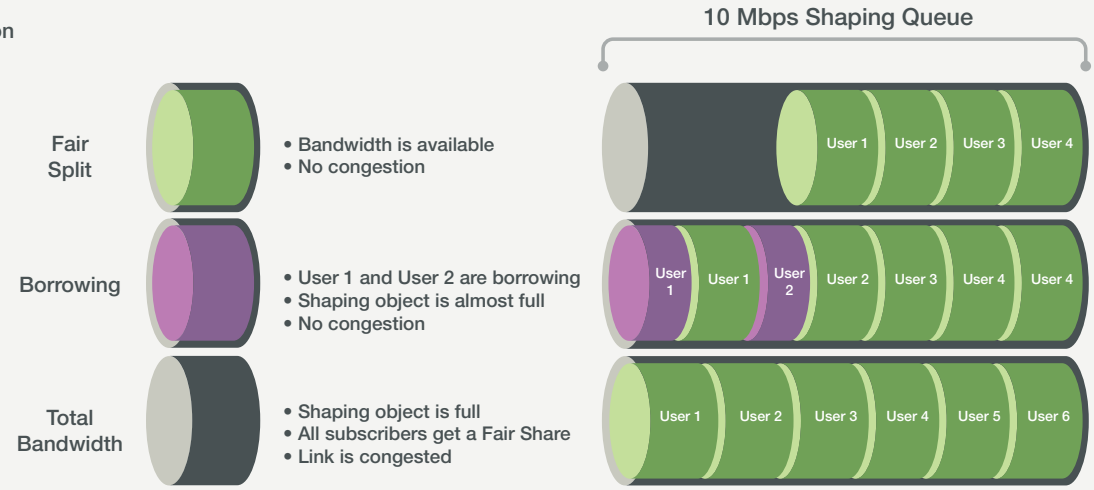


Figure 2

Fair-Split in action with ActiveLogic



ActiveLogic does not set a static limit on each subscriber when allocating the share, preventing wasting bandwidth when some of the subscribers are inactive. Instead, ActiveLogic dynamically calculates the bandwidth limit for each subscriber where the guaranteed level is set according to how many subscribers are active, allowing active subscribers to "borrow" unused bandwidth, and giving bandwidth priority and adjusting limits for returning inactive subscribers.

Subscriber Fair-Split Weighted by Fair Factor Group/Tier (Also called Fair-Split with Fair Factor)

Fair Factor builds on Fair-Split, adding the ability to tier the bandwidth available to subscriber groups (e.g., different service plans), meaning service plan or group awareness is added to the fair usage. With this approach, operators can configure the proportion of bandwidth available for each service plan versus allowing all subscribers equal bandwidth access.

In a network with four service tiers, the highest service tier might be given access to 4/10 of the bandwidth, the next 3/10, then 2/10 and the final plan 1/10 – allowing all subscribers to have access to some bandwidth, but high ARPU subscribers (who are usually paying for more bandwidth) have greater share of the overall bandwidth. As with Fair-Split, any bandwidth not used by a service tier is made available to the other tiers to minimize bandwidth waste. Within the share allocated to a service plan, the resources are equally shared between the users in that plan following Fair-Split mechanism. **Figure 4** demonstrates how Fair Factor assigns bandwidth resources.

In this example **(Figure 3 on following page)**, there are 10 active subscribers accessing a total of 100 Mbps bandwidth. The subscribers belong to different tiers and hence given an appropriate fair factor in order to assign resources according to the tier they belong to. There are two users in Gold, three in Silver, and five in Bronze.

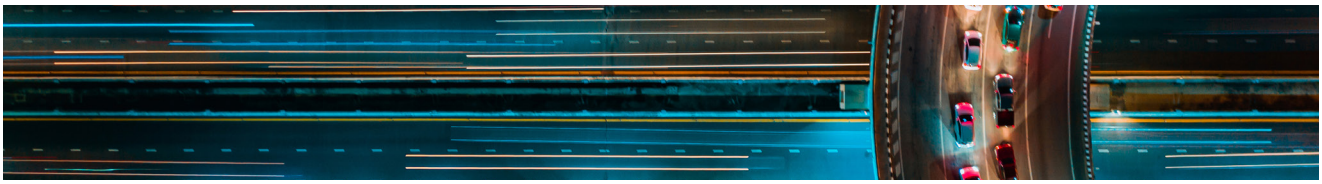
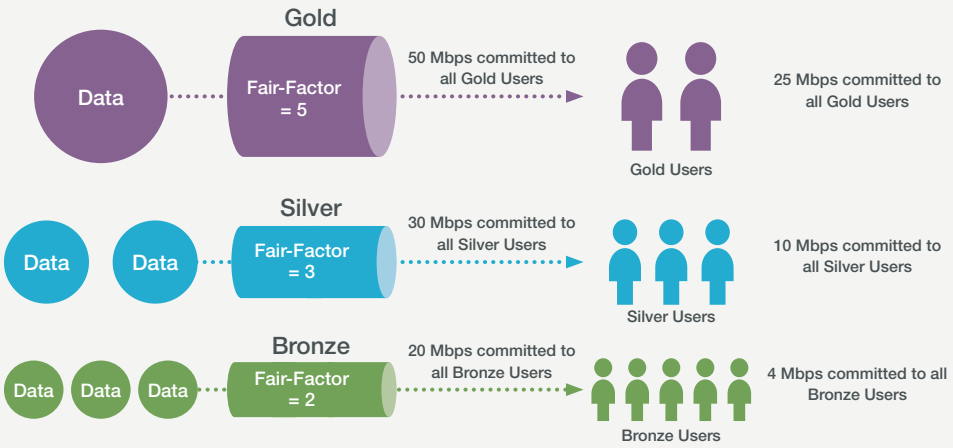


Figure 3

Fair Factor with Fair-Split in action with ActiveLogic

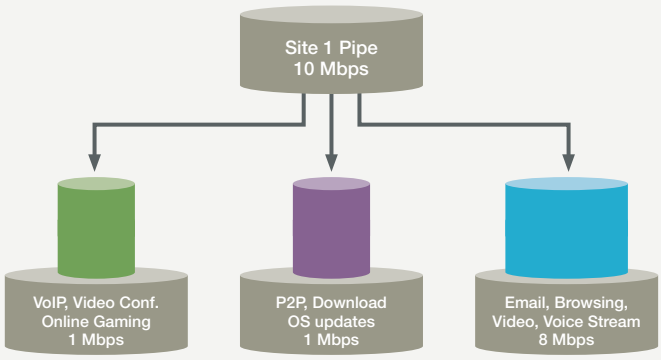


Subscriber Fair-Split with Application Awareness

Not all applications require the same network resources to achieve a high QoE. For example, real-time applications like VoIP and gaming need low latency connections, and streaming video requires high bandwidth to prevent high resolution streams from buffering or stalling. By utilizing a combination of Fair-Split Shaping (Fair Factor with Fair-Split) and Borrowing, ActiveLogic can accomplish a configuration where the bandwidth within each Application or Traffic Class is fairly divided between the subscribers, and also the amount of bandwidth per Traffic Class can be specified to ensure available bandwidth for Real-Time and Interactive Traffic Classes during congestion. This will make sure that a targeted heavy-application cannot use beyond a certain percentage of the bandwidth. **Figure 4** demonstrates this concept.

Figure 4

Application Classes and Borrowing



Taking Fair Factor into account, each Traffic Class will Fair-Split the bandwidth within a Traffic Class based on the configured Fair Factor, as depicted in **Figure 5**.

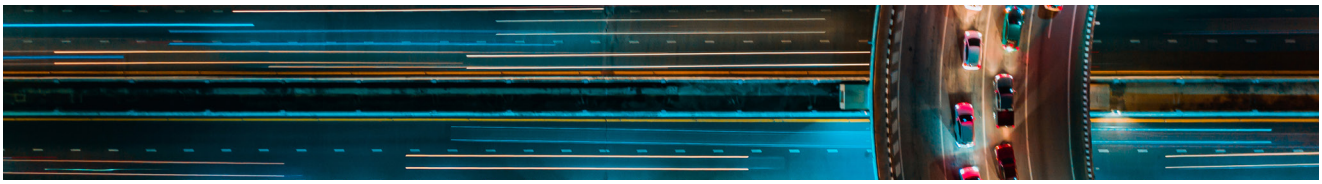
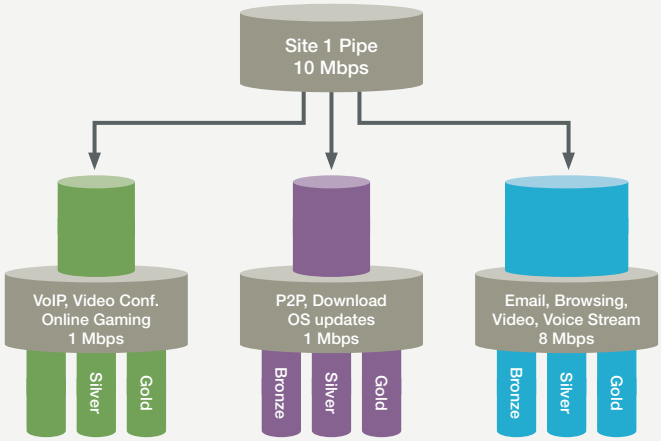


Figure 5

Traffic Classes with Fair Factor

With each traffic class the bandwidth will be Fair Split between the subscribers based on the Fair Split Factor. In this case, Gold gets twice as much as Silver

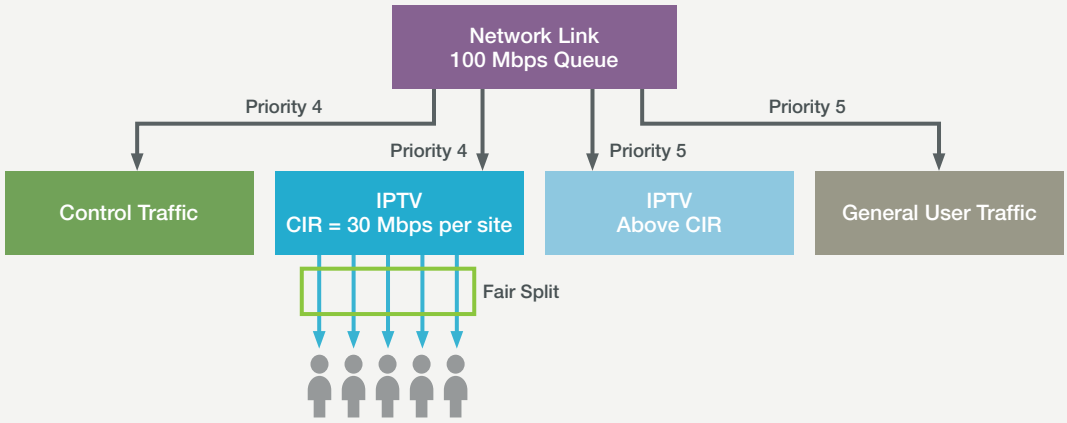


Strict Priority Shaping (Application-aware)

When managing traffic, priority can be assigned to each application type and within each application priority, Fair-Split can be applied to allocate assigned resources to active users. In **Figure 6**, the network link capacity is divided into three application types: control traffic, IPTV, and others. Control traffic consists of low-bandwidth control traffic from the network element and has higher priority (priority 4) over general traffic (priority 5). IPTV traffic is delivered with priority 4 up to its committed rate of 6 Mbps for each site. Excess IPTV traffic that bursts above the committed rate will be delivered with the same priority as general traffic (priority 5). Fair-Split is applied to each of the three application queues, and traffic from each queue is distributed equally amongst all active users.

Figure 6

Link-Level Traffic Management with Application



Multi-layered Shaping and Parallel Queuing

ActiveLogic supports multi-layered shaping schemes that map to various network entities to effectively allocate resource where it is most needed. When multiple layers of shaping are involved, ActiveLogic minimizes packet delay by simultaneously processing user traffic in all applicable queues. In contrast, with conventional multi-layer shaping implementations, where traffic must be processed serially with long additive delay, ActiveLogic parallel queueing minimizes traffic latency through the data plane element by pipelining the shaping processes **(Figure 7)**.

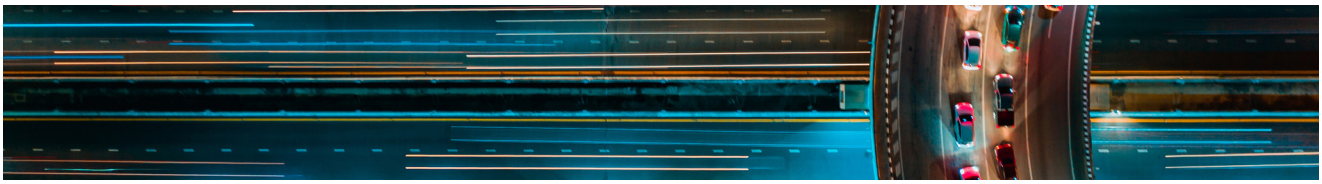
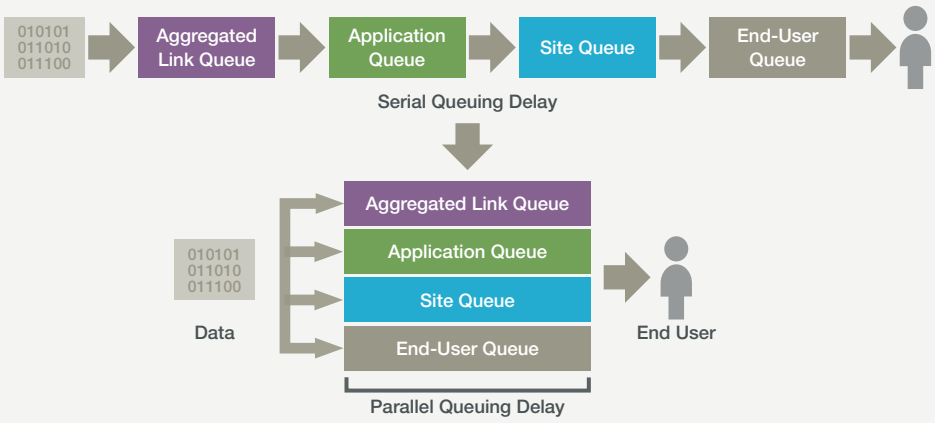


Figure 7

Minimal delay with Serial versus Parallel Queuing



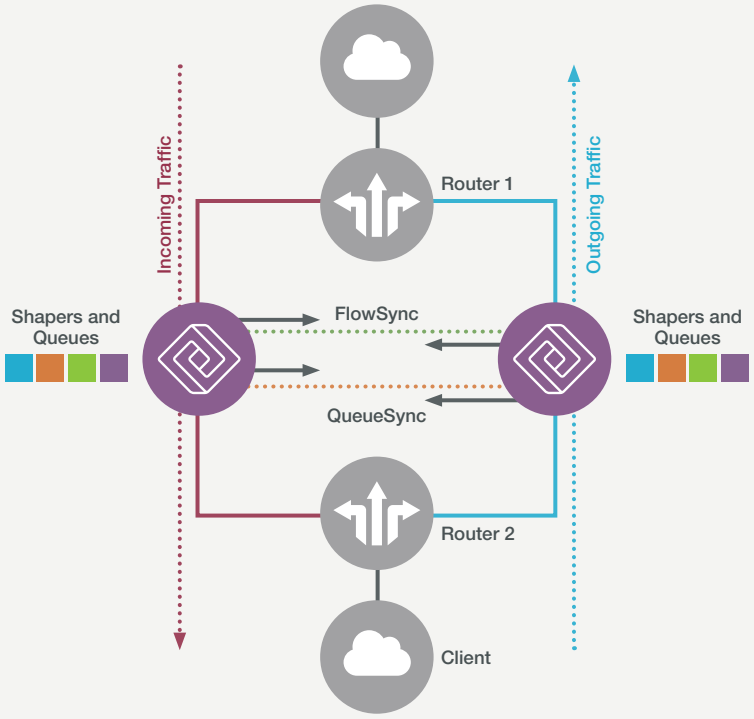
QueueSync – Shaping with Asymmetry or Shaping Scalability

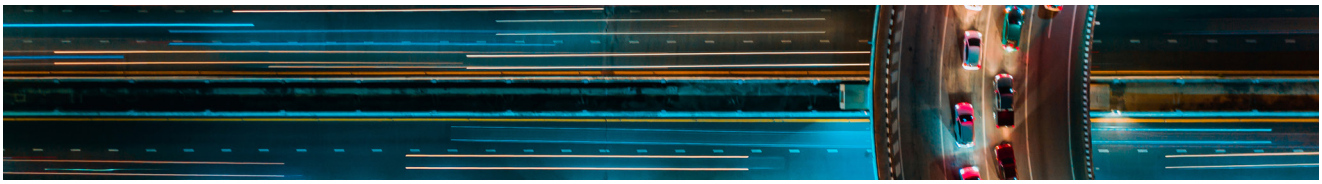
Asymmetric traffic flow is common within most large networks and it creates an undesirable situation for those network devices reliant on monitoring every packet in the customer connection path to properly perform its function.

ActiveLogic Queue Synchronization (QueueSync) technology is a method used when clustering multiple devices in order to provide horizontal user plane traffic scale and redundancy in environments with any of Sandvine's Network Optimization use cases implemented (**Figure 8**). In these environments, subscriber traffic is typically distributed across all active cluster members with a mix of symmetric and asymmetric traffic flows. In order to manage traffic, shaping queue (ShapingObject) information is synchronized across the cluster using a dedicated layer2 network.

Figure 8

Two ActiveLogic nodes with FlowSync and QueueSync seeing asymmetric traffic





SHAPING CONFIGURATION EXAMPLES

A few examples to show shaping in action and how it impacts traffic.

1. Limits and Guarantees with Borrowing (plus virtual queueing)

In this example, the subscriber has a plan with a maximum throughput of 10 Mbps. A Borrowing shaping configuration is implemented to share this per-subscriber bandwidth pool of 10 Mbps among four distinct application categories (VoIP, video, web browsing, and others), while ensuring these application categories have a minimum guaranteed bandwidth.

Also, VoIP services receive special treatment in this configuration scenario, where it was limited to a maximum bandwidth of 500 Kbps, but virtual queueing is employed due to latency requirements of VoIP traffic. This approach ensures that no additional latency is inserted in the VoIP streams by the shaping process, avoiding perceived call quality degradation. Please note: the rule set is built to ensure VoIP traffic is still counted as part of the per-subscriber 10 Mbps pool, while avoiding any added latency due to borrowing from a ShapingObject that does not use virtual queueing.

In this configuration, it is important to avoid oversubscription of the ShapingObject from which bandwidth is being borrowed from. Therefore, the sum of all guaranteed bandwidths should not exceed the maximum bandwidth of the ShapingObject. The guaranteed bandwidth allocated to each application category as described below:

Bandwidth limits and guarantees

Service	Guaranteed bandwidth (CIR)	Maximum bandwidth (PIR)
VoIP	500 Kbps	500 Kbps
Video	5.0 Mbps	10 Mbps
Web Browsing	3.0 Mbps	10 Mbps
Others	1.5 Mbps	10 Mbps

The desired traffic is generated from a single subscriber who started accessing different services as time progresses. In this example, the bandwidth for all TCP traffic (browsing, video, and others) is intentionally set to a high value to force a congestion situation in the ShapingObject. The VoIP traffic (UDP) is set to the CIR (minimum guaranteed rate) to clearly show the effects of virtual queueing on latency. Also, only the inbound direction of the VoIP call is shown. In this configuration, BLUE AQM with default configurations is used.

LiveView is used to monitor the traffic and the bandwidth consumption. **Figure 9** shows the LiveView output when different traffic categories are introduced. In the beginning, the “Others” traffic is able to use all of the available 10 Mbps of bandwidth, since at that time no other services have traffic. As other traffic types are introduced into the mix, bandwidth is shared according to the guaranteed bandwidth allocated to them. VoIP has a guaranteed bandwidth of 500 Kbps and it is able to properly obtain the necessary bandwidth from the pool. In addition, since the VoIP ShapingObject uses virtual queueing, shaping does not add any latency in the VoIP traffic.

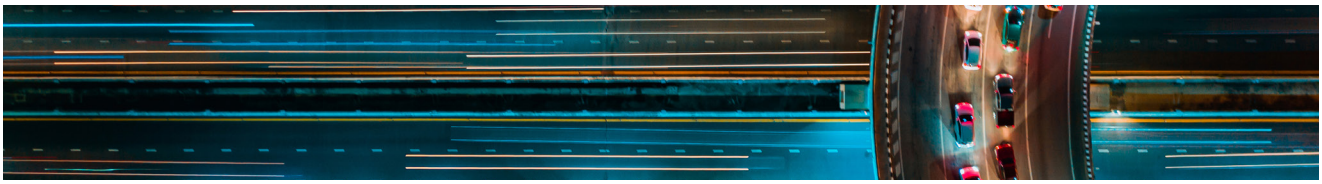
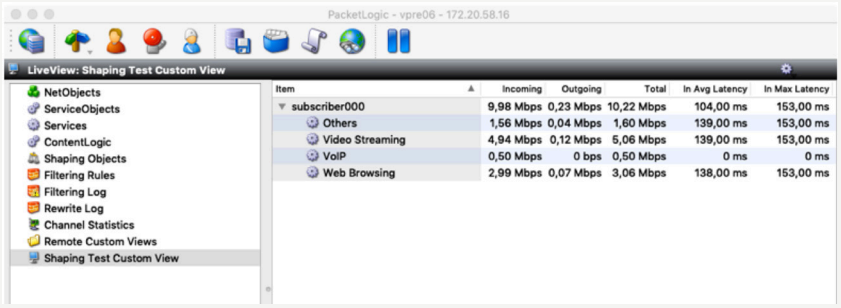


Figure 9

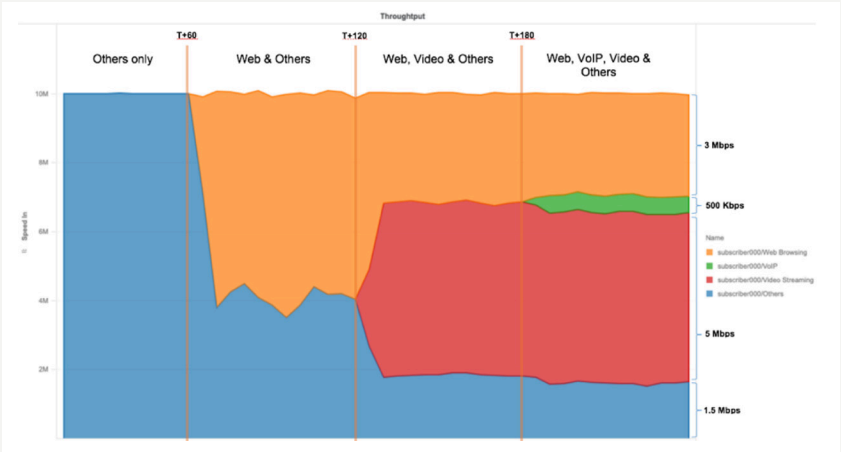
VoIP, Video, Web, and Others traffic types



The graph below shows bandwidth allocation and sharing progress over time.

Figure 10

Traffic behavior as traffic for different services require bandwidth



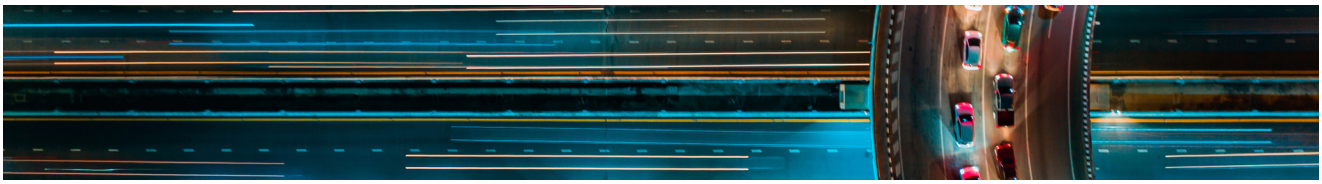
2. Weighted Fair Queueing

In this example, the subscriber has a plan that allows for a maximum throughput of 10 Mbps. A WFQ shaping configuration is implemented to share this per-subscriber bandwidth pool of 10 Mbps among four distinct application categories (VoIP, video, web browsing, and others). The configuration ensures that these application categories have access to a minimum percentage of the full bandwidth in the ShapingObject. Shaping rules are used to assign a priority to each traffic type, and the percentage of bandwidth is set in the ShapingObject. Table 2 lists the priorities and WFQ percentages for this example.

WFQ priorities and percentages for different traffic types

Service	Priority	Percentage
VoIP	4	5%
Video	5	50%
Web Browsing	6	30%
Others	7	15%

Without WFQ, traffic with better priority (lower priority number) is dequeued first, and other priorities start dequeuing only when all better priority queues are empty. Depending on traffic profile, this might starve lower priorities completely. WFQ allows priorities 4 to 9 to have access to a configured percentage of the ShapingObject's bandwidth.

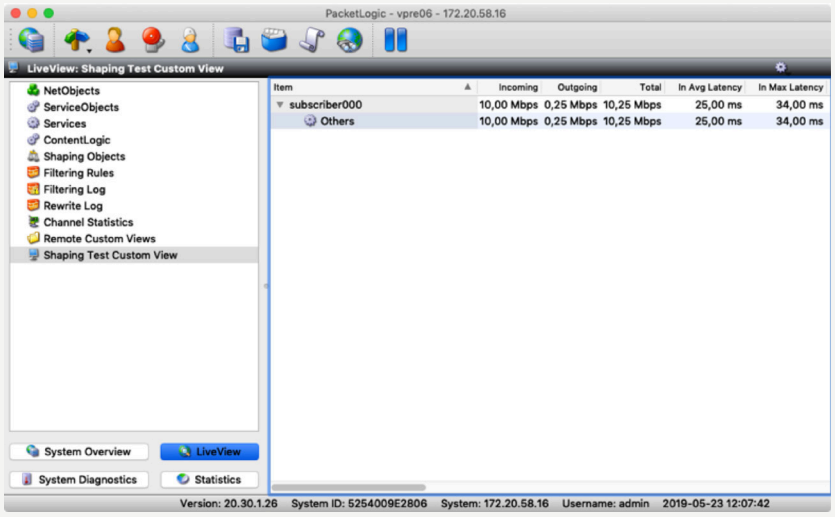


The desired traffic is generated from a single subscriber who started accessing different services as time progresses.

LiveView is used to monitor the traffic and the bandwidth they are consuming. The screenshots below show the LiveView output when different traffic categories are introduced. Notice that it is able to use all of the available 10 Mbps of bandwidth, since no other services (video, web, or VoIP) with better priorities have traffic now.

Figure 11

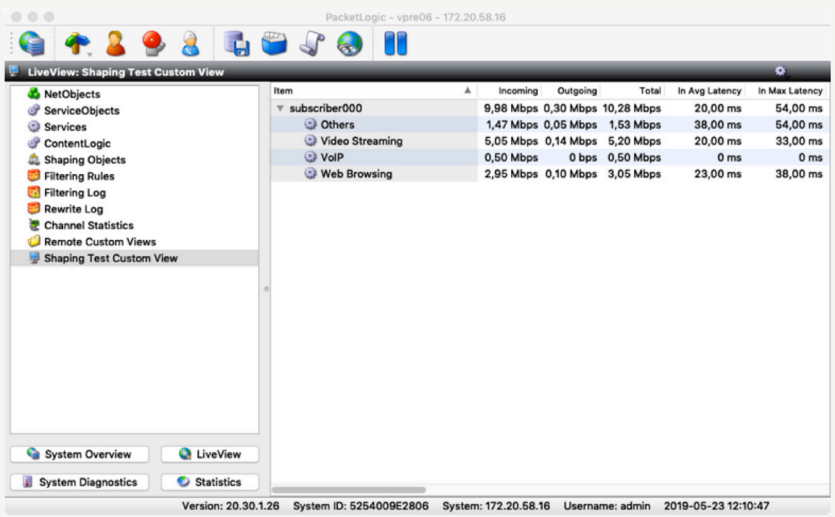
"Others" traffic using the entire 10 Mbps bandwidth, since no other services have traffic

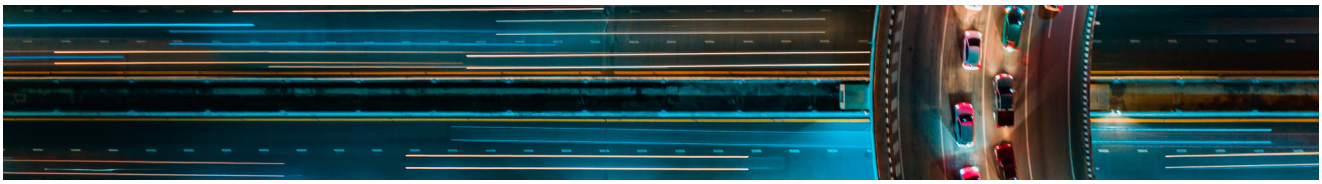


Similarly, when web browsing traffic is introduced, "others" traffic is immediately preempted to its allocated WFQ percentage (15%), as web browsing traffic has better priority and is trying to use as much bandwidth as possible. Without WFQ, the "others" traffic would starve in this case. This is observed with video and VoIP traffic also. When all traffic categories are present, VoIP preempts all other services to their configured bandwidth percentages.

Figure 12

VoIP, Video, Web browsing and Others traffic





The graph in **Figure 13** shows bandwidth allocation progress over time, as traffic for different categories is started at different times.

Figure 13

Traffic behavior as traffic for different services require bandwidth

ABOUT SANDVINE

Sandvine's cloud-based Application and Network Intelligence portfolio helps customers deliver high quality, optimized experiences to consumers and enterprises. Customers use our solutions to analyze, optimize, and monetize application experiences using contextual machine learning-based insights and real-time actions. Market-leading classification of more than 95% of traffic across mobile and fixed networks by user, application, device, and location creates uniquely rich, real-time data that significantly enhances interactions between users and applications and drives revenues. For more information visit **http://www.sandvine.com** or follow Sandvine on Twitter **@Sandvine**.

SANDVINE

USA
5800 Granite Parkway
Suite 170
Plano, TX 75024
USA

EUROPE
Svärdfiskgatan 4
432 40 Varberg,
Halland
Sweden
T. +46 340.48 38 00

CANADA
410 Albert Street,
Suite 201, Waterloo,
Ontario N2L 3V3,
Canada
T. +1 519.880.2600

ASIA
RMZ Ecoworld,
Building-1, Ground Floor,
East Wing Devarabeesanahalli,
Bellandur, Outer Ring Road,
Bangalore 560103, India
T. +91 80677.43333