Quality of Service (QoS)

WHITE PAPER

Why it Matters,
Why Your Network Can't Always Deliver it,
and How to Ensure It Does

BADU networks

- Improving the way the world connects -

CIO Review 20 MOST PROMISING NETWORKING SOLUTION PROVIDERS - 2017

Executive Summary

Network QoS (Quality of Service) has been defined as the ability of a network to deliver the level of performance required for business-critical applications to meet SLAs. The lifeblood of virtually every business is its network, and the impact of poor performance can be devastating. Amazon calculated that a page load slowdown of just one second cost it $1.6 billion in sales each year. In addition, Google found that slowing search response times by just four-tenths of a second reduces the number of searches by eight million per day, leading to a corresponding reduction in ad revenue. In an A/B test conducted by Shopzilla that compared the impact of page load times on conversion, they discovered that faster pages delivered seven to twelve percent more conversions than slower ones. Analyst firm Aberdeen Group found that across a broad range of firms, a one second increase in response time reduces conversion rates by seven percent, page views by eleven percent, and customer satisfaction rates by seventeen percent.

The impact of a slow network on employee productivity can also be significant. A recent survey conducted by flash memory vendor SanDisk revealed that the average employee wastes one week per year waiting for their company's network to respond. It's relatively easy for any business to use this figure to calculate losses based on the number of employees and their average salary. However, the downstream impact in terms of lost sales when project deadlines are missed and new product shipments are delayed is often harder to determine, and may be exponentially greater.

When faced with throughput and performance challenges, companies generally turn to costly network upgrades to increase bandwidth. The upgrades are typically coupled with QoS solutions designed to ensure business-critical applications meet SLAs by controlling four key metrics: available bandwidth, packet loss, latency and jitter. Network QoS solutions do an effective job of prioritizing traffic, preventing packet loss, and reducing latency. However, they are struggling to overcome reduced throughput and the resulting slow performance caused by

jitter – the real-time changes in network traffic flow – that have become increasingly common due to the nature of today's networks and applications, and how they are deployed and used. Network upgrades provide no solution, as the incidence of jitter-induced throughput collapse often increases when more bandwidth is available.

The root of the problem lies in the fact that TCP, the most widely used network protocol, misinterprets jitter as congestion caused by either too many packets flooding the available bandwidth, or a hardware failure somewhere on the path. In response, TCP reduces throughput, slowing down network traffic in an unfortunate effort to prevent data loss even when plenty of bandwidth is available and all network equipment is fully operational.

Increasing Sources of Jitter

Today's streaming services, IoT devices, voice, video and web applications typically transmit data in unpredictable bursts. This means jitter often originates on the servers hosting these applications, before their traffic even enters the network. When applications run in a virtualized environment like AWS, scheduling conflicts between VMs, and packet transfer delays caused by hypervisors managing them add further sources of jitter.

The popularity of cloud services, which Gartner estimates will grow from 17% of all IT spending in 2017, to 28% by 2021, guarantees that virtualization's contribution to jitter will grow correspondingly. In addition, many organizations have adopted cloud-first strategies for deploying new applications. This means virtualization jitter and its attendant throughput collapse will become a major factor in the success or failure of many new application rollouts.

Jitter from physical and virtual server environments is compounded by the volatile nature of Wi-Fi and mobile networks that frequently suffer from RF interference, fading and channel access conflict. An estimated 70% of internet traffic is now consumed over a wireless connection. This high percentage is expected to grow substantially over the next few years with the burgeoning use of cloud services, and the rollout of billions of IoT-enabled devices that already make heavy use of Wi-Fi and LTE.

According to analyst firm Gartner, the number of IoT devices will grow from 6.6 billion in 2016 to 20.8 billion by 2020. Given that many IoT devices utilize Wi-Fi or LTE services, primarily to

upload data to the cloud, the combined impact of IoT, wireless and virtualization jitter on tomorrow's networks will be significant.

In addition, growing upload volume of IoT sensor data, video, images and social media content, is putting pressure on wireless network performance. The reason is jitter increasingly results from upload streams interfering with simultaneous download streams through the same Wi-Fi access point, or LTE eNB. This interference does not occur at the hardware layer, but rather in the transport layer (layer 4) of the OSI network stack.

In an increasingly real-time, virtualized and wireless world, guaranteeing QoS is impossible without a solution that effectively addresses TCP's reaction to jitter. To understand what is required to eliminate the negative business impacts of slow network performance, it's important to look at the capabilities of the solutions that support network QoS.

QoS Solution Overview

In a traditional IP network, each router makes an independent forwarding decision for each packet based solely on the packet's network-layer header, source and destination IP address. The process is repeated by each router at each hop along the way until the packet eventually reaches its destination. These routing table lookups and independent forwarding decisions can add overhead and unpredictability that make it impossible to meet SLAs for mission critical applications. In dedicated WAN links between branch offices and on-premises corporate data centers that traditionally host enterprise applications, QoS has been addressed with Multiprotocol Label Switching (MPLS), introduced in the late 1990s.

MPLS

With MPLS, bandwidth is reserved and network traffic is prioritized by forcing it over a predetermined path to its destination, rather than leaving it to each router along the way to decide the next hop, as is the case with traditional IP networks. The first router in the path does the routing table lookup. The first time a packet enters the network it is assigned to a specific forwarding equivalence class (FEC) reflecting the level of service required. A label is prepended that identifies the assigned FEC which gives the MPLS network the ability to handle packets with similar characteristics, such as coming from specific ports, by directing them over predetermined paths. For example, packets

carrying mission-critical application traffic can be mapped to high bandwidth, low-latency routes across the network to insure SLAs are met.

Each router in the network has a table indicating how to handle packets of a specific FEC type, so once the packet has entered the network, routers don't need to perform header analysis. The label enables packets to be forwarded at OSI Layer 2 (the switching level) rather than having to be passed up to Layer 3 (the routing level), for a routing table lookup first. The packet is simply switched from its incoming interface to the outgoing interface designated by the MPLS switching table. This eliminates the overhead of routing table lookups and independent forwarding decisions at each network hop.

The skyrocketing popularity of SaaS applications and other cloud services has made dedicated MPLS networks that route internet traffic through an on-premises data center too costly and painful for many organizations. This cloud-bound traffic is increasingly eating into available bandwidth and degrading performance for their in-house applications. In addition, once packets leave the on-premises data center, MPLS labels have no impact. This coexistence of on-premises and cloud applications has given rise to the concept of the hybrid WAN. A typical hybrid WAN features an MPLS pipe that connects a branch location to an on-premises data center for in-house system access, and a broadband connection that enables direct internet access to the cloud. In this hybrid environment, SD-WAN solutions have entered the QoS arena. SD-WAN eliminates the backhaul penalties of traditional MPLS networks that first route cloud bound internet traffic through an on-premises data center, and leverages the Internet directly to provide high-performance connections to the cloud.

SD-WAN – QoS for the Hybrid WAN

SD-WAN is the application of Software Defined Networking (SDN) to the WAN. SD-WAN makes it possible to bond multiple WAN connections -- Internet, private dedicated MPLS, LTE or any other transport pipe -- effectively making the best use of bandwidth and reducing dependency on expensive MPLS links. This is accomplished by placing SD-WAN edge devices at on-premises sites and managing them centrally. Overlay tunnels are created on top of the available transport links. SD-WAN is transport agnostic, so it doesn't care whether the transport is MPLS, broadband or a 4G mobile connection.

The key features that make SD-WAN a compelling QoS solution for today's environment include:

- **Application Based Prioritization -** Services can be prioritized at a business application level rather than a network configuration level, as is the case with MPLS. You define the applications to prioritize for your organization, setup classification rules and assign those rules their appropriate priority level and bandwidth allocation. There's no need to do network level configuration, or build complicated QoS policies for each application at every location

- **Dynamic Path Selection** - SD-WAN solutions can move traffic around from path-to-path on the fly based on real-time performance measurements of the available links. For example, latency sensitive application traffic can be moved to the lowest latency line while high bandwidth, more latency tolerant traffic can leverage the higher bandwidth lines. This means that traffic traversing an SD-WAN network can be aggregated across all available links, making it possible to utilize the sum-total of their combined bandwidth. This is massively simpler than MPLS, which requires predetermined routes to be provisioned, and once fixed circuits are in place, making changes is not a point-and-click exercise

- **Built-in resilience** - Traditionally, redundancy and failover have been performed with dynamic routing protocols such as BGP or OSPF. These protocols will switch to a backup path if the primary goes down. With SD-WAN, the lines are all utilized in an active/active state. This means no failover process is required if one line fails, because the traffic continues flowing over the remaining active lines.

Why SD-WAN Can't Guarantee QoS

SD-WAN vendors claim their solutions not only measure packet loss, latency and jitter, but also compensate for them to insure QoS. Some SD-WAN vendors even claim they can replace MPLS. Here's the reality:

- SD-WAN is an edge technology. An SD-WAN can make decisions based on measurements at the edge, but it has no control over what path a packet takes once it leaves the premises and enters the cloud. To an SD-WAN, the network is a black box. If guaranteed QoS is needed, there will be a need for reliable transport. It can be MPLS or any other transport, but not SD-WAN with pure Internet links. Even most SD-WAN vendors recommend keeping an MPLS link in parallel to the broadband link to ensure QoS for real-time traffic like voice and video

- SD-WAN solutions compensate for packet loss with forward error correction. Once packet loss is identified on a path, duplicates of the same packet are sent to have greater assurance that critical data arrives at the destination. At the other side of the SD-WAN connection, the first packet received will be sent along and the duplicates will be dropped

- Compensating for latency due to distance isn't possible using any technology, since data can't be transferred faster than the speed of light under the best possible conditions. What vendors really mean is they can move traffic to a less congested path on the fly, and use WAN optimization techniques to speed traffic by reducing bandwidth usage, or multiplex a single connection over multiple paths

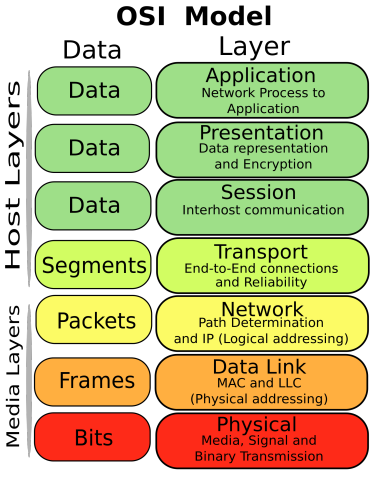
- To compensate for jitter, SD-WAN solutions measure gaps between packets and try to evenly space these packets on the other side, by providing a "jitter buffer" to realign packet timing for consistency. This may work for some applications, but it can also create delays and destroy performance for real-time applications like voice, video, fast data, and IoT applications that stream data in irregular, unpredictable bursts. Packet scheduling is simply another name for delay.

WAN Optimization

WAN optimization vendors, many of whom now also have SD-WAN offerings, focus on accelerating traffic by reducing its volume. They achieve this by employing techniques such as data de-duplication that only sends the latest changes to a data set, rather than resending the entire data set, data compression, and caching to move frequently accessed data closer to its users. These techniques yield some benefits in accelerating traffic by reducing its volume, but they can't do much for data that's already compressed and/or encrypted, and they require access to the payload. In addition, now that up to 80% of all internet traffic is encrypted according to Google, the need for payload access introduces the added overhead of encryption/decryption at each endpoint, as well as the risk of exposing sensitive security key information to third party vendors and their tools.

WAN optimization vendors attempt to address jitter primarily by managing the size of TCP's congestion window (CWND) to let more traffic through a connection, but this ultimately doesn't stop jitter-induced throughput collapse, or the ensuing slow recovery process.

From the perspective of the OSI model, QoS solutions focus primarily on layer 7 - the application layer where SD-WAN software and WAN optimization solutions operate, and layers 2 and 3, the network and data link layers respectively where MPLS labelled packet interact with routers and switches.



Network administrators often try to address QoS problems resulting from jitter at layer 1 (the physical layer) by upgrading bandwidth, only to see it deteriorate further in many cases, as the incidence of jitter-induced throughput collapse often increases with more bandwidth. However, these solutions can't fully guarantee QoS either alone, or in combination due to their inability to deal effectively with jitter - a layer 4 (transport) issue that most vendors only partially address.

TCP's response to jitter becomes the bottleneck over the entire path from the sending server transmitting data over the wired portion of the path, to the wireless last mile, preventing full use of available bandwidth and delivering a poor user experience. Without first tackling jitter head-on, it's not possible to guarantee QoS in today's application environment.

Overcoming Jitter-Induced Throughput Collapse – The First Step in Insuring QoS

In overcoming throughput collapse all IT vendors and service providers focus on tuning the physical layer. They add capacity, introduce packet grooming, which is really packet scheduling by delay injection, or they implement TCP session capacity limits that allow for planned packet loss. Unfortunately, all these approaches have the effect of reducing TCP session throughput. In addition, they fail to address the fact that the server where TCP session traffic originates may also be the problem.

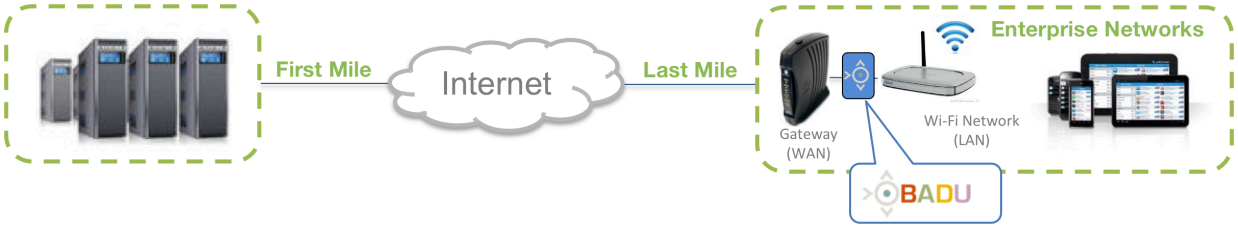
TCP's original design assumption of orderly packet delivery in relatively consistent time intervals made sense when it was introduced over 40 years ago, and still has validity when networks truly become saturated. However, today's streaming applications often generate traffic characterized by short, unpredictable bursts of data that cause significant variation in Round-trip Time (RTT).

TCP's two congestion management schemes; retransmission and congestion window (CWND) management are designed to handle reliability while avoiding packet loss. Between the two, retransmission is by far the more dominant. With each retransmission attempt the Retransmission Timeout (RTO) value is increased, and CWND is reduced on the assumption that the network has become congested. After three RTO's throughput is halved. After seven RTOs throughput collapses because TCP treats the packets as lost rather than merely delayed, and prevents traffic from being sent to guard against further data loss.

This behavior is desirable in pristine networks running traditional in-house business applications. However, today's networks are not pristine, they are messy, unpredictable, and exhibit randomized stochastic behavior. Badu Networks' patented *Warp*TCP™ technology offers the only solution focused squarely on the TCP bottleneck issue for both wired and wireless networks. *Warp*TCP analyzes traffic to determine if congestion is real, and prevents TCP from unnecessarily reducing throughput in response to jitter. *Warp*TCP 's proprietary algorithms estimate actual bandwidth available to each TCP session in real-time, filtering out transient

fluctuations in RTT and packet loss. *Warp*TCP is specifically designed to deal with rapidly changing bandwidth, loss patterns, server loads, and RTT variance, enabling it to do well in volatile environments like mobile and Wi-Fi networks – i.e. today's global network environments. As a result, throughput and performance stay at consistently high levels, even in the face of extreme fluctuations. *Warp*TCP improves both download and upload throughput by as much as 10x in wireless environments, even when the user ventures away from the Wi-Fi AP or mobile eNB, and the connection is subject to greater RF interference and channel access conflict.

*Warp*TCP Architecture



*Warp*TCP consists of two components that work hand in hand to prevent TCP throughput collapse and optimize the use of all available bandwidth to maximize performance:

- A TCP de-bottleneck module that implements *Warp*TCP's proprietary algorithms that determine in real time if jitter is due to congestion, and prevent TCP from reducing the size of CWND when it's not

- A Transparent TCP Proxy that implements TCP session splicing by splitting the connection between the server and the client into two independent sessions. Each spliced server-to-client TCP session is replaced by a server-to-proxy sub-session and a proxy-to-client sub-session. The two sub-sessions are independent TCP sessions with independent control. In most deployments, *Warp*TCP retains the IP addresses and port numbers associated with the original TCP source and destination to map them to the new sub-sessions.

*Warp*TCP's session splicing approach is another key differentiator because it completely breaks the dependency between sending server and receiving client. This independence enables *Warp*TCP to implement its own flow control algorithms based on speed and capacity matching that are far superior to TCP's. With speed matching, the proxy receives as many packets as possible, as fast as possible, buffers them without modification. It then forwards the packets to the client at different speeds and times. Speed-matching enables another performance enhancing feature - opportunistic bursting. Opportunistic bursting allows *Warp*TCP to fill-in unused gaps in bandwidth with packets that would otherwise be stalled. With Capacity

matching, backhaul transports can be configured to provide jumbo MTU frames that *Warp*TCP will utilize in an opportunistic fashion.

In contrast, most WAN/TCP optimization vendors adopt a snoop approach, implementing their proxy as a packet filter, without terminating TCP sessions in the proxy. In a packet filter, the sever-to-proxy and proxy-to-client sub-sessions share the same sequence numbers and the same acknowledgements (ACKs). As a result, they are limited by TCP's inferior flow control. In addition, some vendors inject spurious ACKs, or do other protocol hacks that are not compliant with RFCs. This exposes their customers to major support risks when applying TCP patches, or performing upgrades.

Another key advantage provided by *Warp*TCP's architecture is the enablement of dramatic improvement in the most visible aspect of the user experience that has a direct impact on the bottom line – page load times. Browsers only support establishment of two to four TCP sessions simultaneously, whereas a web page can easily have over 100 objects, each requiring its own TCP session to send and receive data. With *Warp*TCP's session splicing and speed-matching, many more sessions can be handled in parallel and RTT can be reduced significantly. Since TCP connections with the browser are independent of the server connections and can be sent in parallel, pages typically load 2-3x faster.

Flexible Single Instance Deployment

*Warp*TCP requires no changes to clients or servers. It can be deployed at any single point on the network close to the source of jitter. In contrast, most packet filter TCP optimization solutions require their proxies to be placed near the server, or two proxies to deployed - one at each end of the path between the server and the client.

*Warp*TCP can be deployed as a software module or hardware appliance next to an application server in an on-premises data center, between a WAN gateway and Wi-Fi access point, or at a cell tower base station. *Warp*TCP can also be deployed as a VM instance in a cloud environment. In addition, *Warp*TCP requires no access to the payload. It's completely agnostic to the type of content, or whether it's encrypted or unencrypted

Conclusion

Network QoS can only be insured if bandwidth usage, packet loss, latency and jitter can be controlled to consistently meet application SLAs. The MPLS, SD-WAN and WAN optimization solutions that traditionally support QoS operate primarily in the application, network and data link layers of the OSI network stack. They do an effective job of prioritizing traffic, preventing packet loss, and reducing latency, other than that caused by distance, so they have an

important role to play. However, they are struggling to guarantee QoS due to the increasingly jitter-prone nature of today's application traffic, and the fact that the bulk of it travels at least partially over volatile wireless links. The rapid adoption of cloud services, and the proliferation of IoT-enabled devices will only further intensify the impact of jitter and increase the incidence of jitter-induced throughput collapse in the years ahead.

Only Badu Networks' patented *Warp*TCP technology deals directly with the problem of jitter-induced throughput collapse on both wired and wireless networks. *Warp*TCP's unique patented ability to accurately determine in real-time whether congestion exists for each TCP session, and prevent the transport layer's congestion control from reducing throughput when plenty of bandwidth is available, addresses jitter-induced throughput collapse head-on. Implementing *Warp*TCP's approach to managing transport layer congestion control to eliminate jitter-induced throughput collapse, in conjunction with traditional QoS solutions that tackle the non-jitter aspects of throughput and performance, provides QoS insurance for an increasingly jittery future.