Bringing the power of Analytics to improve end-user Quality of Experience

Sangeeta Ramakrishnan, Xiaoqing Zhu, Frank Chan,
Bhanu Krishnamurthy, Cindy Chan, Zheng Lu, Kashyap Kambhatla
Cisco Systems

Abstract

Broadband speeds continue to increase. As operators cope with growing bandwidth demands, expectations amongst subscribers also continue to grow with respect to how well the broadband service performs. Traditionally a broadband system has been judged solely based on how it manages bandwidth. The assumption was that the higher the bandwidth and the lower the packet drops the happier the end consumer would be. While this assumption is fundamentally true, it creates a system that has no clear bounds and therefore it tends to be over-engineered and expensive. Using analytics the "user happiness" can be measured directly and a cost optimized network can be built. The analytics generated can be used for multiple purposes, including troubleshooting, capacity planning, and optimization.

This paper will present the various ways the analytics can be gathered and how it can be used. A primary usecase for this technology is for IP video delivery and troubleshooting any problems associated with its delivery. Operators spend a large amount of OpEx dollars for troubleshooting their networks and services. Analytics are key to being able to quickly triage and identify sources of problems.

Another way the analytics can be used is for capacity planning. Today much of capacity planning is highly dependent on network interface utilization. We will instead present a novel method to go beyond utilization levels and in fact estimate bandwidth demand on the network. Improved capacity planning techniques can help operators spend their Capex dollars in an efficient manner to target network spend on the areas that need it the most, thereby being cost-effective while yet improving subscribers' experience.

Finally we will present ways in which the analytics can be leveraged to optimize the efficiency of the operator's network. The optimizations can range from video-aware cable modem load balancing, to WiFi optimizations that enhance broadband delivery. Further these analytics can be used to optimize IP video Quality of Experience thereby significantly improving the bandwidth efficiency of operators' networks. Results from such optimizations demonstrating up to 40% improvement in stream packing efficiency will also be presented.

Besides outlining the various ways Analytics can be used in operator networks, we will also present a proposal on how this can be realized with a Software Defined Networking (SDN) approach. Given recent technology advances in the industry with respect to Analytics, Big Data, and the advent of Software Defined Networking, this is the right time for the cable industry to adopt Analytics to improve the service offered to subscribers while yet improving the cost effectiveness of offering such services.

INTRODUCTION

According to Cisco Visual Networking Index (VNI) Global IP traffic has increased more than fivefold in the past 5 years, and will increase nearly threefold over the next 5 years. Video as it is dominates the Internet traffic today, but that trend is only going to accelerate further with 80% of Internet consumer traffic expected to be video by 2019 [1]. The vast majority of video being delivered on the Internet uses HTTP Adaptive Streaming (HAS), also known an Adaptive Bit Rate (ABR) video [3].

In the ABR video delivery systems the video is encoded at various bitrates called profiles. Additionally the video is segmented into short fragments each of which is a few seconds long (typically between 2-10 seconds). At every fragment boundary, the ABR client can dynamically select the rate profile to request. The client typically switches to a lower-rate profile, to avoid buffer underflow during network congestion. ABR is widely deployed with Apple HTTP Live Streaming (HLS) [4], Microsoft Smooth Streaming [5], and Adobe HTTP Dynamic Streaming (HDS) being the most popular ABR streaming protocols. More recently Motion Picture Experts Group (MPEG) has issued the Dynamic Adaptive Streaming over HTTP (DASH) [6] specifications standardizing the ABR delivery. However client rate adaptation logic has been left outside the specification to allow innovation by vendors in this area.

ANALYTICS

The advent of Adaptive Bit Rate (ABR) video has changed the traffic characteristics on Service Provider networks. With ABR video, the bandwidth usage on the network adapts up and down to availability and shortage of bandwidth respectively. This makes it harder for operators to estimate end user Quality of Experience (QoE) simply based on BW utilization of interfaces, like it was done in the past. With ABR video, there can be multiple interfaces with comparable network utilization but very different levels of QoE.

Instead of using interface utilization, by examining the level of video oversubscription, a better understanding of end-user QoE can be derived. In our work we have defined video oversubscription as a ratio of the aggregate bandwidth demand for video on an interface to the aggregate bandwidth utilized by video on the interface, with higher numbers of oversubscription implying a worse QoE on such interfaces. In `Figure 1` the blue bars show sample interface utilization for 10 different interfaces. The green bars show video oversubscription for each of those same interfaces. As seen below, while interface utilization looks comparable across several interfaces, the video bandwidth oversubscription is quite different amongst them. The QoE on Interfaces 6 and 8 are significantly higher than others and hence subscribers on those interfaces experience much worse QoE than on other interfaces.

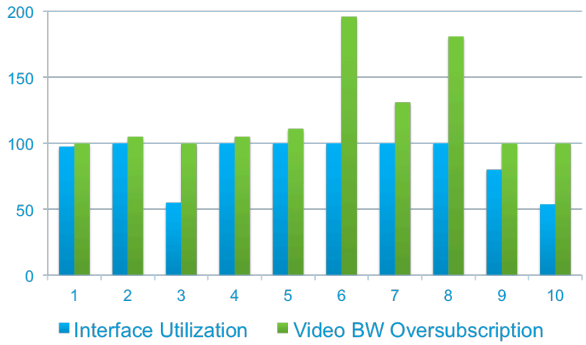

Figure 1 Interface Utilization versus Video Bandwidth Oversubscription

Such an oversubscription metric not only provides an insight into the state of the video delivered on the interface, but also on the overall state of the interface. With vast

majority of traffic on the DOCSIS downstream interfaces being TCP traffic, assuming TCP flows are reasonably sharing bandwidth amongst them, the Video Bandwidth Oversubscription metric not only provides insight into the state of the video flows for which such visibility is available, but also for the remaining TCP flows on those interfaces.

Besides the above-described metric, interfaces to clients can be used to measure buffer stalls, and buffer levels as observed by clients. Aggregating such metrics across clients on a single CMTS or a single CMTS interface can be useful in gauging the health of those interfaces.

ARCHITECTURE

We propose a Software Defined Architecture (SDN) for gathering these analytics as shown in Figure 2. SDN is defined as an architecture that is characterized by the separation of Control and Data planes, and uses an open standard protocol to communicate between them. In an SDN architecture a controller provides network layer abstraction to the applications above it. SDN promises flexibility and rapid innovation by virtue of the fact that Applications can be built almost independent of the underlying dataplane hardware, and need not support network-element-specific protocols, such as COPS, IPDR etc. More information on SDN can be found in [6].

The Video QoE Application collects information from various elements of the end-to-end architecture, from streamers, clients, and various network elements, including the CMTS. The Video QoE application combines information from both video and network domains to provide valuable analytics, that can be used in multiple ways as described in the following sections.

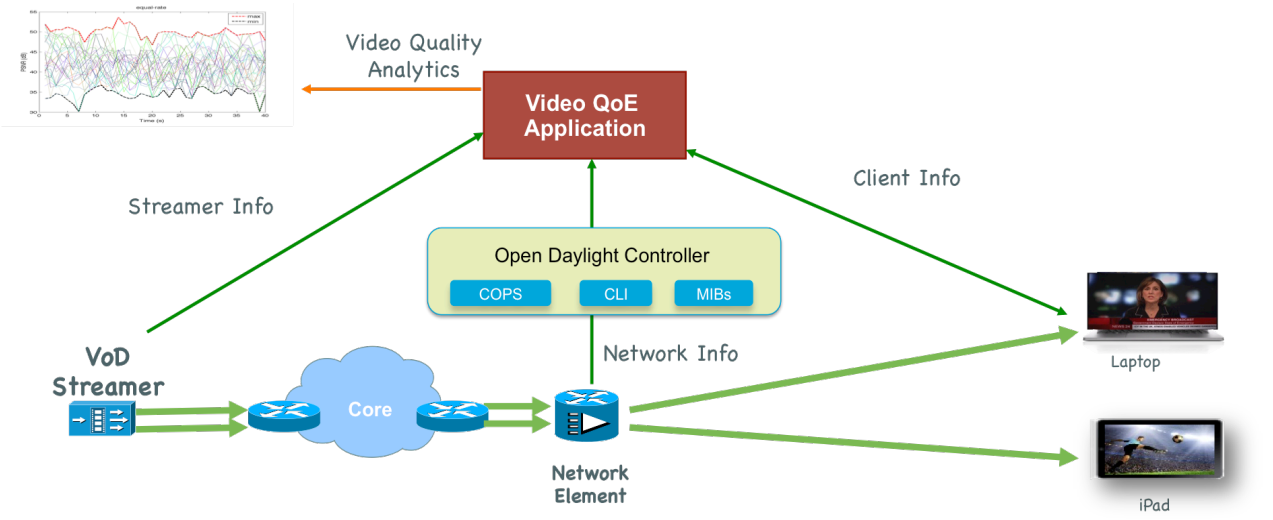

Figure 2 SDN based architecture for Video QoE Analytics

CAPACITY PLANNING

Operators can use the above-described metrics such as oversubscription for capacity planning purposes. It provides better insights into the true state of the network and can thus help them target network upgrades in areas where upgrades are truly needed versus

spreading it out evenly. This enables them to delay capital expenditure and also focus upgrades in a timely manner in the most problematic areas.

With plans afoot to move to virtualization, these kinds of metrics can also be used to spin up additional resources for certain Service Groups over others. Such methods can be used to deploy "virtual capacity" only on a need basis, thereby applying these spare resources selectively to Service Groups where it is needed.

TROUBLESHOOTING

Operators today spend significant amount of OpEx dollars on troubleshooting the network. When customers complain about the video quality they are experiencing, having QoE metrics per user, per network segment etc. can help troubleshoot and root-cause the problem more easily.

For example if only 1% of users are experiencing poor QoE, it may not be clear how big a problem it is and how much resources to expend to troubleshoot the problem. However if the poor QoE users' network path is examined, it may become very obvious that vast majority of poor QoE issues are on a single CMTS or on a single interface etc. This can provide an excellent starting point for operators to dig deeper to troubleshoot and fix the problem.

Although the above-described example limited itself to the CMTS as the only network element, one can see how this can be extended across multiple network elements. In fact in the age of big data, all types of information about flows can be logged, and machine-learning techniques can be applied to identify problem areas. This should help rapidly narrow down problem areas when troubleshooting large scale systems, which may have multiple problem areas, co-existence of which may make it harder to root cause the problem in traditional/manual ways.

NETWORK OPTIMIZATION

Having a better understanding of the QoE of users can help operators improve the QoE of users on their networks. In fact there are two sides to such optimizations – QoE can be improved while keeping network utilization comparable, or QoE can be maintained while reducing network resources required to meet the QoE needs. More realistically as demands on operators' networks grow, they can pack more users and streams on to the same network while maintaining QoE. These types of optimizations can be achieved in a number of ways, and a couple of such usecases are described below.

Video QoE-aware Cable Modem Load Balancing

DOCSIS specifies CMTS to CM interfaces that enable cable modem load balancing across interfaces. It is performed in a couple of ways in today's networks. The first being static cable modem load balancing, where modems are assigned to channels as they come online. In static load balancing typically the counts of cable modems on interfaces are balanced. The second type of load balancing is called dynamic cable modem load balancing and in this approach cable modems are moved across interfaces after they are on-line. Although DOCSIS specifies the messages the CMTS must use to communicate load balancing instructions to cable modems, the standards do not specify how the CMTSs make these load balancing decisions. Typically CMTSs use the interface utilization along with cable modem counts to balance modems across interfaces. With ABR video growing or shrinking to fit the available bandwidth, interface utilization has

diminished in value as a metric to use in load balancing decisions.

Instead a video-aware cable modem balancing application could use the video oversubscription metric to determine which interfaces are more/less oversubscribed to determine in what direction to move the cable modems. Applications can also use the same metrics to evaluate whether the load balancing decisions made were effective in improving QoE or not.

As shown in Figure 3, with the proposed SDN architecture, the Video QoE Application can publish a QoE API, which the Cable Modem Load Balancing Application can use to adjust its decisions to take video QoE into account. This in fact shows the advantage of building applications in a SDN framework and enabling value-added functions across such Applications by leveraging the API exposed by Applications.

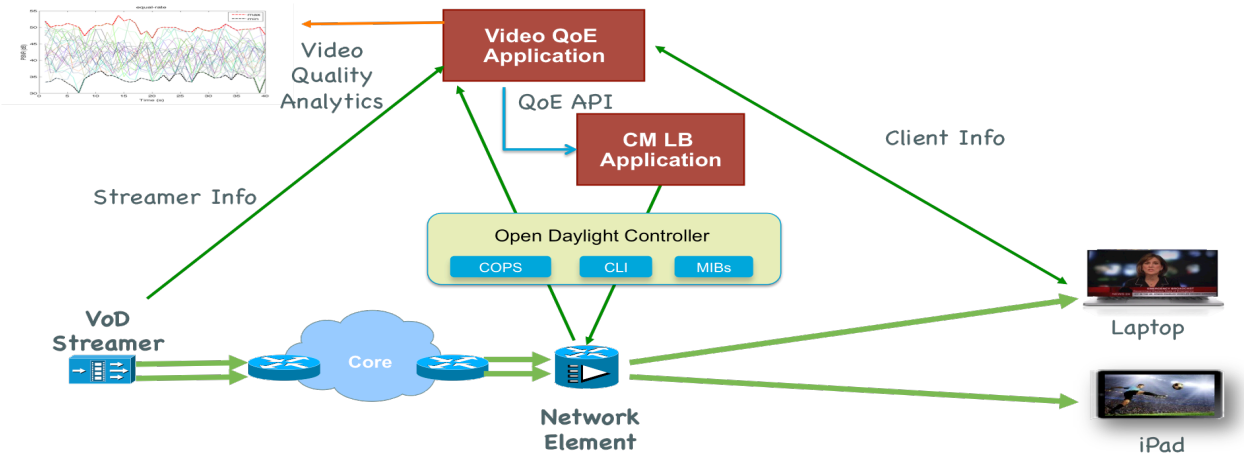

Figure 3 Video QoE Aware CM Load Balancing

Video QoE-aware WiFi Radio Resource Management (RRM)

With increasing number of WiFi devices in the home, and operators' deployment of residential gateways with built-in WiFi, it is becoming more important to manage the WiFi radio resources efficiently to improve subscriber user-experience. As described in [8] WiFi RRM works by collecting various types of information from Wi-Fi Access Points (APs) that it manages and analyses this information and determines whether to apply any changes to the operating parameters of the APs that it manages. Some of the parameters that it may change include Transmit Power, Channel assignment, Channel bandwidth, etc.

The RRM algorithms have generally been envisioned to be fairly static – changing only once a day or few days etc. What we propose in this paper is an enhancement to RRM by which the video QoE information is taken into account in the decisions made by the RRM algorithms. This will likely necessitate more frequent RRM updates, but this can still be done in the order of minutes (not milliseconds). Conceptually this can be viewed as the equivalent of dynamic load balancing as described in an earlier section, except applied to WiFi radio resources. This kind of video-aware RRM can improve video QoE for subscribers.

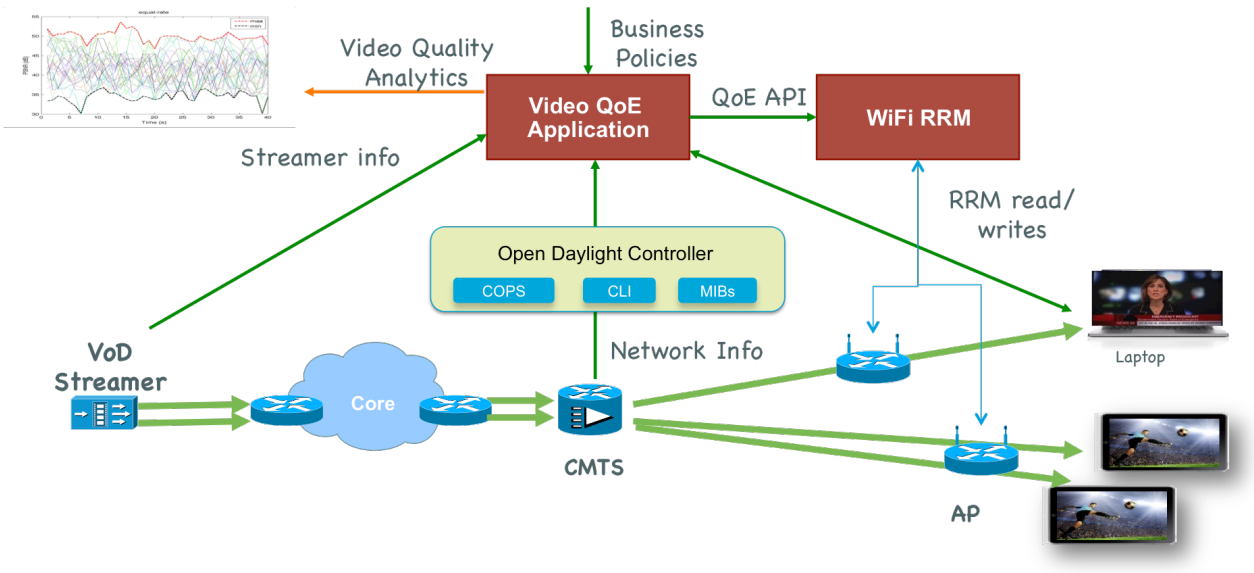

Figure 4 Video QoE Aware WiFi Radio Resource Management

Video QoE optimization over DOCSIS

The proposed SDN architecture can be extended to further optimize the video quality by ensuring fairness in how bandwidth is shared across the DOCSIS interface. In existing HTTP Adaptive Streaming architectures, each client operates completely independently of each other and therefore in a greedy fashion. There is no attempt made to share bandwidth in a fair manner across the competing clients. In the recent past there have been several attempts [9], [10], made in developing client rate adaptation algorithms that share bandwidth fairly across competing clients. In all of the referenced work, fairness across clients is defined as equal bandwidth share for each client. An additional goal of these efforts has been to minimize/eliminate rate profile oscillations of the clients over time.

To compare the behavior of clients with and without our QoE application, we tested 16 clients competing for bandwidth on an 80Mbps link. The clients we used were VLC (an open source client) [11] with a simple rate adaptation logic, and a more sophisticated client with superior rate adaptation logic called Probe-And-Adapt (PANDA) [9]. **Figure 5** shows the rate profile oscillations of VLC clients. **Figure 6** shows the same setup except with the PANDA clients. Clearly the PANDA rate adaptation algorithm is superior to VLC's as seen by the reduced rate oscillations. Next we repeated the experiments with the Video QoE application programming equal rate allocation on the network element, thereby ensuring that each of the flows gets equal bandwidth share on the congested link. The rate profile selections of the VLC client under this scenario are shown in **Figure 7**.

The vertical dashed line indicates the time when the network programming was applied. The contrast is stark in that after network programming is applied, the rate oscillations are almost completely gone. From a rate oscillation and bandwidth fairness perspective this approach far outperforms the scenario with even a highly intelligent client such as PANDA as shown in Figure 6.

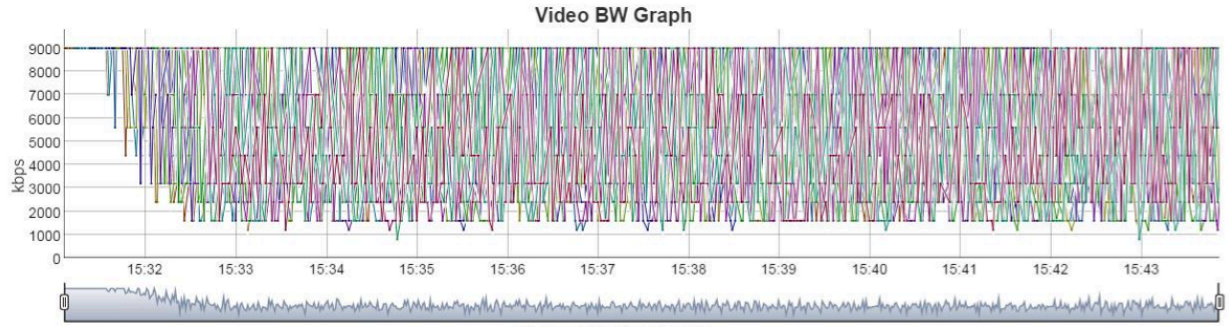

Figure 5 Rate profile selections of VLC clients - baseline

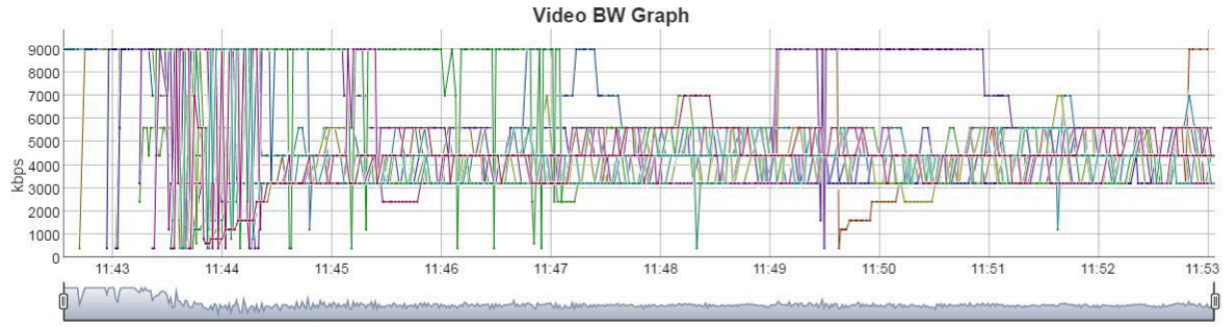

Figure 6 Rate profile selections of PANDA clients – baseline

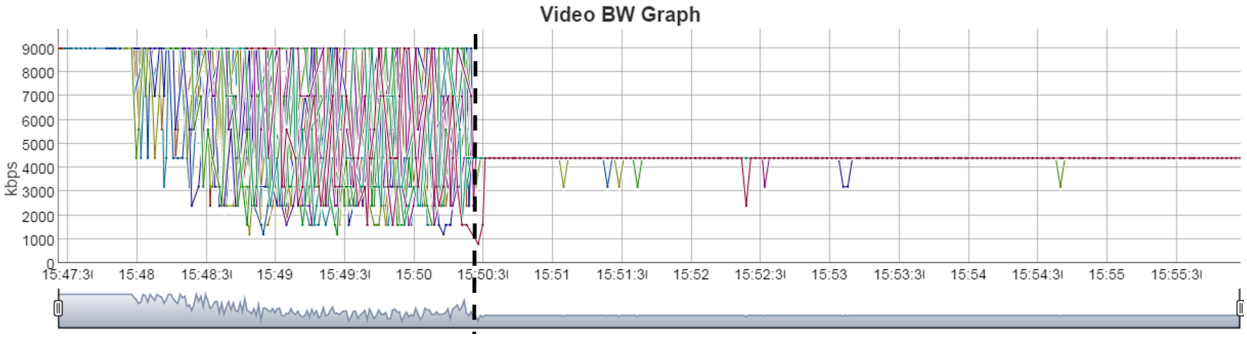

Figure 7 Rate profile selections of VLC clients – with equal rate allocation in network

It is interesting to note that if indeed equal share of bandwidth is an important goal, it is very easily achievable via the proposed SDN architecture. The Video QoE Application can program the CMTS (via PCMM) to allocate equal bandwidth share for each of the streams. This causes the clients to adapt to the perceived network bandwidth. In fact such a straightforward mechanism not only achieves equal bandwidth share, but also helps eliminate rate oscillations in clients' profile selections due to the steady nature of bandwidth available to them. Moreover this architecture can be leveraged to apply business policies that achieve a better fairness model by allocating bandwidth based on content/device type etc.

Beyond equal share of bandwidth, fairness could be defined as comparable perceptual video quality across clients sharing a congested link. Such a goal can again be achieved by our above-defined architecture by re-programming the network at regular

intervals to ensure fairness across clients. That is, instead of bandwidth equally across clients or class of clients, bandwidth can be allocated in a manner to equalize quality across flows. Anyone familiar with encoding technologies will be familiar with the fact that the bits required to encode a video stream at a certain quality will vary depending on the complexity of the video scene being encoded. Or put another way two different videos may achieve very different video qualities even when encoded by the same encoder at the same target bitrate. This property is the basis for the widespread use of variable bit rate encoding in broadcast TV today.

With the migration to Adaptive Bit Rate video, however generally the industry has fallen back to Constant Bit Rate encoding. This loss of efficiency was not a major problem as long as ABR video constituted a small portion of the overall network traffic. Given the popularity of ABR video and its continued growth on DOCSIS networks with the advent of IP video, it is imperative that ABR video operates in as bandwidth efficient a manner as possible.

There have been recent attempts to move in this direction, by large video streaming providers such as Netflix. They describe a per-title video encode optimization method in [9]. This approach is an explicit recognition of the fact that different bitrates are required to achieve the same quality on different titles (content). Hence they propose using different rate profiles for each title. So very simple content (example animated titles) may use lower bitrates compared to more complex titles. While this approach is an improvement over the existing fixed bitrate ladder used in most ABR delivery systems, it suffers from two problems. The first is that the encode may be optimized to that title, but there is nothing in the network that ensures that bandwidth is apportioned fairly per title. Put another way in the face of congestion both flows (complex and simple titles) will settle at the same bitrate, resulting in a much better visual quality for the low complexity title and poorer visual quality for the high complexity title. Our proposed SDN architecture can be leveraged to solve this problem by ensuring the network allocates bandwidth proportional to the title complexity. For example when uncongested, the Video QoE App can allocate the bandwidth for that flow corresponding to the highest bitrate profile available for that title. As the network gets congested (due to say start of busy evening hours) the Video QoE Application can move each of the flows down to the second highest bitrate profile for the title that is being viewed.

The content diversity property is exhibited in the rate-quality tradeoffs depicted in Figure 8 below, where each line depicts the quality of a video encoded at different bitrates. The fact that different lines exhibit different slope means that some videos achieve high quality rapidly at lower rates, while others struggle to reach comparable qualities even at the highest bitrates. Generally speaking the lowest graphs correspond to the most complex videos that are harder to compress and the highest graphs correspond to the simplest videos that are the easiest to compress.

The quality metric we have used is SVQ (Stream Video Quality) a Cisco proprietary non-reference video quality metric, which is lightweight to calculate and has demonstrated via internal investigations high correlations to subjective mean-opinion-scores. The SVQ score ranges from 1 to 10, with 10 being the highest perceived quality. Typically, a score below 6 corresponds to “bad” visual quality whereas a score above “9” is considered “very good”. Irrespective of which quality metric is used, it is well accepted that rate quality trade-off curves do indeed vary from video to video.

The second problem with the per-title encode optimization is that it doesn’t account for the fact that video complexity not only

changes from one title to another, but also from minute to minute within the same title. This is exhibited in the rate-quality trade-off curves in Figure 9, which shows similar patterns as the one above for content diversity, except here the quality depicted is per fragment. So even within 16 consecutive fragments of video (each of 2 seconds) the quality achieved at each bitrate varies significantly.

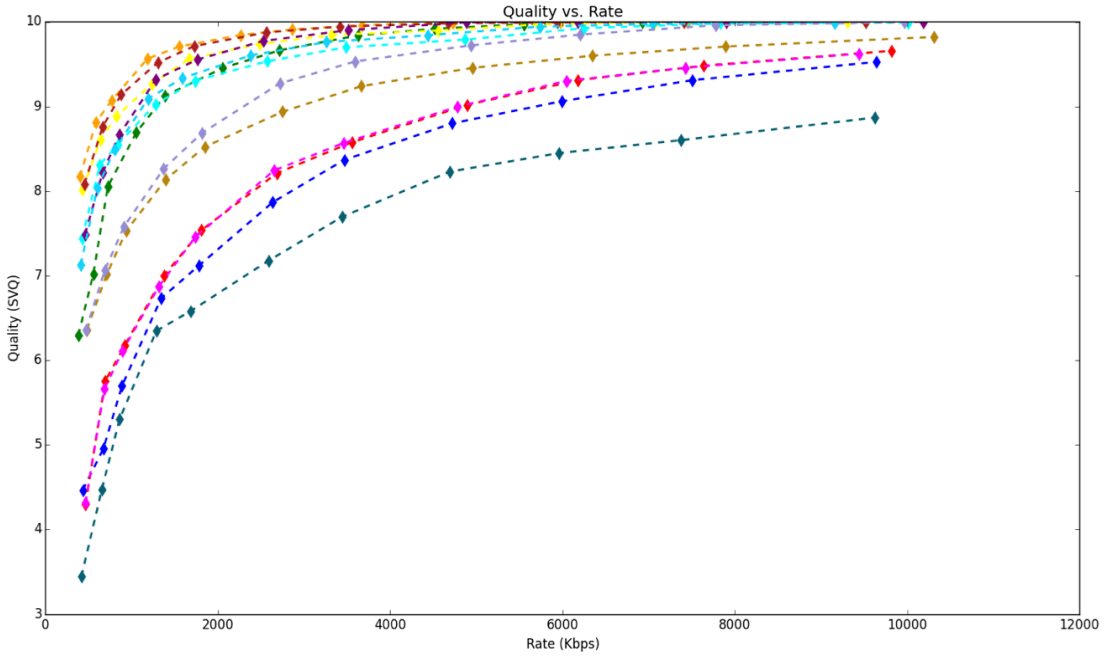

Figure 8 Rate quality trade-off curves for different videos

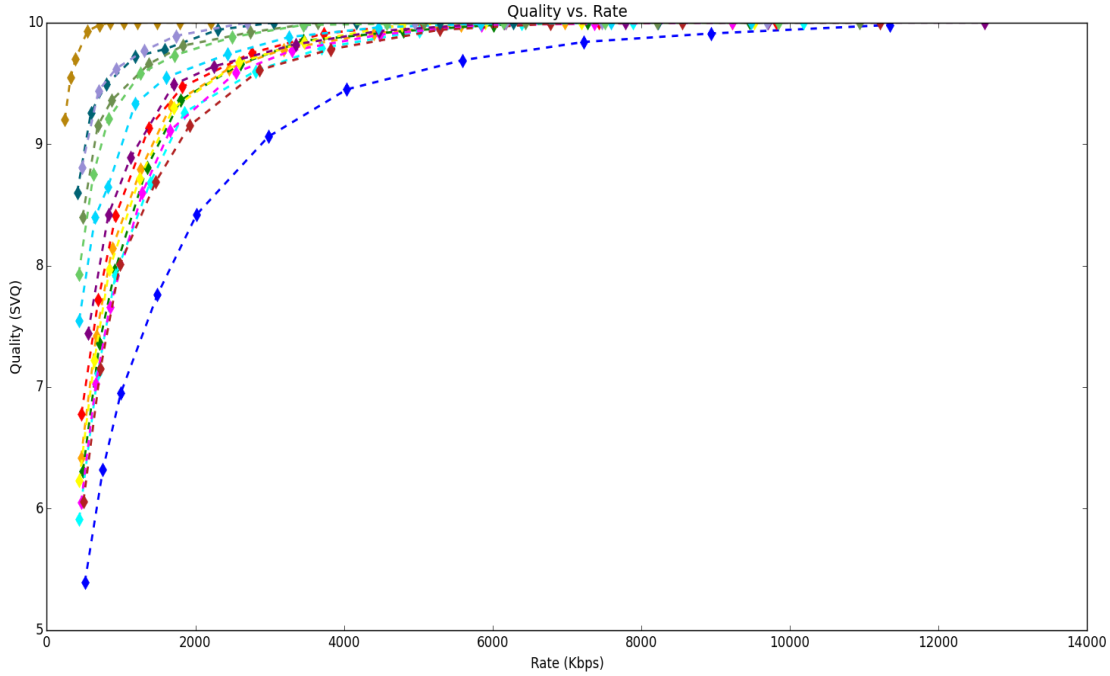

Figure 9 Rate quality trade-off curves for different fragments within the same video

The Video QoE Application can also solve this second problem of content complexity variation over time, within the architecture depicted in Figure 2. By making the Video QoE Application aware of the varying complexity of the video streams via metadata, it can now account for the changing complexity of the content over time, besides the content complexity across the different flows. By taking into account both of these diversities the Application can periodically reprogram the network to account for these differences. ABR clients are already built to adapt to varying network bandwidth, so they will adapt to the network bandwidth made available to it. If it is feasible to modify clients they can in fact be made aware of the decisions made by the Video QoE Application thereby vastly simplifying its bandwidth estimation and associated rate adaptation logic.

We conducted tests with such an optimization method applied in a testbed with a link bandwidth of 100Mbps. We varied the number of clients competing for bandwidth from 15 to 35 clients at any given time. The Video QOE Application was used to collect the minimum and maximum quality across all the clients for each run of the test. We ran 3 flavors of algorithms, where one is baseline, where the Application does not program the network at all, and simply collects and reports analytics. The second case was the "equal-rate" case where the Application programs equal bandwidth for each of the flows on the network element. Finally the third case was where we turned on our optimization algorithm that not only reprogrammed the network but also communicated with the clients to influence their rate adaptation decisions. Results from the above tests are shown in Figure 10. Our optimization method yields the highest minimum video quality compared to other methods at all number of clients. It is not surprising that our method also achieves the lowest maximum quality across all clients. This is in fact fair, because as an operator, the goal is indeed to keep as many subscribers satisfied as possible, which is achieved by ensuring all clients stay as high a quality level as possible without dropping below a threshold. In fact from the graphs it can be observed that, suppose an operator wanted to maintain a minimum quality of 8, they could only pack at most 15 VLC clients on their network, with equal-rate allocation method of Video QoE Application they could pack about ~25 clients on their network, and with the outlined optimization approach they could pack 35 clients on their network. The optimization approach improves stream packing efficiency about 40% over an equal-rate approach. Therefore a significant improvement in stream packing efficiency can be achieved while maintaining video quality. Alternately operators can leverage such technology to improve QoE across their subscriber base.

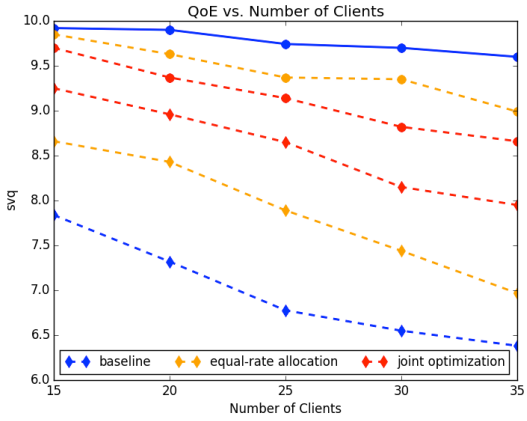

Figure 10 Minimum and maximum quality with varying number of clients

SUMMARY

We have presented methods by which operators can gather video QoE analytics leveraging an SDN architecture. The value of these analytics in capacity planning and troubleshooting has been outlined. Finally optimization methods that leverage the

analytics to significantly improve network efficiency have been presented. All of the above capabilities have been built on top of an extensible SDN architecture.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of various colleagues at Cisco Systems including Joel Schoenblum, Gareth Bowen, Tankut Akgul and Ziv Nuss.

REFERENCES

1. Cisco Visual Networking Index: Forecast and Methodology, 2014-2019 White Paper http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html
2. "HFC capacity planning for IP Video", Sangeeta Ramakrishnan, Cisco Systems, SCTE 2011.
3. History of Move Networks, Available Online: http://www.movenetworks.com/history.html
4. Apple Inc., "HTTP Live Streaming Overview," https://developer.apple.com/library/mac/documentation/networkinginternet/conceptual/streamingmediaguide/Introduction/Introduction.html
5. A. Zambelli, "IIS Smooth Streaming Technical Overview," http://www.microsoft.com.
6. MPEG, "ISO/IEC 23009-1:2012 Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats," 2012.
7. Software Defined Networking, Open Networking Foundation, https://www.opennetworking.org/sdn-resources/sdn-definition
8. CableLabs Technical Report, "Wi-Fi Radio Resource Management (RRM)/Self Organizing Networks (SON) Technical Report", http://www.cablelabs.com/wp-content/uploads/specdocs/WR-TR-RRM-SON-V01-1409261.pdf
9. Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and Adapt: Rate Adaptation for HTTP Video Streaming at Scale," IEEE Journal on Selected Areas in Communications, vol. 32, no. 4, pp. 719–733, Aug. 2014.
10. S. Akhshabi, S. Narayanaswamy, A. C. Begen, and C. Dovrolis, "An Experimental Evaluation of Rate-Adaptive Video Players over HTTP," EURASIP Journal on Signal Processing and Image Communications, vol. 27, no. 4, pp. 157–168, May 2011
11. VLC Media Player, http://www.videolan.org/
12. Netflix Blog, " Per-Title Encode Optimization", December 2015, http://techblog.netflix.com/2015/12/per-title-encode-optimization.html