TCP Rate Control with IFD

Dr. Lawrence Roberts
Anagran Inc.
April 30, 2009

Introduction

Since packet networks were started with the ARPANET[1] in 1969, they have not managed the rate or quality of individual flows (file transfers, voice calls, etc.), instead, they have depended on output queues to restrict the total capacity of traffic passed to a port. Multiple queues have helped provide less loss or delay for packets with various priorities, but this still does not allow control of an individual flows rate or quality. As a result, no flow is protected from delay or loss caused by other flows, and there is no way to provide rate guarantees or priorities for large numbers of projects, departments, or traffic types. Thus, we still have major quality problems with voice and video, serious traffic imbalances due to unfairness (e.g. P2P), and substantial delay variation in interactive activities like web access.

Most of the quality problems that impact IP traffic are caused by TCP since TCP flows generally grow until the network takes some action to signal them to slow down. The general method of signaling to TCP senders to slow down is packet discard in the network. This paper will examine the behavior of TCP flows and their control by network equipment. Three methods of control will be considered; queues with tail drop, queues with Weighted Random Early Discard (WRED), and Intelligent Flow Delivery (IFD).

IFD is a new method of carefully managing the rate of every TCP flow so as to insure the total traffic does not create any overload, no delay or delay jitter is added, and TCP flows run smoothly without any major rate variation or stalls. Just like Tail Drop and WRED, TCP packets are discarded to signal the sender to slow down, but the difference is in which packets are discarded.

The benefits from this new approach to packet traffic control are numerous and allow networks to finally support extremely demanding new applications, eliminate the artifact and distortion in voice and video, improve the performance of data applications, guarantee capacity for priority applications, relieve networks from the imbalance caused by multi-flow applications like P2P, and allow each project, department, or application to share a common network without being impacted by the other classes of activities.

Network Overload Control Today – WRED or Tail Drop

The method currently used to allow IP networks to manage potential overload at any choke point is a balance between the specified action of TCP in the end-user equipment, and the delay and discard performed in the output queue of the network equipment.
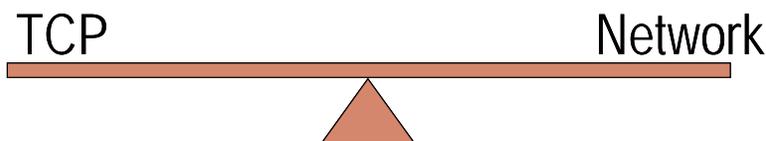
Figure 1: TCP and the Network operate to maintain load balance

TCP normally keeps increasing its rate until it encounters delay or packet loss, at which point it slows down. Output queues add delay which may be sufficient, and if not, they discard packets when the queue fills. These actions are sufficient to ensure that some of the TCP flows will slow down sufficiently to not overload the choke point. Briefly, the rate for a TCP flow between computers with typical buffer configuration is approximately:

$$\text{TCP Rate (Mbps)} = 376 / (1 + \text{Loss} / 0.5\%) / (\text{RTT} + 4)$$
Where RTT=Round Trip Delay in ms and Loss is the packet loss %.

Inspecting this, one can conclude that TCP has a maximum rate limited by the loss and the RTT. Typically the upper limit is 20 Mbps for no loss and small RTT's like 15 ms. However, most trunks carry lots of flows and together they will tend to fill any trunk until queuing delay and loss do occur, at least occasionally. Measurements every five minutes will usually not show overload, in fact a trunk with an average utilization of 50% will often have many short overload peaks causing delay and packet loss. Only in the core, where the traffic is statistically smooth, can overcapacity prohibit any momentary overload and thus also avoid any packet loss.

If a flow does not happen to burst when several others do, it may avoid any discard and continue to grow. But if a flow does happen to burst when others do, it may lose several packets and stall. Stalls occur when the TCP sender is informed by the receiver that several packets were lost in the same round trip cycle (one RTT). Then the TCP design has the sender stop totally for a timeout period. Thus, the impact on an ensemble of flows is to cause a wide spread in flow rates, some stalled, and some growing to their maximum rate. The graph in Figure 2 shows a trace for each of 50 flows being controlled by a WRED queue into a 20 Mbps link. Note that each flow increases until it happens to suffer one or more discards, and then its rate drops and the process repeats.
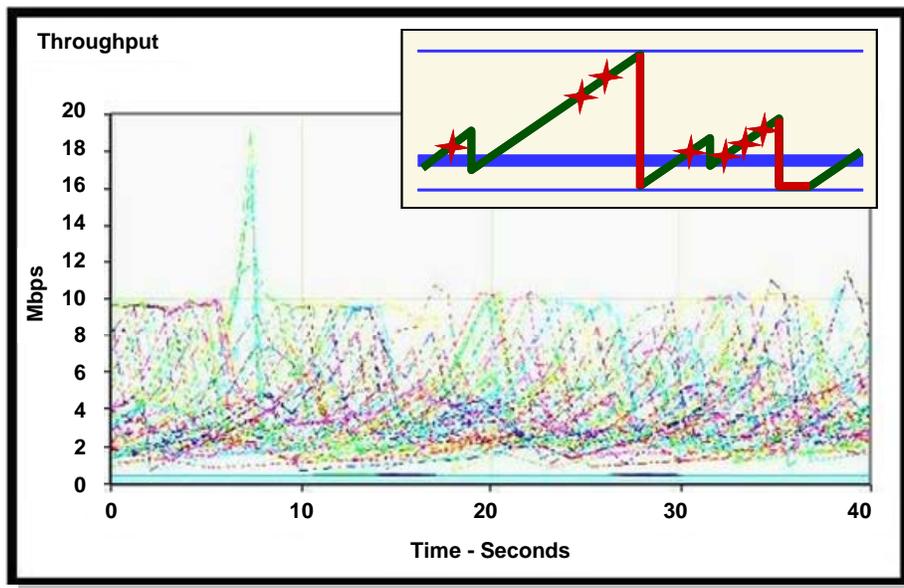

Figure 2: 50 TCP flows controlled by WRED

Looking at the drawing in the upper right corner, losing one packet cuts the rate in half, two sends the rate down to slow start (rate cut to near zero), and 3 or more packet discards in a cycle causes a stall (zero rate for a timeout period). Looking at the collection of flow rate traces in Figure 1, many of the rates drop to slow start and some are stalled, sitting at zero rate. This behavior is typical of both WRED and Tail Drop queuing since as the TCP rates increase and packets overflow the queue buffer, many packets from the same flow can be discarded at once. WRED improves somewhat on Tail Drop since packets are discarded randomly from the queue before it overflows, leading to less multi-packet drops for the same flow. However, as can be seen in figure 2, the problem still exists.

Another result of WRED or Tail Drop load control is to cause considerable delay and delay jitter in all the traffic. When packets build up in the output queue, the delay incurred can be anywhere from 1-40 ms. When a voice flow and its ACK return encounter two choke points in each direction, the delay can vary by four times 40 ms or 160 ms. This magnitude of jitter results in poor voice quality. Usually voice is put into separate higher priority queues to reduce the jitter. However, if other traffic attempts to look like voice, the result will be the same as if one queue was used. The delay jitter also impacts live video and here the rates are so high that extra queues are not the solution; often a separate network is required. A second problem exists for Video on Demand which usually uses TCP for error free delivery. By storing enough video before starting, the impact of packet loss and delay jitter can be minimized, but the data rate must keep up with the video. Since TCP slows down with delay and loss, the actual delivery rate for video may be slowed below critical. This causes frame freeze until the slip buffer is refilled. The combination of moderate network and queuing delays combined w/ queuing induced jitter makes delivery of HD video over TCP impractical. A simple example of this is a 4 Mbps HD video being delivered over 750 miles but with two 40 ms choke points in the round trip path. The maximum data rate would have been 26 Mbps with no congestion or queuing delay, but when the queuing delay occurs, the maximum rate falls to 3.9 Mbps and the video stalls from time to time. Thus, the absolute output queuing delay can be a large problem as well as jitter.

The Alternative – IFD

To address the growing volume and variety of network traffic and the inherently unpredictable nature of network usage in general, Anagran has developed an advanced Flow Rate Manager which monitors and rate-controls all network traffic across up to four 10 Gbps Ethernet connections at line rate. A flow is simply an end-to-end network activity such as an image download, a voice call, a video stream, or a part of a web page retrieval. The Anagran Flow Rate Manager is an extremely compact, low power system that has perfected the ability to precisely control the rate of every TCP flow individually. Through continual, rapid measurement of the output rate on every port, VLAN, and class, the system can intelligently adjust the rates of every flow to:

- Eliminate overload at upstream or downstream choke points (no jitter or loss)
- Guarantee quality for streaming media (IPTV, voice)
- Improve interactive response time for web traffic and gaming
- Allow more traffic to flow with quality over a given trunk, or the same amount of traffic over a lower speed trunk

Flow Manager Design

The design of the Anagran Flow Manager is unusual since the rate control is at the input, not the output.

Input Process: Looking at Figure 3, data arrives at an input port and a flow identifier is extracted. For IPv4, this is the five fields; source address, destination address, protocol, and if available, the source port and destination port.

Flow State Memory | Assign Rate, QoS, Output Port, & Class | Processors

Input | Load Measurements | Output

Switch

Discard

Rate of Each Flow Controlled at Input | Traffic measured on both the output port and in up to 4000 Classes
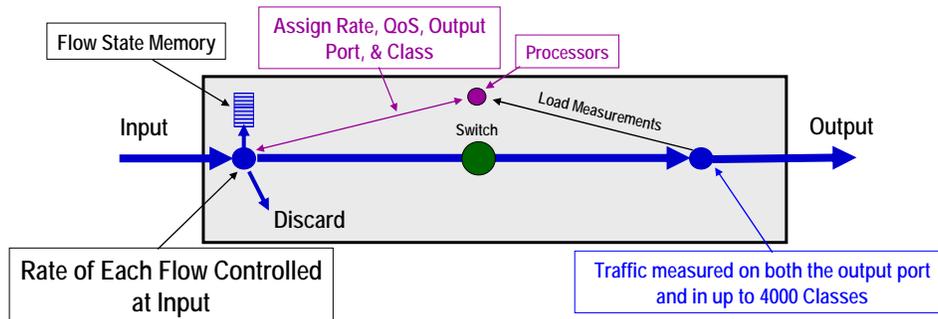
Figure 3: Flow Manager Structure

These fields are hashed together and looked up in the flow table. The lookup is extremely fast since it is an exact match, and can easily be done at line rate (10 Gbps) with the shortest packets. The first packet of a flow will not match an existing entry. When there is no match, the packet header is sent to the central processor to determine the destination, flow type, and "fair rate". This information is returned and the packet is forwarded to the destination. If there is a match, the flow record is updated with the new information (bytes, time) and the current rate is recomputed. Comparing this rate with a "fair rate" a rate control decision is made.

Fixed Rate: If the flow type is "fixed rate" (like much UDP) then the "fair rate" is a maximum, above which packets will be discarded. However, so long as the flow stays below this maximum rate, all packets are forwarded over a high priority path. If there had not been capacity for this maximum rate at the output, the NPU would have returned "reject", all packets received for this flow discarded, and an ICMP packet returned to the sender indicating "no path", which tells the application to try again later. This should be extremely rare but is necessary to protect the output from overload. Fixed rate flows like video and voice are thus guaranteed minimal delay and no loss once accepted.

Available Rate: If the flow type is available rate (typically TCP) then the first time the flow exceeds the "fair rate", one packet will be discarded to tell the sender to slow down as shown in figure 4. Since the sender will not learn about this discard for some fraction of a Round Trip Time (RTT) then no more discards are made for a RTT.
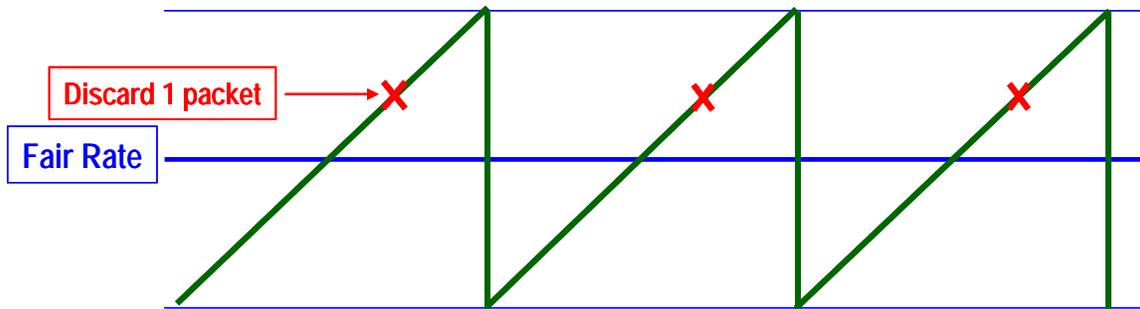
Figure 4: Available Rate Discard Process

This process is substantially superior to the random discards of the queue based routers and switches. TCP only sees a single discard and thus goes to half rate and then increases its rate linearly. With random discards, there is a high likely hood of more than one packet being discarded and then TCP reverts to slow start or stalls. The stall period increases exponentially with repeated multiple packet losses, so in more highly congested systems the stall can become minutes long, enough to make one give up. Also, some popular operating systems allow multiple stalls to totally kill the flow, requiring the user to start over. Also, the percentage of packets discarded is lower than in the random case, often about 0.5% when policing is required.

Note: this mechansim works with any protocol that has a congestion avoidance mechanism including encrypted P2P protocols.

Switching: Accepted packets are switched through a non-blocking fabric to the designated output. One benefit of input rate control is that the traffic entering the switch fabric is well controlled and unlike conventional output policing systems, both the input and the output traffic can be assured to be limited to the port rates, except for small momentary overloads as predicted by queuing theory for random arrivals. The fabric has sufficient queuing for the output overloads of perhaps 20 packets at 95% utilization. Thus no input or output queues are necessary.

Output Process: When a packet arrives at an output module, it also has a flow table which has been filled when the flow was originated. This table provides information as to which port the packet is destined for, encapsulation information, flow type, VLAN info, class info, and perhaps a modified DSCP value. After being properly formatted, the packet is passed to the output chip and its byte count is added to either the fixed rate or available rate counters for its class, VLAN, and port. These counters are processed every 50 milliseconds to determine a "fair rate" for the available rate counters. Each port, VLAN, and class has a configurable maximum rate (perhaps less than the port speed) and the available rate traffic must fit within the capacity left between the fixed rate usage and the maximum. The process here is quite complex but the bottom line is that fair rates for each virtual segment (class), each VLAN, and each port are computed so as to not load any of these more than 99% and to average about 92% utilization so long as traffic is available.

Central Processor: All the measurements from the outputs flow back to the central processor. It also receives the initial flow setup requests from the inputs as well as update requests every 128 packets or every second for every flow. For a new flow it first examines the packet header to determine the best port, VLAN, and class. 8,000 ACL commands are available to specify this as well as a full layer 3 routing table (in routing mode). Another step is to compute or re-compute a "fair rate" for the flow. This is determined based on the measured fair rates for the port, VLAN, and class it is in, and modified by a priority which adjusts the "fair rate" it receives. The priority can be specified by the ACL commands, the class it is in, and the activity/status of the subscriber

address. Thus, every flow receives a fair rate which will ensure the port, VLAN, and class are all operating at the highest utilization, under 100%, that they can.

Intelligent Flow Delivery (IFD) Performance

With IFD controlling all the TCP flows to maintain a fair rate (based on total traffic and priorities), the flow rates only vary 2:1 and do not stall. Using the same test as shown in figure 2 for WRED, figure 5 shows the 50 flows all operating smoothly at about 400 Kbps and not stalling.
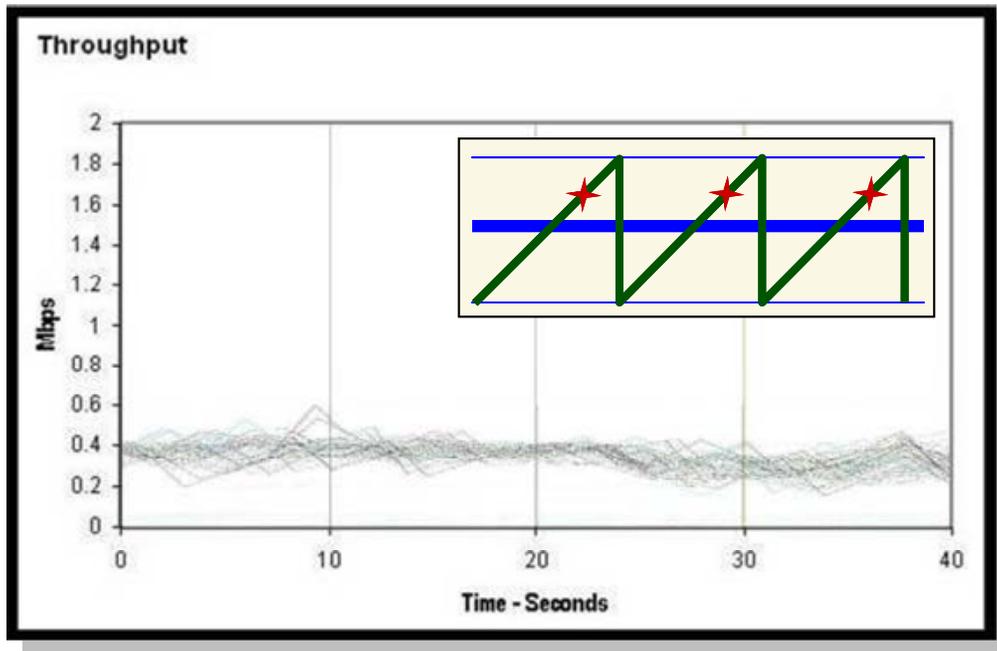

Figure 5: 50 TCP flows controlled by IFD

In this example, the flows all have the same priority and thus operate at approximately the same rate. This dramatically improves the perceived responsiveness for activities like Web browsing where many parts of a web page all must be received before the page can be displayed. With exactly the same total throughput, if a page has 40 parts and uses 40 flows the probable time for all 40 to complete is the important measure. Then the page can be displayed. In the WRED example in figure 2, the average time for 40 flows to complete is 3 times as long as with IFD where all flows are the same rate. One stalled flow can delay the whole page. This is explored further in the next section.

In the even more annoying case where the operating system drops flows if they have too many stalls, the performance improvement is that the pages always complete faster, but even more significantly, there are no "hung" web page requests. This is a major benefit.

Web Access Response Time

Measuring the effect of a router's output queue on TCP flows that load the queue, all of which are sending the same size block over the same path, reveals a broad distribution of block

completion times. Some flows get through with no discards and some get hit hard and are very slow. For typical web traffic, a web page must wait for the last of 20-40 flows to complete.
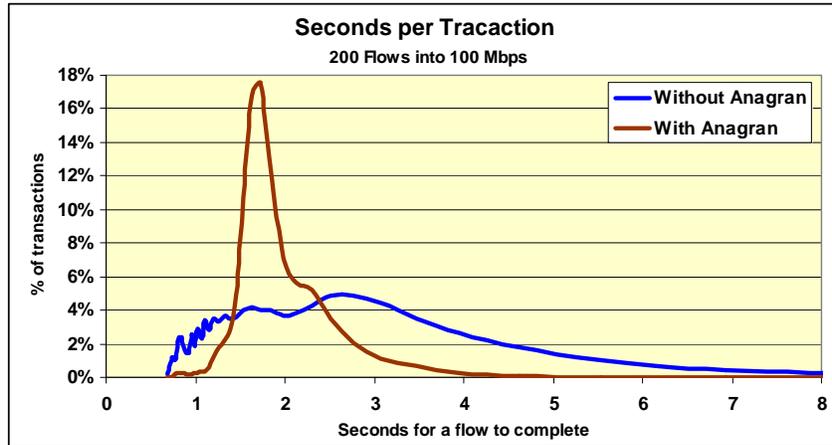

Figure 6: Histogram of Web Access Time (From lab tests)

As can be seen in Figure 6, the standard output queue delay/discard process leads to a very broad distribution of web access times, from 2/3 of a second to 14 seconds. Both distributions have the same average transaction time, 1.7 seconds, but very different variances. Thus, with load control using an output queue, some flows achieve a high rate, but at the expense of others which are stalled and complete very slowly. With this broad distribution of delays, the web page gets completed three times slower than when the Anagran flow manager is inserted into the path. This is because the Anagran system controls all similar flows to run at the about same rate, with a very narrow distribution of block transfer delay. Thus all blocks arrive at about the same time and the page completes three times faster with the same average and peak traffic. The difference is the flows are treated fairly, and to the end user, the web page retrieval is a much faster, real-time experience.
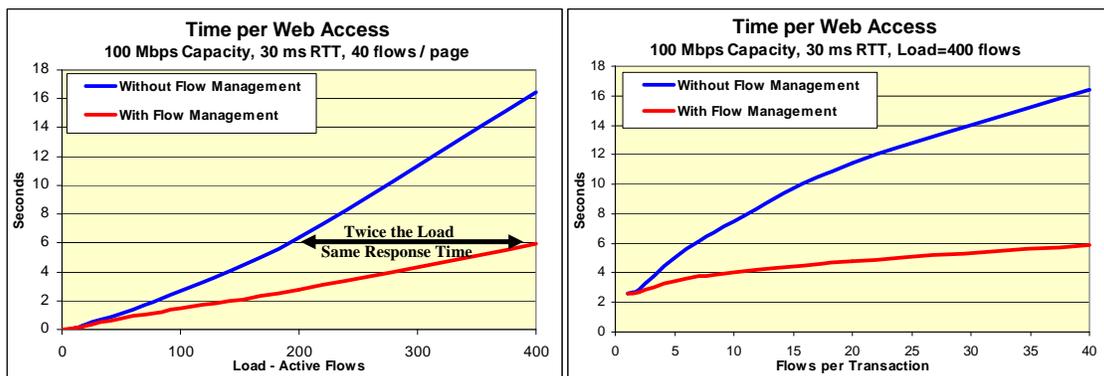

Figure 7: Laboratory test results with a 100 Mbps path with a RTT of 30 ms

The test results when the test traffic is being sent into a 100 Mbps router port is shown in Figure 7 when the router is alone (blue), and when the Anagran Flow Manager has rate-controlled the traffic before the router to just under 100 Mbps (red). With 40 segments per web page and only

400 active flows, the TCP traffic expands to fill the trunk and, as expected, is controlled to about 92 Mbps. However, this is sufficient to cause the average web page to take 3 times as long as it would if all flows are controlled to be at the same rate. Additional experiments showed that if the Anagran controlled the total peak rate to 50 Mbps, the web page response time would be the same as with the conventional router at 100 Mbps. The corollary to this can be observed on the right in Figure 7 where the response time is the same (6 seconds) when the load with Anagran is twice (400 flows) the original load (200 flows). These results demonstrate the power of equalizing the rates of interactive traffic since the *peak capacity requirement for interactive traffic can be cut in half or the response time tripled*.

Video & Voice IFD Processing

Video and voice are either classified as fixed rate or available rate in the Flow Manager. Classification is done based on the header of the first packet or on the flow behavior after monitoring the first portion of a flow. For example, voice may have a unique DSCP code or it may be identified after the flow has been monitored for a shot period and noting the average packet size is small and the flow rate stabilizes at around 100 Kbps.

Fixed Rate

If the traffic is to be viewed in near real-time it should not be slowed down. VOIP uses the protocol UDP which generally tells the network that this stream is real time and should not be delayed or discarded. However, except in expensive routers with multiple queues, UDP is delayed and discarded along with TCP traffic when overload occurs. Both delay jitter and packet loss hurt voice quality thus voice quality is often poor when voice is mixed with data traffic and only one queue is available.

With Flow Management and IFD, streaming video and voice are usually sent "fixed rate". Fixed rate only means the peak rate is fixed, the short term average can vary significantly. Capacity is not guaranteed for the peak, instead the average of all the fixed rate flows is measured and a new fixed rate flow may optionally *not* be admitted if its peak rate plus the current average would exceed a threshold. In this very unusual case, an ICMP message is returned to the sender stating "no path available, please try again". If the actual average peaks slightly into the remaining capacity, the "available rate" traffic is slowed slightly, keeping the video and voice lossless.

IFD does not queue traffic to slow it down, instead it is continually controlling the rate of every available rate flow to keep the output within the configured rate. Thus, instead of introducing variable delays of up to 40 ms, the maximum delay is in the microseconds. This tiny delay does not impact video or voice, but multiple delays of 40 ms does.

Thus, streaming video and voice do not suffer loss or delay from the sudden capacity reduction which occurs at the network edge as the data enters the expensive last mile link or WAN link. Instead the fixed rate traffic is streamed through and the available rate flows are all rate controlled to ensure the switch or router at the edge is never overloaded.

Available Rate Video on Demand

Video on Demand (VoD) typically uses TCP protocol which was designed to let the network set its rate. TCP keeps increasing its rate until it either is limited by the round trip delay or it receives a discard from the network. Routers tend to add delay in their queues first before discarding. This reduces the discard frequency since TCP's natural rate limit is proportional to the inverse of the Round Trip Time (RTT). Adding 40 ms to a 1000 mile path each way increases the RTT from 16 ms to 96 ms which in turn decreases the TCP flow rate in a normal PC from 20 Mbps to 3.8 Mbps, a five to one decrease. This is effective for rate control, but adds 80 ms of jitter to the stream and for higher rate video, may limit the maximum delivery rate. HD TV typically exceeds 4 Mbps and when it does, an RTT of 96 ms would cause a freeze frame event. However, buffering can make most standard rate video work well except when discards occur. One discard per round trip time slows down the flow rate to one half and it build back up. This is normal and does not cause a problem since the average rate can be maintained. However, two discards can cause a return to "slow start", a return to the slowest rate and slow growth back to full rate. For HD video this would take 600 ms in our 1000 mile example with no router delay and 6 seconds with the added 80 ms of router delay. Thus with 2 discards and delay (which generally is related) there is likely to be a frame freeze. However, this is not the worst event since when a high rate TCP flow bursts, it sends many packets rapidly. Then when the router queue fill, if it is one of the unluckily one to just arrive with a burst, it may lose three or more packets. Then TCP goes into stall mode and waits one to many seconds before starting slow start. This is a stall, which many of us see on Web accesses at some reasonable frequency. This can cause a substantial freeze of the video.

Hopefully this helps to show how as video moves from YouTube to HD the problems with TCP video will increase rapidly. However, this is easily fixed with IFD since it controls TCP in a totally different way. Instead of all flows being mixed into one or more delay queues, each flow is controlled separately. This is done by computing the fair rate for each flow based on measuring the output rate and dividing it up between the flows based on their priority. If a flow exceeds its fair rate by a margin, one packet is discarded to tell it to slow down. Multiple packets within an RTT are never discarded so it goes to half rate and grows back. This process keeps each flow at the right rate such that the downstream choke point capacity is never exceeded, but if traffic is available, is loaded to over 90%. Without added delay, slow start or stalls, the problems cited above do not occur and HD video could be streamed 5600 miles, across the US.

If the desire is to protect VOD from all speed related problems (such as HD across the world) then fixed rate can be used even though the protocol is TCP. The DSCP or source address must somehow flag these flows, but once identified, they can be treated as fixed rate with any appropriate peak rate. Generally the sender is carefully controlling the flow rate so the TCP is being source controlled and will not grow beyond the peak rate.

Test Results

Tests have been conducted with mixed data, voice, and video test traffic going through a router into a 100 Mbps trunk. One case is run with the traffic arriving on a 1 Gbps trunk to the router. In the second case the traffic first goes to the Anagran Flow Manager and then to the router.
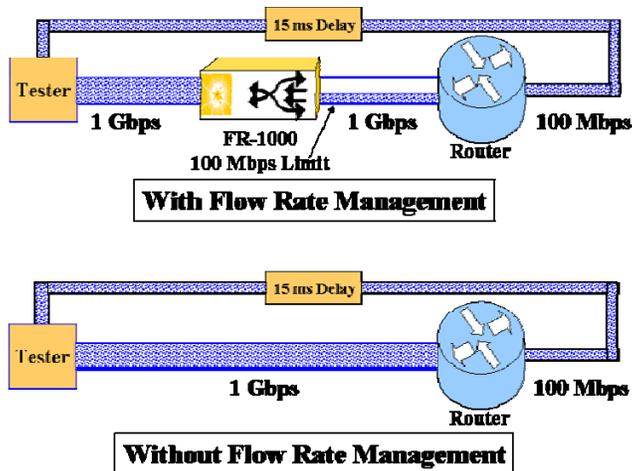
Figure 8: Test Setup for Video, Voice and Data test

The test generators in this test measured the video loss and reported it in each case. Voice traffic measured the Mean Opinion Score (MOS) which looks at delay jitter and loss. Data traffic was increased from 0 to 100 flows. At about 15 TCP flows, their natural rate became high enough to start causing router discards and from there on things get progressively worse without Flow Management.
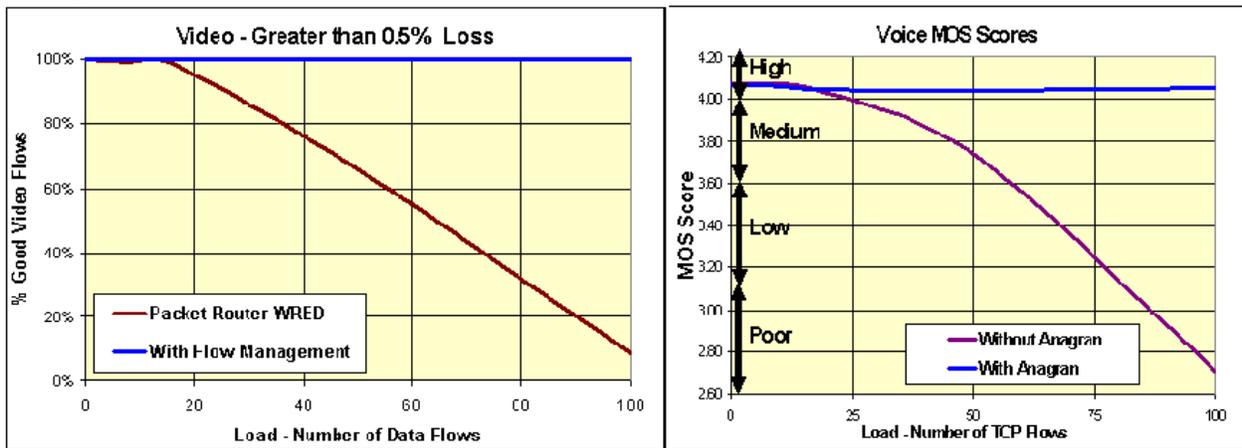

Figure 9: Test Results for Video and Voice

As the video graph shows, the % of good video (less than .5% loss) becomes so bad with only 100 data flows that virtually no video is acceptable quality whereas with Flow Management 100% stays loss free. For the voice, the MOS score stays high quality with Flow Management, but without it, the MOS score fall rapidly, again to all poor at 100 data flows. This shows how the video and voice can be maintained at superior quality even when mixed with data and encountering a choke point at the network edge.

Comparison with Multi-Queue Systems

Historically, high QoS networks have been designed using high end routers with multiple queues and Weighted Round Robin (WRR) priority queuing. This is expensive compared to best effort switches or routers. When properly tuned, these networks will keep the video and voice from having the same loss as best efforts traffic. However, the tuning is a continual task since too

much video in the video queue can still suffer loss. Also, delay jitter is not eliminated but will be reduced. The main problem is the CAPX and OPEX is far greater than using best effort equipment plus the Flow Management. Also, the quality is often not ideal with most multi-queue schemes since the video is still subject to delay and loss although less often.

IFD and Call Admission Control (CAC)

Perhaps the most important difference IFD makes for video is that too many video flows on any link will cause even a good multi-queue system to discard packets from all the video flows. For example, the addition of the 26th HD video to a 100 Mbps DSLAM or WAN trunk will cause all videos to have a 4% packet loss (poor quality) and adversely impact all data traffic.   IFD solves this broblem by rejecting the last video that arrives which would cause such degradation. This concept of Call Acceptance Control (CAC) has always been available on circuit switched systems, but never on packet systems. This is a protection which is critical with the growth of video.

Conclusion

Anagran's Products have strong capabilities to apply priorities to the rate of each flow to automatically adjust the mix of traffic to optimize the bandwidth consumed and ensure quality. The key functions involved are:

- **Equal rate per similar flow** decreases the time per web access to 1/3 or allows the peak rate for interactive to be cut in half with the same response time
- **Intelligent Flow Delivery (IFD)** allows precise rate control of every flow including applying a priority
- **Tiered Behavioral Traffic Control (BTC)** commands can apply lower and lower priorities to a flow as it grows in size
- **Flow Priority** is a major new network capability where a low priority flow gets capacity only to "fill the higher-priority valleys"
- **Virtualization** is a new capability where a group of flows for projects or application areas are each put into classes, thus allowing guaranteed capacity, maximum capacity, and/or different priorities per virtual group.

Applying these basic capabilities to a typical mix of traffic allows anywhere from 20% to 50% reduction of the peak rate which means either more traffic on current trunks, or less capacity required for the same amount of traffic. In fact, using IFD insures that there is no random packet loss or delay jitter at any load, thus if a major overload does occur, all TCP traffic still operates smoothly without stalls and the voice and video still maintain full quality. All flows will be slowed down but the quality of TCP transactions, voice and video is maintained.