Flow Rate Management

Dr. Lawrence Roberts
Anagran Inc.
September 24, 2008

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.

Flow rate management is a new method of designing network equipment where these problems are eliminated through the use of per flow rate control. If every flow through a system is carefully rate controlled, the total traffic capacity for each destination address, each flow class, each VLAN, and each port can be controlled to not exceed a specified rate limit. Not only can the maximum rate be controlled, but if the traffic exists, the utilization of each of these categories can typically be held between 90% and 99%. This can be done for any overload without causing delay jitter and random packet loss, the usual result of any overload in packet networks today.

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 – "TCP/Output Queue"

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.

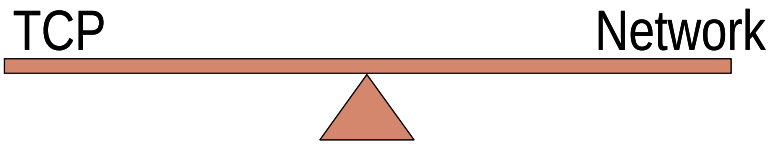


Figure 1: TCP and the Network operate to maintain load balance

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:

TCP Rate (Mbps) = 376 / (1+Loss / 0.5%) / (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 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. 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. However, due to large flows having more packets at risk, there is a global effect where the high rate flows decrease their rate and the low rate flows increase their rate, leading to the average result of **equal capacity per flow**. This is only true for long term averages and the variance can be very large.

This result (equal capacity per flow) was not specifically planned, but was accepted as a good result in the 1970-1990 timeframe. At that time both voice calls and remote data access to a computer both used one flow each direction per person Thus equal capacity per flow resulted in equal capacity per person, a reasonable and attractive result. Today however, computers generate the flows and there is no fixed relationship between one person and one flow. The computer application can just as easily use 100 flows and thus receive 100 times the capacity of more typical applications like FTP which use one or a few flows. This unfairness has been utilized by P2P applications and could soon be used by other applications if not fixed. This can be fixed using flow management and will be addressed later.

Another result of "TCP/output-queue" 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 is considered to be poor voice, thus 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 los and delay jitter can be eliminated, 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. Even moderate distances plus network queuing delays can easily make HD video impractical to

deliver with TCP. 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 – Flow Management

To address the growing volume and wide 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 flows 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. Extremely compact and consuming very little power, the Anagran Flow Rate Manager is the first 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

This paper will lay out the current and potential ways this rich control system can be used to minimize the total peak capacity required in the presence of a wide variety of traffic, thus permitting more traffic to be placed on any trunk while at the same time preserving the required QoS for each type of traffic.

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 2, 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.

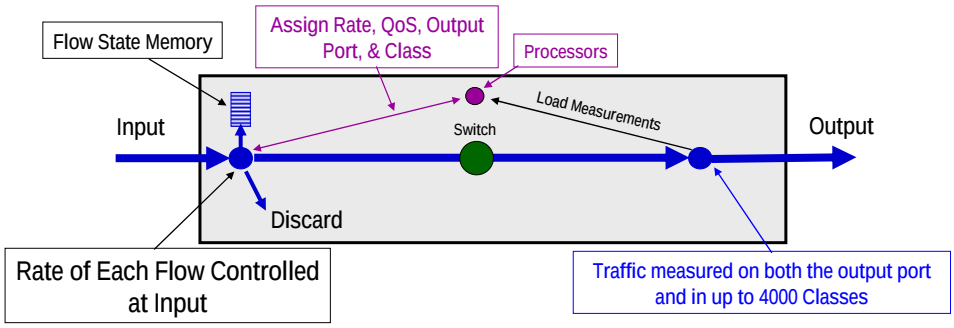


Figure 2: Flow Manager Structure

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. If 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 (typical TCP) then the first time the flow exceeds the "fair rate", one packet will be discarded to tell the sender to slow down. 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.

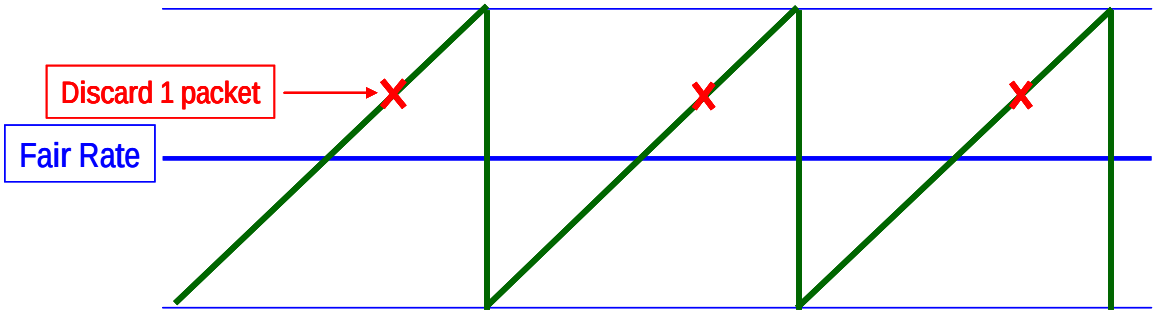


Figure 3: Available Rate Discard Process

This process is substantially superior to the random discards of the conventional output buffer. 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. (I see this often on my broadband service). Also, the % of packets discarded is lower than in the random case, often about 0.5% when policing is required.

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 queues are necessary.

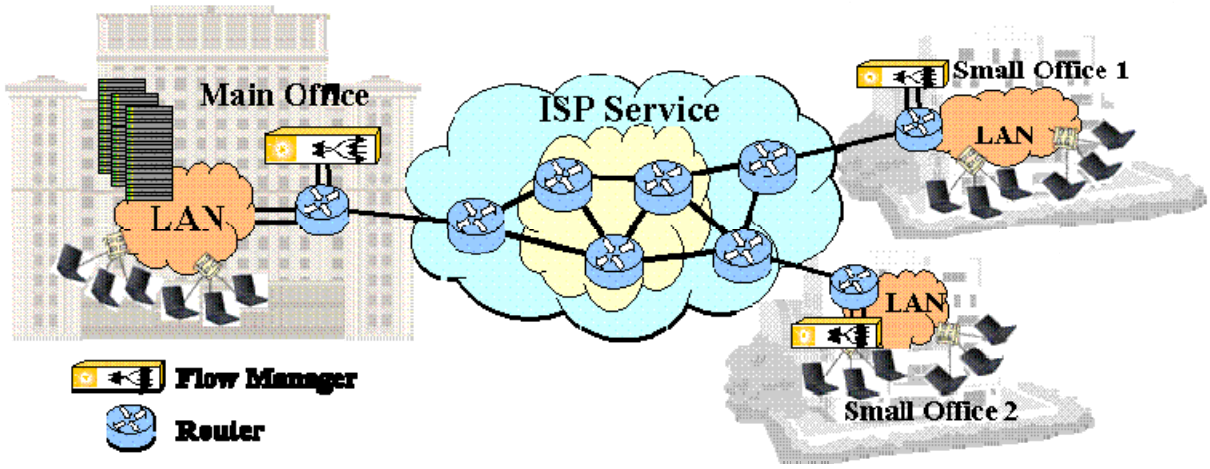
Output Process: When a packet arrives at an output module, it also has a flow table which has been filled when the flow was originated. This table provides information as to which port the packet is destined for, encapsulation information, flow type, VLAN info, class info, and perhaps

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.

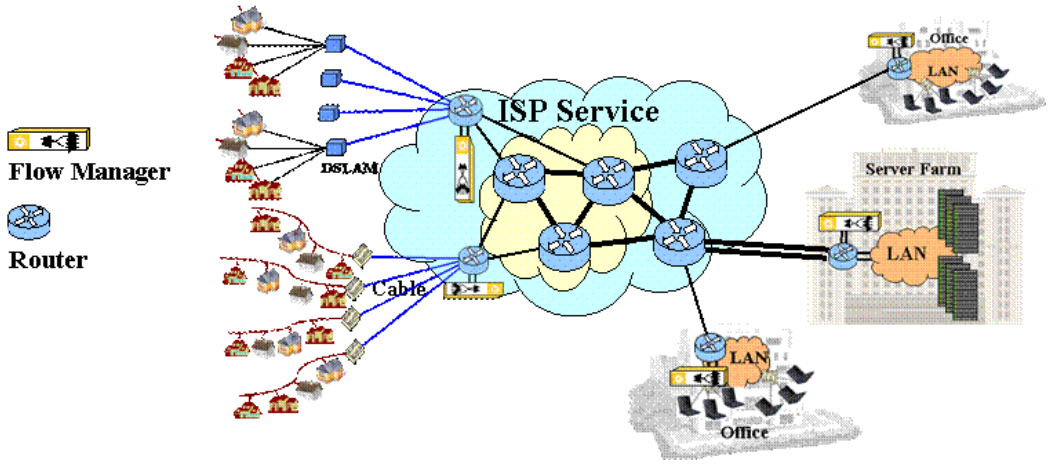
Where Flow Management Goes in the Network

Flow Management controls overload in any IP network. In networks today, over-provisioning is used in ISP and large private network cores as an expensive means to curb the ill effects of network overload. It is also used in homes and in corporate LANs and WANs. Since virtually all congestive overload occurs at the network edge, that is where Flow Management is required. Since it takes more than 2:1 and often 3:1 overcapacity to avoid overload, over-provisioning is expensive and typically not feasible at the ISP edge or the corporate edge. In the Core or the LAN overcapacity is typically used and is effective. Thus, the area where we find quality loss is the network edge, and this is therefore where Flow Management needs to be added to the network. Since the Anagran flow manager only needs to be at these two places, it improves the maximum useable capacity of the expensive links at the edge, and saves money when added to the network while improving and preserving quality. In a corporate network, the edge is at the router supporting the links to the WAN or to ISPs.



Flow Management in the Corporate Network – At the edge

In Internet Service Provider (ISP) networks the edge is where their home users connect to the network. In corporations, flow management works best inside the corporation at server farms, corporate aggregation sites such as regional campuses, and over links to ISPs. It is usually ***not*** needed within the ISP.



Flow Management in the ISP – At their home user edge

Flow Management Operation/Benefits

Flow rate management provides a number of new tools, never before available, to recognize and control the rate, quality, and responsiveness of different classes of traffic. In all cases the results with Anagran will be compared to systems where the gross traffic level is being controlled with packet routers or switches where the control of TCP traffic is done using output queues.

TCP slows down with delay or discards. If too much traffic arrives, routers and switches put the traffic into an output queue to first delay it and if this is insufficient, to discard packets. However, both actions are harmful to TCP traffic and of course also to any voice or video that happens to be in the queue. For TCP traffic, delay slows all the traffic, even the low-rate flows that are just short interactions. It also discards multiple packets from "unlucky" flows that just burst, causing them to go to slow-start or stall.

On the other hand, with an Anagran Flow Manager inserted into the path, the peak traffic is rate-controlled to just under the output port speed of the subsequent router so that no packets are delayed or discarded in subsequent routers and switches. The Anagran system accomplishes this by a process called **Intelligent Flow Discard™ (IFD)**, which controls the rate of each flow independently so that the peak rate is between 95% and 99% of the designated rate – the class or VLAN maximum rate for the correct port on the next router or switch.

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.

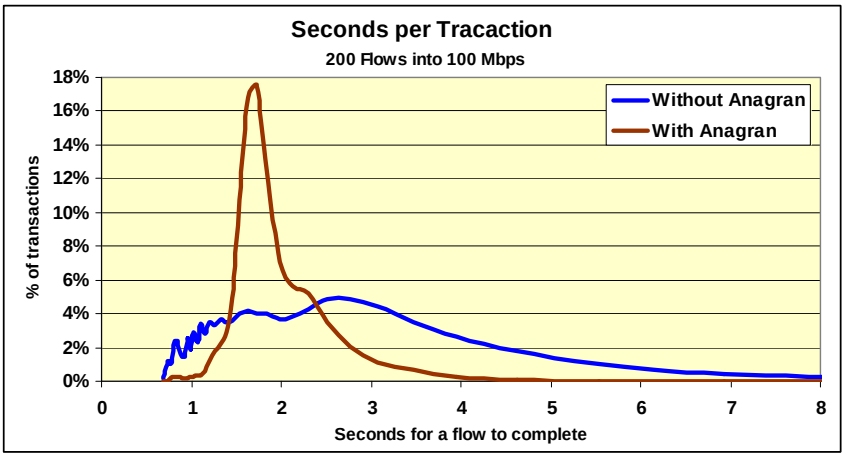


Figure 2: Histogram of Web Access Time (From lab tests)

As can be seen in Figure 2, 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.

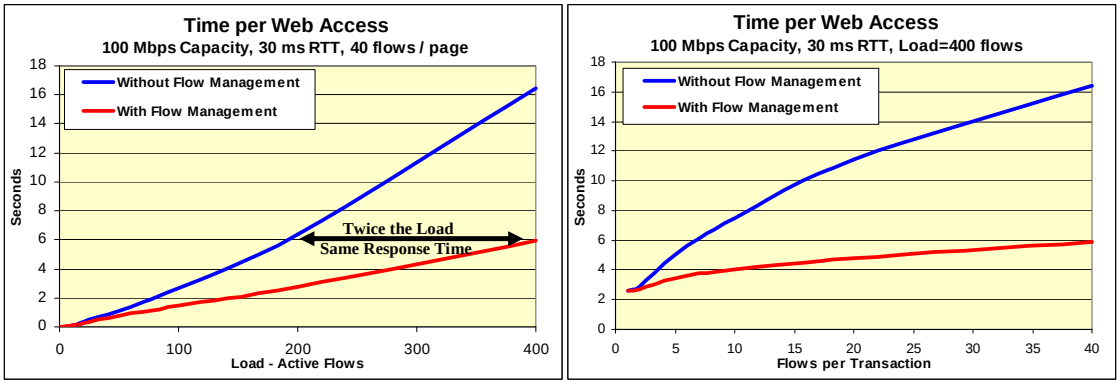


Figure 3: Laboratory test results with a 100 Mbps path with a RTT of 30 ms

The test results when the test traffic is being sent into a 100 Mbps router port is shown in Figure 3 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 3 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.***

Non-Time Critical Traffic – Bulk Transfers

While it is critical for short gaming traffic and Web transactions to be rapid and thus allowed the highest rate they can achieve, the various kinds of file transfers, FTP, long Email attachments, and P2P downloads do not require the fastest transfer rate all the time. It is valuable if the modest sized Email attachments and Web documents of the 1-10 MB size range get delivered in seconds, not minutes, but there is no need to maintain high rates every second. Then, for longer file transfers of the 100 MB or greater variety the rate can vary inversely to the load since the delivery time is in minutes or longer.

With conventional network equipment using output buffer discard technology, the rate for transferring a file increases with the length of the file due to slow-start, until it reaches the maximum feasible rate for the TCP buffers and the round trip time (RTT) involved, typically 10-20 Mbps for Windows at modest distances.

With Anagran, without any special prioritization rules, Intelligent Flow Discard (IFD) limits the rate of all flows in the same class to about the same rate, independent of the file size. See the example in Figure 4. Thus, without any special attention, the longer files are rate-constrained to fairness with the shorter flows. This ensures the short gaming and web access flows grow as fast as needed without any discards. This is already an improvement over conventional operation since short flows are promoted and large flows don't get the majority of the capacity. Then by setting **Behavioral Traffic Control™ (BTC)** rules, the priority of the larger file transfers can be decreased slowly as the flow continues.

The way this works is that rules are set that say "If the flow has transferred more than X bytes set the priority down to Y". Flows with lower priorities get proportionally lower rates. This lets the network ensure that Email and PDF files get through as fast as possible (without hurting the interactive) but the true bulk transfers like P2P get deferred until the smaller transfers do not need all the capacity – this is known as the "fill the valleys concept". The BTC rules used in Figure 4 were priority 1 for flows up to 100 KB and then used progressively lower priorities down to 0,1. The effect of any of these strategies is dependent on the momentary mix of traffic. When there is lots of traffic of all sizes, the high priority traffic takes over and the interactive uses the channel. When there is little interactive, the rates of the larger file transfers are increased. Anagran always attempts to keep the channel full. This case is shown in Figure 5 and as a result the rates of large file transfers have been increased since there is no interactive traffic at the moment.

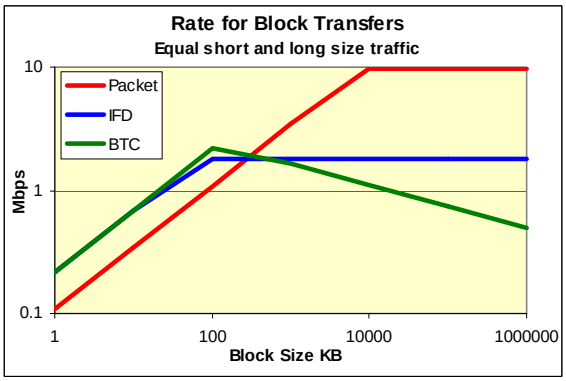


Figure 4

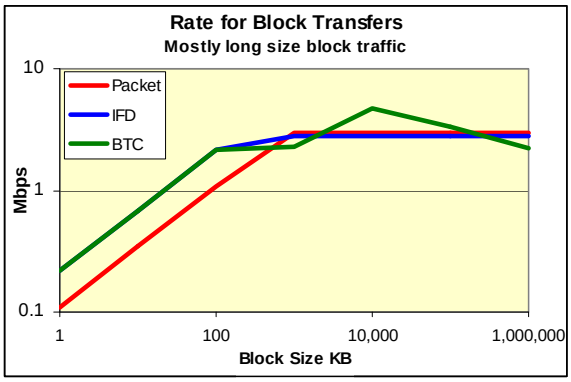


Figure 5

Examples of Bandwidth Reduction

There are many different goals that can be pursued using the flexibility of Flow Rate Management. A few will be explored below.

Reduce Channel Capacity, Same Traffic

In this example, the desire is to reduce the channel capacity required and get the same traffic across the channel. The current average utilization is 50%, but the peak often hits over 100% which causes delay and as we discussed above, it also causes the interactive rates to spread which reduces the response time for Web access. When Anagran is added into the path, using IFD and a slight BTC reduction of priority for bulk traffic, the channel capacity required can be reduced by 20%.

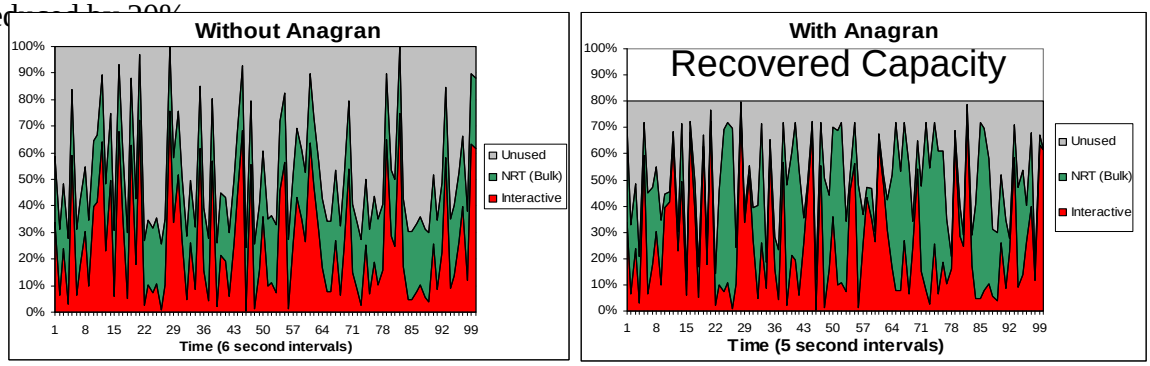


Figure 6: 20% Channel Capacity Supporting the Same Traffic

As one can see in Figure 6, the bulk traffic has been slowed down when the interactive traffic peaked and sped up during slack periods so as to "fill the valleys". Also, the Web access time has been improved.

Reduce Interactive Peaks – Same Traffic

As was shown earlier, the equalization of the flow rates for interactive traffic allows the peak rate to be reduced by 50% and still maintain the same multi-flow response time. In the following example, reducing the peak interactive rate by 25% can reduce the capacity required and also

obtain improved response time. Setting the bulk traffic to a lower priority at the same time minimizes bulk traffic during interactive peaks and "fills in the valleys" when the interactive is low. This combination can save considerable capacity while improving response time and passing all the traffic. Network bandwidth costs can be significantly reduced while end users enjoy a much improved quality of experience.

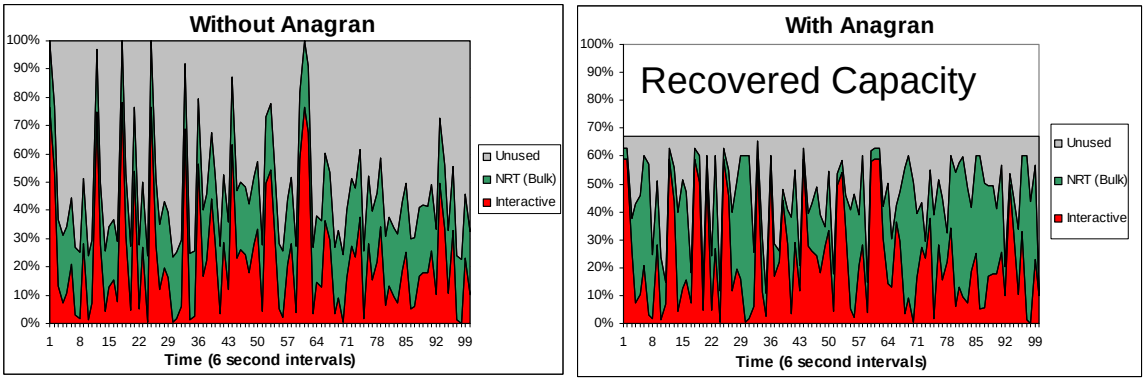


Figure 7: Reduced Interactive peaks 25% and Capacity Reduced 33%

In Figure 7 the interactive traffic averages 23% and the bulk 23% in both graphs. The Interactive peaks have been reduced from 78% to 58% whereas instead of staying at 23% all the time, the bulk has been reduced to 4% when the interactive peaks and increased to 60% when there is spare capacity. This way the total capacity has been reduced by 33% and the unused capacity reduced 33% from 54% to 21%.

Reducing the Bulk

When confronted with a lot of bulk traffic, it is sometimes necessary at peak hour to reduce the average P2P and bulk traffic, making it run slower until off-peak hours. Doing this provides even greater capacity savings.

In the example shown in Figure 8, the P2P and bulk traffic have been reduced to 50% of what was there before; down from 22% average to 11%. Also the interactive peaks have been controlled down to 55% of their prior 81% to 44%, a shift that just maintains the web response time due the equalizing of the flow rates. Of course the interactive traffic has been maintained at the same average of 25%. The result is the ability to control the traffic to fit into 50% of the prior capacity with no packet delay or random loss. The slowing of the P2P was greater than the shorter bulk traffic using the concept of tiered BTC commands to assign lower and lower priorities as the file transfers became longer. However, without the more powerful technique discussed next, some of the long Email and FTP tasks will also be slowed during this peak period.

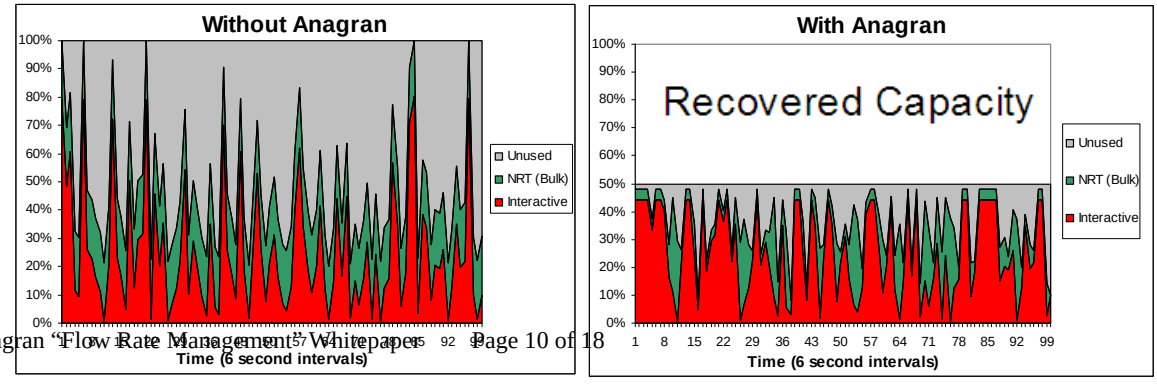


Figure 8: Maximum Capacity Reduction of 50%, Slowing P2P/Bulk to 50%

*Summary of Per-Flow Techni*ques

The examples above show how basic flow rate control can improve web response time ***and*** reduce the capacity required by simply and automatically adjusting the rate of each flow. The basic capabilities of IFD and BTC allow this powerful level of control over the total traffic load. TCP traffic usually keeps growing to the maximum rate it can, and intelligently controlling its rate rather than using random discards and unnecessary delay substantially improves network performance.

Video & Voice 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.

Fixed Rate

If the traffic is to be viewed in near real-time it should not be slowed down. IPTV 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. This causes delay variance which cannot be large for IPTV or voice; it can be tolerated by near real time video streaming through the use of a small buffer. However, loss causes artifacts since a portion of the compressed image will be lost; the higher the compression, the greater the loss.

With Flow Management, 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.

Flow Management 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 (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 should 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 Flow Management 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.

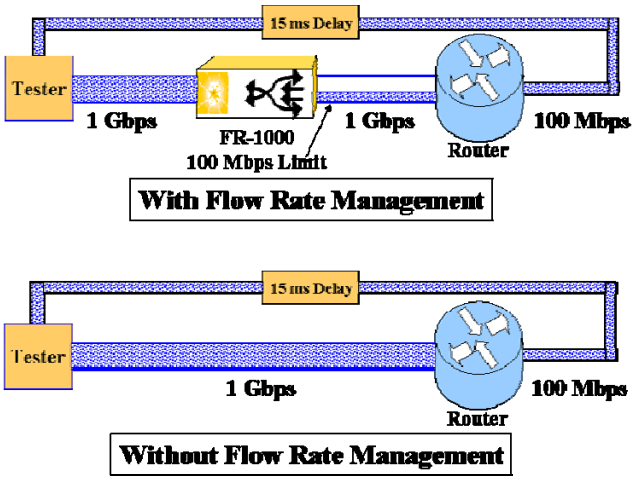


Figure 9: Test Setup for Video, Voice and Data test

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.

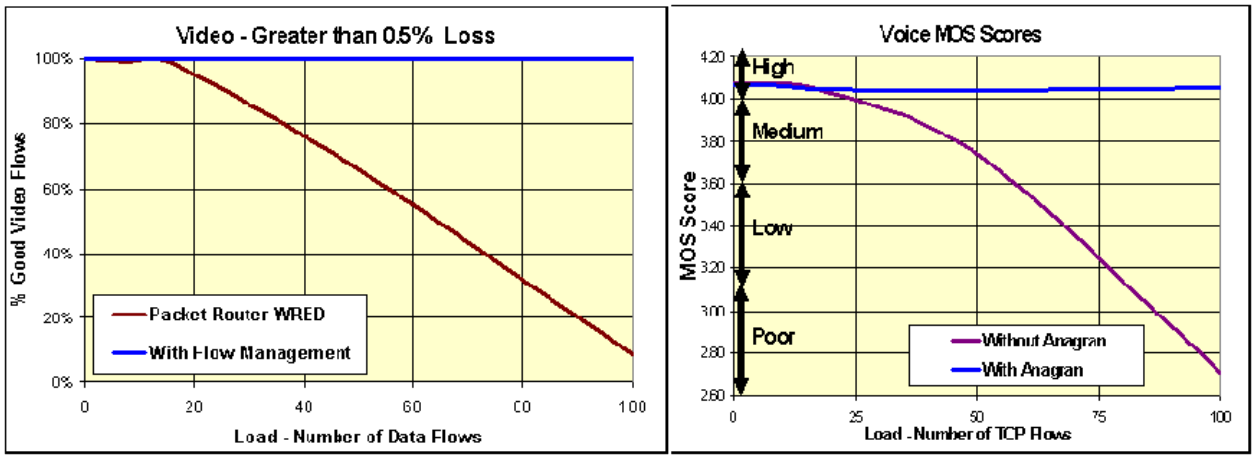


Figure 10: Test Results for Video and Voice

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.

CAC: Perhaps the most important difference is that video is becoming very popular and too many video flows on any link will cause even a good multi-flow 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) plus kill all data traffic. This is avoided totally with the Flow Management in-line since it rejects 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 often critical with the growth of video.

Using Internet Service for Corporate Video and Voice

Often, for Video conferencing and VoIP between corporate sites, special high QoS network services or leased lines are used to avoid the potential quality problems in a standard Internet service. These problems are almost always at the edge putting video or voice onto the link to the network or coming off the network into the corporation's site. Both of these problems are solved by using Flow Management at each corporate site. The same system controls the overload in both directions and thus even the ISP's potential problem of discarding or delaying traffic going onto a link is managed by the Flow Management system at the corporate site. The question remains, what about inside the ISP's network? Here, the input load is well known and is far more stable than on any edge link. Not only does the law of large numbers help but the total input capacity is known so that the core capacity can be increased in a smooth fashion to ensure a specific overcapacity even at peak hour. Video and voice will move better and without loss in large networks, so all the other traffic helps. Using this approach, the long haul capacity can be obtained for a far lower cost (most likely the lowest) and high quality maintained for the video.

Controlling P2P Traffic

In addition to the important benefits that have been outlined for rearranging flow rates to reduce the network capacity required and ensuring quality for voice and video, there are some even more powerful techniques that specifically apply to controlling P2P traffic. The first of these is in itself a complete solution to eliminate the unfairness caused by P2P programs which use multiple flows to obtain more capacity than other users. TCP allows many forms of unfairness. Simply:

1- Unfairness with distance – long delay slows TCP
2- Unfairness due to random discards – some stall, others thrive
3- Unfairness with multiple flows – the more flows, the more capacity consumed

Anagran can fix all three of these unfairness issues.

1- Anagran assigns the same rate to all similar flows, independent of distance
2- Anagran eliminates unfairness with discards since all similar flows receive the same rate
3- Multi-flow unfairness is the one that most P2P capitalizes on today. A race is ongoing between the P2P developers and the **Deep Packet Inspection (DPI)** systems to try and

outsmart each other. However, unless this basic unfairness is fixed, the end game is clear. **P2P can and will win against DPI by encrypting everything and making the flows look identical to some interactive traffic**. This is still in the future since making flows as short as Web accesses is inefficient and making them as slow as voice is also inefficient but with enough flows, they will eventually become undetectable as P2P, first with DPI and later with BTC. Therefore, a new technique is required where all the traffic from any source address or destination address is measured and used to detect P2P and to control the total rate that user is allowed. This is the next subject.

Rate Equalization per Address and P2P User Identification

What is desired to deal with the unfairness of multiple flows from one user is to accumulate a byte count and flow count for each user IP address and use that information to somehow control the traffic rate that user is allowed on a dynamic basis. The end goals are:

- Adjust each users composite usage rate based on their total bandwidth consumption to equalize the capacity provided to all active users
- To detect and classify P2P traffic by watching the number of flows per user, the total traffic per user, and the size of the flows

Rate Equalization

Although this is a simple concept, implementation is much more difficult. It is certainty possible to keep in memory the byte and flow count for 100,000 users which might be using a 10 Gbps link. However, the addresses must be looked up and bytes counted - perhaps every packet time. Then there needs to be a smooth and effective way to slow traffic for that user, hopefully one which does not stall or kill the flow. However, this is quite easy in the context of a flow manager where most of these capabilities already exist.

Anagran's Flow manager achieves its compact size and power efficiency by streaming packets through and working mainly at the next higher level, flows. The P2P problem with multiple flows is primarily observable at the next higher level, the traffic and flows to or from one user, typically identified by one IP address. Since byte counts for every flow already are kept, it becomes easy to accumulate the byte count per address as well. These are then averaged over a suitable period like 15 minutes (configurable) and used to adjust a priority for that user. The higher the 15 minute traffic rate, the lower the priority. This priority is then combined with the other priorities and used to control all the flow rates to that user. The result is that all users on one common resource like a cable, DSLAM, Internet access link, or wireless base station receive the same average capacity.

Of course, the basic policy of equality can be modified by additional policies and priorities such as the payment level and other factors. Also, since there needs to be an averaging period to account for normal traffic bursts, the effect will be slightly delayed. However, this does not create a loophole; a user cannot get around the impact if they want to download a video. Thus, this new technique, "Rate Equalization per Address", provides a major new capability to control P2P and address the inherent unfairness issue. Given the attractiveness of any unfairness, it is likely that this TCP unfairness will be more widely used in the future and it is important to

address it now. At least then in the future, software developers will optimize their systems' compatibly with the communication system and not capitalize on an unfairness that costs someone else money.

P2P User Identification

Additionally, it is important to accurately detect and classify P2P traffic, so that the P2P flows can specifically be rate controlled as distinct form FTP or Email. Anagran has found in testing these techniques in an office environment, that P2P users can be identified with high accuracy if during any four second period the user has more than 5 flows and more than 50 Kbps of traffic. P2P always triggers this rule whereas the other main case of multiple flows, Web Access, does not since the traffic rate is lower due to slow-start. Of course the rule can be adjusted based on the environment.

Once the probable P2P user is identified, the actual P2P flows can be selected based on additional BTC rules. These rules are ones like: the flow has transferred more than 3 MB and has an average packet size over 1000 bytes. If it was not for the fact we had already determined the user has multiple flows operating, this rule might have selected Email and FTP also but with the multi-flow rule up front, this rule picks out only the P2P transfers.

These features are currently being tested in the Anagran system. It is designed to support all the users or stations that would be expected to be supported by an Anagran FR-1000 in a local service area or campus environment.

Comparison with Deep Packet Inspection (DPI)

DPI has recently been employed to try and control P2P traffic. However, this approach is failing due to several factors:

- DPI is unable to recognize encrypted applications or keep up with new signatures as application fight back. Controlling only 60-70% of the P2P allows the other 30% to fill the pipe. This is a no-win race
- DPI has difficulty controlling the rate of flows, even after detecting them. Sending resets kills legitimate applications and discarding ACK's often stalls or kills the flow. Slowing a flow precisely and smoothly takes a much different approach.
- Digging into each packet is not only invasive but very expensive.

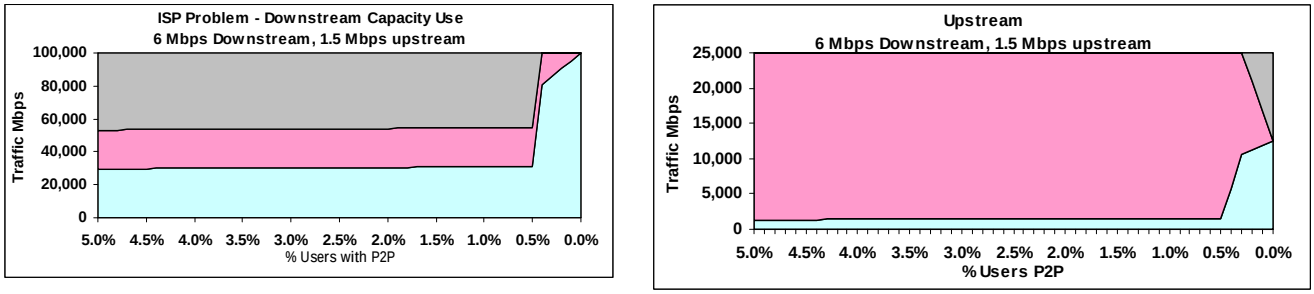


Figure 11: Impact of P2P on Asymmetric DSL Service

The example in Figure 11 is of a ISP using DSLAMs, each of which supports 8000 subscribers on a 100 Mbps trunk. If over 0.5% of the subscribers are using P2P which DPI cannot detect, then the P2P users congest the lower capacity upstream. The congested upstream slows the acknowledgements required for the downstream users. This limits the downstream to about 50% utilization with the P2P users consuming 23.5 % and the normal users getting 30.6%of the trunk capacity.
When instead, the capacity allowed to each subscriber is equalized with flow management, all the users become the same as normal users and each receives
The concept of DPI leads to an arms race between the DPI vendor and the application engineer. This is expensive and will always fail some of the time. Since P2P expands to use all available capacity, any failure leads to congestion of the limited capacity by the remaining P2P. Also, legitimate P2P applications will find it hard to get even their fair share of capacity.

Example of P2P Impact on DSL even using DPI

The following example is of a ISP using DSLAMs which support

Comparison with Hard Rate Limits per User

Another technique that has been tried is to set a rate limit per user and police all traffic if this is exceeded. This clearly is not dynamically responsive to changing conditions where under light loads the rate limit restricts the total Internet utilization and under heavy loads (many users) the P2P users receive an unfairly large capacity. Further, the policing may seriously harm video or voice and limit short interactive bursts. This also has been expensive, both to buy and to operate.

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 Discard (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. When P2P is present these functions allow the P2P

traffic to be placed at a lower priority so that it does not unfairly consume capacity. To control or monitor P2P the following functions permit a one-time total solution to P2P unfairness:

- **Rate Equalization per Address** allows traffic to be distributed equally to all users (addresses), independent of the number or type of flows they are using, ***eliminating the unfairness of P2P*.** This totally eliminates the need for DPI in controlling P2P.
- **P2P User Identification** allows the P2P users to be very cleanly identified through the total user activity and put into a P2P class. Then their P2P activity can be identified by flow and managed by priority or management policy.