33 Bandwidth Management (Edge)

If you have not read the book (*Performance Assurance for IT Systems*) check the introduction to "More Tasters" on the web site http://www.b.king.dsl.pipex.com/ to understand the scope and objectives of tasters.

The primary focus of this taster is on the use of devices that sit on the periphery of a network with the purpose of getting the best throughput / performance out of the available bandwidth by various priority and queueing techniques. It commences with a short review of methods by which the load on the network can be reduced; concentrates on Edge Bandwidth Managers; and it includes a brief overview of network-centric approaches to QoS and Traffic Engineering, primarily as a means of providing some context to Edge Bandwidth Managers.

33.1 Starting Terminology

ATM (Asynchronous Transmission Mode) evolved from Frame Relay. It uses small, fixed 53 byte cells that reduced the processing overhead and thereby the latency.

Bandwidth Manager is the name given to software that resides in an "Edge Device", which controls and shapes network traffic

DiffServ stands for differentiated services, i.e. a mechanism for classifying different types of traffic which are each handled differently by the network to provide the required quality of service.

Edge Devices, as the name implies, sit on the periphery of the network, usually adjacent to a router.

MPLS (Multi Protocol Label Switch) is an IETF (Internet Engineering Task Force) initiative that generates information (the "label") which is written to packets (sitting between layer 2 and layer 3) and is used to expedite their forwarding across the network.

QoS (Quality of Service) is the general term that is used among the networking fraternity to describe mechanisms that attempt to meet service level requirements for network performance.

Traffic Engineering is a term that is used for techniques that optimise the performance of a network. On its own, it is concerned with the overall picture, not with individual types of traffic.

33.2 Technology Outline

33.2.1 Reducing Demands on the Network

When addressing network performance the first step should be to ensure that the resources are used sensibly; wanton and indiscriminate use should be avoided wherever possible. A summary of some of the techniques that can be employed to reduce network traffic follows. See other tasters for more detail, particularly *Network Basics* and *Web Server and Cache Server*:

- **Caching.** There are several forms of caching. Web cache servers that are strategically placed around network can satisfy read requests (mainly for static content), thus avoiding longer network round-trips to central web and FTP servers. This type of caching can also be done at the client-end in the first instance, using standard Internet Browser caching. Bespoke client-side application caching facilities can be particularly fruitful, albeit possibly expensive to develop for non-static content

- **Compression.** There are a number of techniques in this area. Some protocols can compress headers on-the-fly. Examples include PPP (Point to Point Protocol) which is used on slow dial-up links and RTP (Real Time Protocol) which is used for multi media conferencing. Data can be pre-compressed on the server before it is served up; a number of web server products allow static content to be served up that has previously been compressed when it was originally placed in the content store. The degree of data compression that can be achieved will depend on the algorithm that is used. In the area of images, “lossy” algorithms can achieve significant compression ratios at the expense of some loss of definition (20:1 or greater can be viable); whereas lossless algorithms manage lower compression ratios (frequently 2:1 or less) but retain full definition. It obviously depends on the requirement but lossy compression can be adequate in many circumstances. Some protocols, particularly where large amounts of bandwidth can be consumed, provide compression; RDP (Remote Desktop Protocol), which is used in thin client technology products, is an obvious example

- **Content Filtering.** At a system-wide level this can include the discarding of any data that is perceived as being of no use. Obvious examples are obscene material and spam. At an application level the objective should be to limit transmission to essential data. A simple example here is filtering the results of a query centrally, probably by a database server, rather than sending a superset of the data that the client-side must subsequently filter

- **Limiting Access.** Access lists can be used to regulate which users can obtain access to data at all, or possibly they can be used to allow some users to see summary information but not the detail

- **Protocol Accelerators.** TCP throughput can be an issue over slow links, e.g. satellite, due to long round-trip times. This can constrain the use of the available bandwidth. XTP is an IP-based, connection-oriented protocol that can be used to improve performance over such links. It employs more intelligent retransmission algorithms than standard TCP and generally reduces traffic overheads. It can be implemented in the form of gateway devices to the problem part of the network. Some of the features in XTP are finding their way into later versions of TCP, e.g. SACK and T/TCP.

33.2.2 Bandwidth Management Edge Devices

The main focus of this taster is the use of Bandwidth Management Devices (BMD) to control the flow of network traffic, primarily though not exclusively for outbound traffic that is about to go over the network. These devices sit on the periphery of the network, e.g. in a data centre or in a building on the client-side, typically adjacent to an edge router. Solutions in this space are either software-only or they come in the form of a black box (hardware and software).

The basic requirements of a Bandwidth Manager are to:

- **limit the overall bandwidth usage.** An example of this requirement is where the WAN link that the adjacent router is connected to has limited bandwidth. In this situation the router is likely to become overloaded, and it will start to discard packets. Constraining the bandwidth between the BMD and the router to the speed of the WAN link will prevent this happening

- **divide traffic into classes**. Different types of traffic may require different amounts of bandwidth and different priorities. For example, there may be three classes: Voice over IP (VoIP), FTP, and HTTP where VoIP is given the highest priority and is allocated the majority of the bandwidth

- **subdivide a class into separate "per-flow queues".** Simple class granularity may be too coarse, e.g. one session may be getting more bandwidth than another within the FTP class. This mechanism can be used to provide a guaranteed amount of bandwidth to individual sessions

- **flexible and dynamic use of bandwidth.** If a class (or sub-class) is not using its entitlement at a particular point in time it should be possible to redistribute any unused bandwidth to other classes or sessions

- **congestion control.** It should be possible to make use of the self-healing features in a protocol, e.g. TCP, to discard packets when congestion occurs within the Bandwidth Manager itself. The objective is that the sender will retransmit the discarded packets at a reduced rate until the congestion abates. Ideally, any reduction and subsequent increase in the packet transmission rate should be graceful, as opposed to bursty.

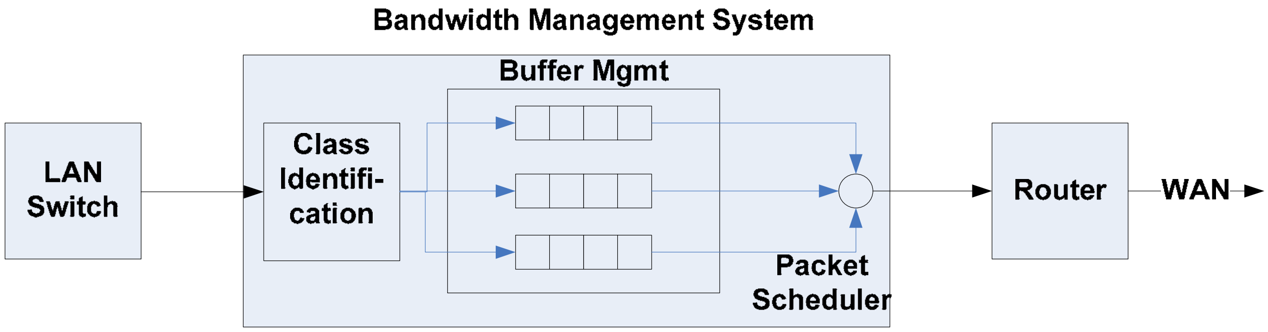

Figure 33-1 Structure of Bandwidth Manager

As shown in Figure 33-1 there are three main elements to a Bandwidth Manager:

- Identifying packets, and placing them into nominated queues, each of which has a specified bandwidth requirement

- Buffer Management handles the various queues. Each queue corresponds to a single class (CBQ) or to a single sub-class where control is required at the "per-flow" or session level (PFQ)

- And the Packet Scheduler, which uses the bandwidth requirement of each queue to control how packets are dispatched to the router, and hence out to the network.

33.2.3 Network-Centric Approach to QoS

The objective of this section is simply to provide some context to the techniques that are used in Edge BMDs. There is no attempt to compare the two approaches. Network-based QoS is a complex topic that warrants a taster of its own. For the moment I will limit discussion to a brief overview, making reference to MPLS (Multi Protocol Label Switch), a relatively recent technology, which is currently "flavour of the month".

One of the initial objectives in Quality of Service (QoS) for IP networks was to cater for a "per flow" level of service, which was termed IntServ (Integrated Services). However, this approach was overly complex and more importantly not scalable, as RSVP, one of the underlying network system protocols, had to distribute "per flow" QoS guarantee requests across the network and state information for each flow had to be maintained at each hop across the network. A more pragmatic approach resulted in DiffServ (Differentiated Services), which essentially equates to the simple class split that Bandwidth Managers offer. The class is marked on the packet itself, called the DSCP (DiffServ Code Point). QoS behaviour of a packet is decided at each node in the network, termed the "Per Hop Behaviour" (PHB). The types of PHBs are: BE (Best Effort), i.e. no special treatment; EF (Expedited Forwarding) for high priority traffic; and a number of AFs (assured forwarding).

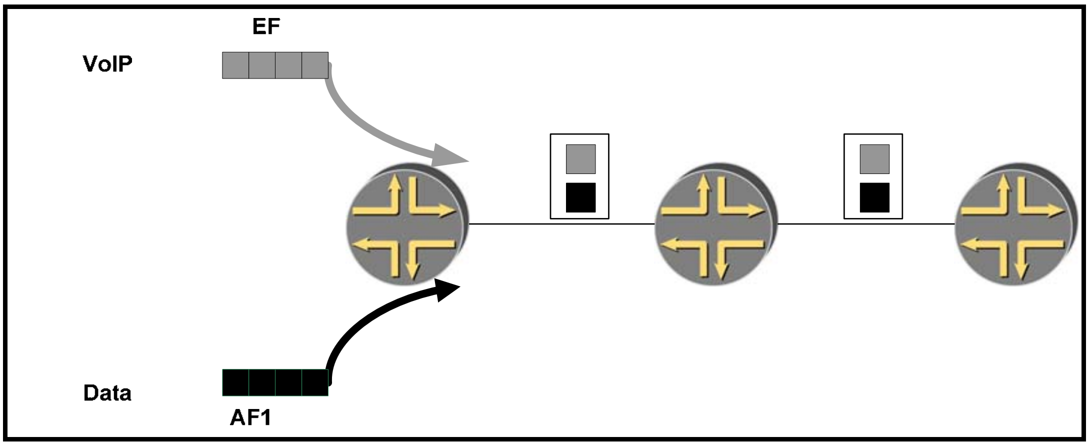

Figure 33-2 DiffServ Queues

As shown in Figure 33-2, in a DiffServ domain the class identification is marked on the packet at the entry point and is used at each hop across the domain. VoIP is the higher priority EF while data is lumped in the lower AF1. Although DiffServ provides a class-based priority system, it cannot guarantee Quality of Service if the network path has insufficient resources. Traffic Engineering features are used for this purpose.

The basic concept behind MPLS is that a label is written on each packet at the ingress node of an MPLS network domain. This label is used for DiffServ and for routing the packet across the domain. The main advantages of MPLS are:

- The processing that is required to generate the label contents is only necessary at the ingress node, in contrast with IP networks where routing processing is necessary at each step across the network. This means that ingress nodes need to be more powerful than other routers

- The chosen path across the domain, termed the label-switch path (LSP), can be based on overall QoS considerations rather than on the more rudimentary shortest path objective of IP networks.

Factors that can be used to influence the decisions on the choice of path include: bandwidth requirements, the types of link that can be crossed, e.g. low latency links cannot be used, the permitted number of hops, and a path priority (higher priority LSPs can pre-empt lower priority LSPs). To compute a path each router must advertise its status and resource availability using link state protocols such as IS-IS and OSPF. Once decided, the path must be reserved; this is done via RSVP. Figure 33-3 shows a simple split by priority and bandwidth requirement. All links are 150Mbps. Higher priority traffic requires 100Mbps and it is allocated the shortest LSP path A-C-D-G. The lower priority traffic also requires 100Mbps but it is given the longer LSP path B-C-E-F-G.

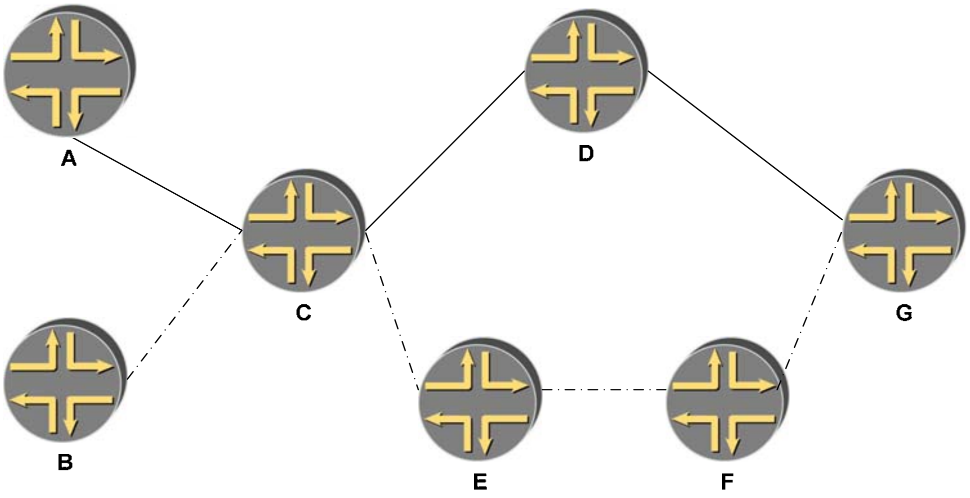

Figure 33-3 Bandwidth Allocation

One of the selling points of MPLS is that the provision of paths that satisfy QoS requirements means that it is not necessary to provide the excess capacity that IP networks may require to provide the same level of service. Of course, it is all well and good to plan Quality of Service; it is a totally different matter to deliver it. It is necessary to monitor traffic to ensure that it stays within the reservation bounds that have been set. One method is to use “LSP Policers”, installed at the ingress node. Misbehaving traffic can be dropped or marked (for special attention).

33.3 Performance

The performance objectives criteria of Bandwidth Managers include:

Performance / Accuracy. The performance of the Bandwidth Manager must ensure that it meets the primary objective of allocating the bandwidth according to the specified rules. Performance is more likely to an issue where there are many queues (when using PFQ), where significant delving is required at layer 7 to classify traffic correctly, and where any lack of memory may lead to frequent congestion with the result that a noticeable percentage of packets are discarded.

Fairness is the equality of bandwidth provision within a class. Obviously, simple class division (CBQ) will not provide the necessary functionality to meet this objective. Where PFQ is used, problems of fairness are likely to be more evident where the class has limited bandwidth available for distribution.

Handling Congestion. There are two techniques for handling packet congestion. Random Early Detection (RED) uses a probabilistic function of the average length of each queue to decide if an incoming packet should be dropped. It is best used with a protocol such as TCP where the sender will reduce its transmission rate in the light of packets being dropped. RED can result in bursty traffic, particularly in limited bandwidth situations, i.e. it can be somewhat heavy-handed, as it results in the halving of the window size by the sender (the amount of data that is allowed to be outstanding in the network before an acknowledgement (ack) is required). The second technique is TCP Rate Control (TCR), which is arguably a more refined approach. Here, the behaviour of the sender can be controlled by using window sizing and “ack pacing”, a method that controls the rate at which acknowledgements are sent back to the source and hence the rate at which the source can send packets. The use of TCR may be problematic if the source is over the other side of the WAN or Internet and there are delays within the network itself that results in dropped packets, as this may affect what the BMD calculates to be a reasonable (just in time) pacing. Whatever technique(s) are used, a key objective should be to ensure that retransmission rates are low.

Avoiding jitter in time critical workloads, e.g. Voice over IP and video. This can be caused by a high priority packet getting to the packet scheduler only to discover that a large (say) FTP packet is currently being dispatched. The lower priority packet cannot be pre-empted once it is in transit. A method of minimising this problem is to use MSS-Clamping. MSS stands for maximum segment size. In essence, the Bandwidth Manager can interrupt the initial hand-shaking between the source and destination when the size of a packet is agreed upon, and it can impose a smaller size which will help to reduce any jitter.

Borrowing Bandwidth. It should be possible to make full use of any unused bandwidth in another class (inter-class borrowing) or within a class (intra-class borrowing). The software implementation of some products may limit the amount that can be borrowed, i.e. they cannot borrow all of the unused bandwidth that may become available.

Robustness. It is important that the software behaves well in adverse conditions such as the loss of packets over long distance sessions.

Resilience. In a highly available system / network the Bandwidth Manager should also be capable of being fully redundant by the use of two devices that are directly connected to each other. The basic mechanism of having an active and a totally passive standby device with the use of heartbeat checking to detect failure is unlikely to be satisfactory from a performance perspective, as the failure of the primary device may well introduce a noticeable delay while the standby takes over, which may well cause stability problems. A better approach from the perspective of providing a seamless failover is where traffic is copied to the standby node which processes the traffic, just as the active device does, but it does not forward the traffic while it is acting as the standby. A further refinement is where both devices are active (which obviously demands some form of load balancing), and while they each process all the traffic (as above), they only forward their own traffic.

33.4 General Observations

The main observations on Bandwidth Managers are:

- There is a danger that the facilities are considered to be a way of "getting a quart out of a pint pot", i.e. that they will provide a permanent solution to an under-sized network. In reality, the techniques allow for the smoothing out short-term performance issues, plus fair and optimum use of the network; they are not a panacea for a basic lack of bandwidth

- Iterative tuning may well be required to realise optimum performance for the demands that your particular systems place on the network

- There appears to be little in the way of published benchmark results for this type of device

- Beware what comes in a black box appliance, i.e. what dictates the model size. There is a tendency for vendors to equate a given model in their range with a particular bandwidth requirement. However, requirements for fine granularity of queues, e.g. for per flow queueing (PFQ) and detailed SLA reporting, may necessitate a larger model which has sufficient memory to accommodate the required buffer space

- Some products provide data compression. However, the use of this feature means that the software must also sit on the destination side of the network to uncompress the packets.

- The reporting facilities that are offered with this type of product may be required to monitor actual performance against the objectives. However, reporting, although comprehensive in some products, can make for a significantly more expensive solution.

33.5 Further Reading

Wei, H-Y., Lin, Y-D., *A Survey and Measurement-based Comparison of Bandwidth Management Techniques*, IEEE Communications Surveys & Tutorials – 4th Quarter 2003, Volume 5, No. 2. This is essential reading, comparing products and providing a summary of benchmark results.

Wei, H-Y., Tsao, S-C., Lin, Y-D., *Assessing and Improving TCP Rate Shaping Over Edge Gateways*, IEEE Transactions On Computers, Volume 53, No. 3, March 2004. This is a detailed paper.

Knight, J.P., *Review of Bandwidth Management Technologies,* Loughborough Computing Services / JANET Bandwidth Management Advisory Service, December 2003 is a very useful document.

Minei, I., MPLS DiffServ-aware Traffic Engineering, http://www.juniper.net provides a useful introduction to the practical use of MPLS.

Web sites of note include:

http://www.bmas.ja.net is the web site of the Bandwidth Management Advisory Service for JANET. It contains much useful information.

http://www.icir.org contains references to papers on CBQ, RED, TCR, *et al*.

Companies that supply XTP products include Mentat (http://www.mentat.com) and Network Xpress.