

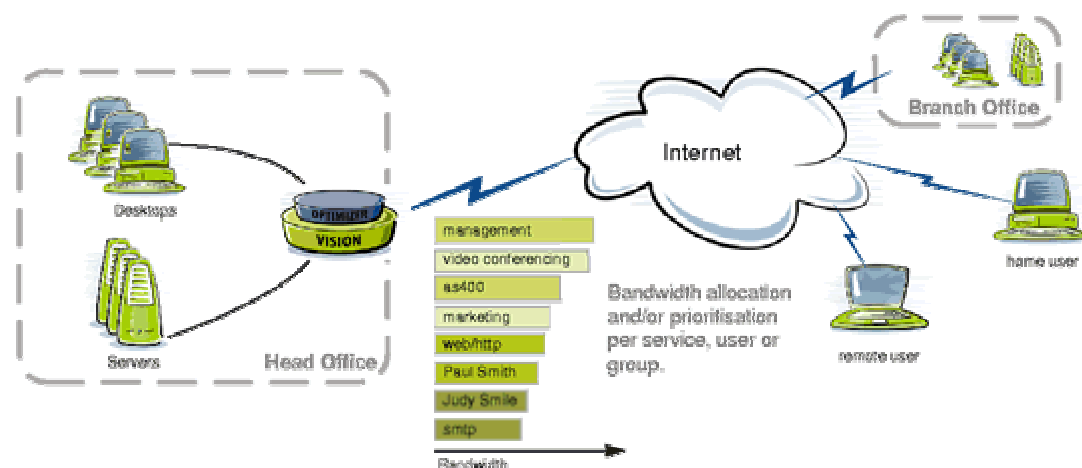
White Paper Traffic Optimisation

Traffic optimisation is the general term given to a broad range of techniques designed to enforce prioritisation policies on the transmission of data over a network link.

Most of us will at some time or another have experienced the effects of network latency and queuing. A common experience is the sometimes frustrating delay before characters are echoed when Telnet-connected to a remote host, for example –to the AS400 accounting package. Any of you who have been forced to use a low speed dial-up PPP or SLIP connection will have seen the effect that a file transfer has on interactive traffic over the connection. The file transfer easily consumes most of the link bandwidth, forcing the Telnet data to be queued, waiting for a free slot before being transmitted across the link. When a large file transfer takes place over a low speed connection the user experience for web browsing becomes unacceptable.

When optimising, the focus is on user experiences rather than kilobits per second.

Proactively create a specific behaviour on your link rather than allowing traffic to perform in an unpredictable nature.



This problem occurs because the datagrams containing the file transfer data are given equal priority on the link to the Telnet or real-time datagrams. No consideration is given to the type of data contained within the datagram when queuing it for transmission; queuing is performed on a "First In, First Out" (FIFO) basis, and the datagrams are scheduled for transmission on a "First Come, First Served" basis. When a new datagram arrives at the queue, it is added to the tail of the queue; when the link bandwidth becomes available, the datagram at the head of the queue is transmitted.

Traffic shaping allows us to implement a specific policy that alters the way in which data is queued for transmission. Datagrams associated with file transfers are generally quite large, often MTU sized, while datagrams associated with interactive sessions like ssh or Telnet are often quite small.

All data takes time to transmit over a network connection; the larger the datagram, the longer it takes. As a result, if you have to wait for a large datagram associated with a file transfer to be transmitted before your Telnet keystroke can be transmitted, you will perceive considerable delay.

Additionally, because file transfers don't need to wait for human input to continue, you can be sure that at any time you hit a key in your Telnet session, there are already not one, but a number of datagrams from the file transfer session sitting in the queue and further compounding the delay. Ultimately, it will take the same amount of time to transmit all of a set of datagrams across a network link, no matter what order they are transmitted in, so at some point you may decide that it is worth trading off a small amount of delay in completing the file transfer for more normal response times for your interactive sessions like ssh or Telnet by giving them higher priority.

The Traffic Shaping engine acts as a powerful bandwidth manager, transparently regulating network traffic. You can identify, group, and apply a data rate to any desired traffic. The traffic shaper offers flexible data rates and accommodates burst traffic, while smoothing out the peaks. Connection oriented data rates implement bandwidth policies and quotas for workstations, applications, hosts, TCP connections.

Traffic Management requires 4 revolving stages:

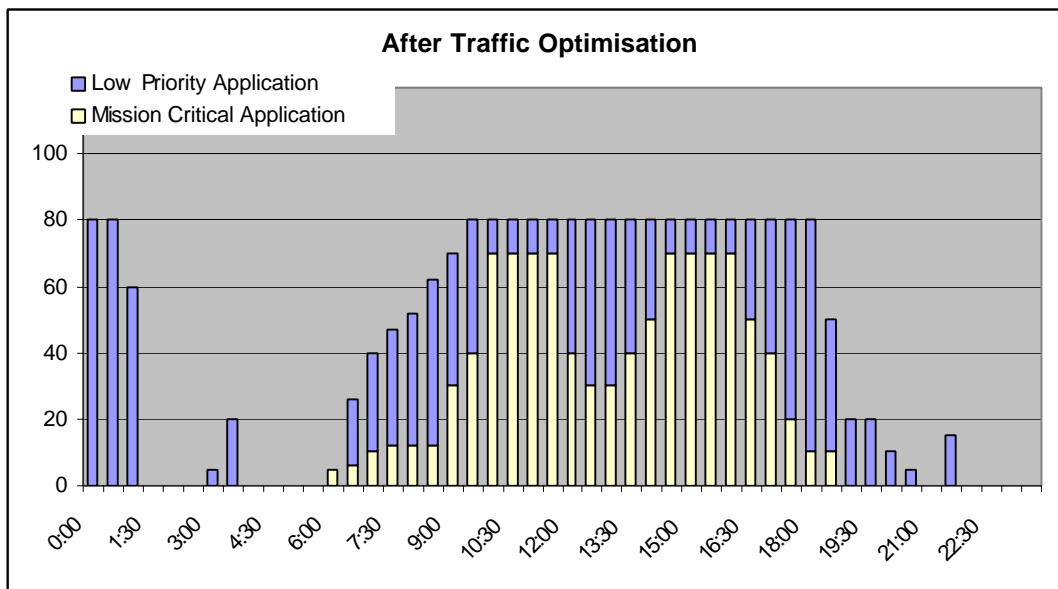
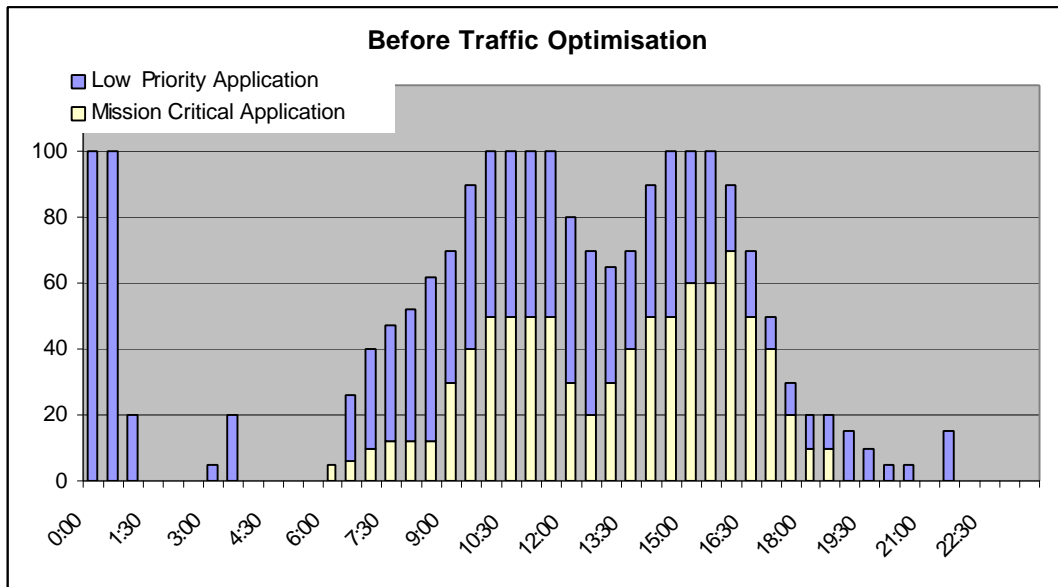
Plan
Provision
Monitor
Optimise

How does it work?

When data is received by the optimiser it is:

1. **Classified:** using access lists (filters) the data is classified. Typical classification groups may be named "High Priority", "Real-time" or "Low priority"
2. **Labelled:** Once the data has been identified it is labelled with the appropriate priority number.
3. **Queued:** The packets are placed into the corresponding queues. The high priority queues are serviced first.
4. **Processed:** The scheduling mechanism checks the queues in a round robin fashion and empties the high priority queues first. Therefore a high priority packet which arrives 'after' a low priority packet is processed first. This way FILO and FIFO queuing techniques are manipulated to provide true delivery of Quality of Service (QoS).

Instead of competing for bandwidth, the Mission Critical Application is now utilising as much bandwidth as required and network response times have been reduced. Low priority applications like email and backups have not been effected (from a user point of view) when spread over a larger time slot. This efficiency can be obtained by understanding the nature of applications in real-time networks.



The link is now operating at 80% utilisation, meaning QoS has been met and more users can be added without an upgrade on the network. This also allows for unusual usage of the system (eg additional enquiries following a promotion). Most organisations are likely to find that their Internet bill is reduced due the payment plan offered by their ISP (eg metered or burstable pricing).

This approach has been used by carries for many years. By traffic shaping or optimising a network it has been proven that costs can be reduced whilst improving the user experience.