QoS (Quality of Service)

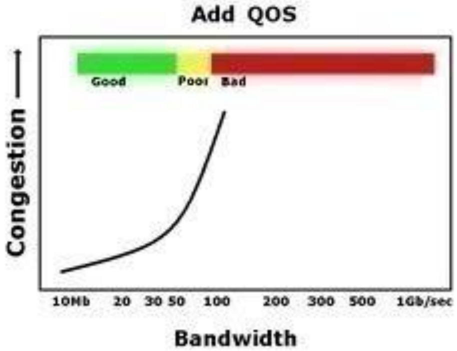
QoS stands for Quality of Service. QoS is a generic name for a set of algorithms which attempt to provide different levels of quality to different types of network traffic.

Queuing

One method of implementing QoS is to utilize some sort of advanced queuing algorithm.

Simple networks process traffic with a FIFO (First In – First Out) queue. Packets which are received first are processed first.

QoS comes into play when the network admin wants to treat some packets differently than others. For example: e-mail packets can be delayed for several minutes with no one noticing, while VoIP packets cannot be delayed for more than a tenth of a second before users notice a problem.
In a true FIFO system, all packets are stored in one queue. An advancement over FIFO queuing is Fair Queuing (FQ). In an FQ system, each type of packet is stored in a it’s own queue. FQ isn’t terribly useful, but a variation of it is. That variation is Weighted Fair Queuing (WFQ). In a WFQ system, each queue can be given a different priority level. That is where QoS really begins.



Class Based Weighted Fair Queueing

Improvements to WFQ include Class Based Weighted Fair Queuing (CB-WFQ), where each type of traffic is assigned to a class and each class is given it’s own queue. CB-WFQ allows for easier queue management.

Heirarchical Weighted Fair Queueing

Hierarchical Weighted Fair Queuing (HWFQ) is another improvement to simple WFQ. In HWFQ, the network device monitors the worst-case packet delay for each queue and adjusts to queue priorities automatically.

Random Early Detection

Random Early Detection (RED) is a an algorithm which simply drops packets if too many are being received. This causes the devices which are sending the packets to notice a problem and reduce their transmissions.

Weighted Random Early Detection

An improvement to RED is Weighted Random Early Detection (WRED). WRED is RED which utilizes the IP headers priority value to determine which packets to drop.

Traffic Shaping and Rate Limiting

Another method of implementing QoS is traffic shaping. In traffic shaping, traffic from each source is monitored for bandwidth utilization. When traffic from a specific source is too high, packets from that source are then queued (delayed).

Rate Limiting

Rate Limiting is an improvement on traffic shaping. With Rate Limiting, the packets are not simply queued, the packets can also have their IP priority levels altered or they can be dropped altogether.

Source: http://www.tech-faq.com/qos.html