



Techniques for Improving Bandwidth, Latency, and Loss Across the WAN

WAN Acceleration Overview

WAN acceleration is increasingly becoming more important to enterprises as a key enabler of strategic IT initiatives, including branch office server and storage centralization and disaster recovery. Recent advances in technology have enabled WAN acceleration to transition from a tactical fix to a strategic IT investment. However, these advances have also led to increased confusion in the marketplace. Never before have there been more WAN acceleration products, leveraging a wider variety of technologies, with varying levels of results.

There are a variety of WAN acceleration techniques that can improve application performance across distributed offices. Some focus on maximizing bandwidth utilization, others address latency (the time delay between when something is sent and when it is received), and still others address network integrity issues that can prevent the effective delivery of packets across the WAN.

The goal of WAN acceleration is to address **all** of these issues — bandwidth, latency, and packet loss. Following are the most common techniques that address these challenges, resulting in maximum application performance across the WAN.

WAN Deduplication

The most efficient way to accelerate the transfer of information across the WAN is not to send it in the first place. This is the major principle employed by WAN deduplication (also known as data reduction). A WAN acceleration appliance equipped with this technology will examine all data in real-time prior to it being sent across the WAN, and store this information in local data stores. Whenever duplicate information is detected, instructions are sent to the appropriate appliance to deliver the information locally instead of resending it across the WAN.

Using this “network memory” technique can eliminate over 90% of WAN traffic under the right circumstances. It provides various levels of improvement based on the application environment and the repetitiveness of the traffic. For example, interactive web traffic may see a 10-fold performance improvement, while large data backups may see a 100x improvement. Performance naturally increases in environments with lots of duplicate information and after data reduction appliances have had an opportunity to “memorize” the network.



Compression

Compression is used to reduce the bandwidth consumed by the traffic travelling across the WAN. **Payload compression** uses algorithms to identify relatively short sequences that are repeated frequently over time. These sequences are then replaced with shorter segments of code to reduce the size of transmitted data. **Header compression** can provide additional bandwidth gains by reducing the information using further specialized algorithms.

The gains realized vary depending on the mix of traffic on the WAN but are fairly consistent across different solutions. Text can yield between 2–5x compression ratios. On the other hand, precompressed content like zip files will generally yield no gains. Typically, enterprises will get 1–2x more bandwidth deploying compression techniques on the WAN.

Latency Mitigation

The time it takes for information to go from a sender to receiver and back is called network latency. Since the speed of light is constant, a minimum latency is directly proportional to the distance travelled between the two network endpoints. For example, it is common to see >100 milliseconds latency when communicating between the United States and Asia, or even between both American coasts.

However, latency is also impacted by queuing and processing in routers and other network elements along the path. Many file, email, and document management systems leverage the Transport Control Protocol (TCP), which has a variety of congestion control functions that can actually introduce quite a bit of latency. WAN acceleration devices often leverage a variety of TCP acceleration techniques to overcome this, including selective acknowledgements and adjustable window sizing. There are also “chatty” protocols such as Microsoft’s Common Internet File System (CIFS) which perform poorly in the face of latency. Applications like this require hundreds or even thousands of round trips to successfully transfer a single file. This is typically not an issue when file servers are deployed on the same Local Area Network (LAN) as clients. However, when CIFS is used across a WAN, as is the case when branch offices are accessing file servers located within a centralized data center, both latency and bandwidth constraints can adversely impact file sharing performance. To overcome this, different approaches have been adopted, including **read-aheads** and **write-behinds**, whereby requests are pipelined on behalf of the client to eliminate round-trip delays.

Network Integrity (that is, loss mitigation)

Even when the physical layer of a WAN is error-free, some technologies and provisioning practices still lead to packet loss at the network layer. In fact, it is possible to see network packet loss rates as high as 8% in some MPLS and IPVPN networks. When this type of loss is coupled with high latency and the retransmission and congestion-avoidance behavior inherent to TCP, it is not surprising that application performance suffers across a WAN. When doing replication and other applications that require the transfer of a lot of continuous data, as little as 0.5% packet loss can cause effective throughput to drop to <1 Mbps — regardless of how much WAN bandwidth is actually available.

Forward Error Correction (FEC) is a technology that is well known for its ability to correct bit errors at the physical layer. This technology is often adapted to operate on packets at the network layer to improve application performance across WANs that have high-loss characteristics. FEC works by adding an additional error recovery packet for every “N” packets that are sent across the WAN. This FEC packet contains information that can be used to reconstruct any single packet within the group of N.



If one of these N packets happens to be lost during transfer across the WAN, the FEC packet is used on the far end of the WAN link to reconstitute the lost packet in real-time. This eliminates the need to retransmit the lost packet across the WAN, which dramatically reduces application response time and improves WAN efficiency. An advanced implementation will dynamically adjust FEC overhead in response to changing link conditions for maximum effectiveness in environments with high packet loss.

Packet Order Correction (POC) is another useful technique to overcome out-of-order packet delivery. POC works by resequencing packets on the far end of a WAN link “on the fly” to avoid retransmissions that occur when packets arrive out of order. By performing the functionality in a dedicated WAN optimization device (as opposed to an end station or router), enterprises have the scalability needed to handle high volume, high throughput data streams with minimal added latency.

Quality of Service (QoS)

In an effort to maximize WAN utilization, most enterprises will oversubscribe their WAN links. When demand exceeds the capacity of a WAN link, and traffic is contending for the same limited resource, less important traffic (such as web browsing) may take bandwidth away from business-critical applications. To prevent this, some WAN acceleration solutions implement Quality of Service techniques to classify and prioritize traffic based on applications, users, and other criteria.

QoS typically involves three primary functions, which can all have a significant impact on application performance:

- **Packet Marking**, which is a means whereby network elements provide different levels of service to different packets, based on markings in the IP header.
- **Application Classification**, which enables different priorities and handling instructions to be applied to individual types of traffic.
- **Queuing and Shaping**, which improves traffic delivery through various congestion points in a network using queuing policies, dropping disciplines, and service disciplines.

Conclusion

There are various challenges that can adversely impact WAN performance, including limited bandwidth, high latency, and packet loss. In many instances, these issues are intertwined and require a combination of WAN optimization techniques to ensure maximum application performance. Using a subset of the required tools can lead to underwhelming or unexpected performance results. As an analogy, doing WAN deduplication without loss mitigation is like going 100 MPH on a freeway riddled with potholes. Similarly, doing latency mitigation without QoS is like driving a Ferrari on a heavily congested road.

The most effective WAN acceleration solutions use numerous optimization techniques to address all WAN challenges at a macro level. By understanding these optimization technologies— one can better understand WAN acceleration and the role it can play throughout the enterprise. ■