High performance WAN

Are you moving in the right direction?

wanstor

Introduction

Wide area networks (WANs) are critical to the IT infrastructure underlying all business-critical applications, be that in data centres or in the public/private cloud. But for large, distributed business and not for profit organisations, maintaining network links to each branch office can be costly.

Many IT teams fall into the trap of thinking that WAN optimization and acceleration is as simple as boosting the bandwidth in a slow office or dropping a single device into the network. However at Wanstor we know this isn't the case and for best performance a WAN strategy requires bringing together multiple networking and security technologies together.

In this article we will cover some of the most common WAN strategy challenges that IT teams are facing. We offer some practical advice which you can implement into your organisation straight away.

Can reduction and compression technologies solve insufficient bandwidth problems?

The explosion of apps consuming LAN speeds has put pressure on WAN sites that do not have access to unlimited bandwidth. It has been obvious for some time that data compression is a key technology for reducing such stress.

Generally data compression technology works well for most data, except for real-time multimedia (e.g. video conferencing) - which is already compressed and can't benefit from simple compression techniques. WAN optimization products implement compression in many ways, including:

Standard compression: This method takes streams of data and sends a reduced version of the content across the circuit, saving bandwidth. Standard compression in a WAN environment has many intricacies, including the choice of algorithm, how compression works across streams, and the interaction between compression and encrypted traffic.

Caching: This technique reduces data by maintaining a stored version of recently requested data objects (typically, files or email attachments) at the remote side of the connection. If a data object is requested a second (or third, or fourth) time and it is in the cache, then that copy is returned, eliminating the need to re-transmit the object from the central site to the remote site.

Caching is especially useful in environments where file sharing is done across the WAN, or where the email server (typically Exchange) is located at the central site and not the remote site.

Deduplication: This approach reduces data by detecting duplication in streams of bytes. Deduplication is a term from the world of storage and backup systems

The actual details relating to each of these algorithms is used and whether the vendor calls it caching or deduplication is mostly irrelevant. One important difference is that caching nearly always requires a hard disk of some sort to hold cached data, while deduplication is handled on in real time without any persistent storage.

At Wanstor we suggest Network Managers interested in data compression techniques to reduce bandwidth, evaluate products only by putting them into place in their own networks and comparing the results. The most important detail about data compression is that it requires two devices, one on either end of each WAN circuit or virtual connection.

Compression product makers have tried to mitigate the need for deployment and management of hardware by providing compression devices as virtual machines. The idea is to offer compression software that runs directly on end-user devices, and to introduce many other WAN optimization and acceleration techniques into their products to provide more all-in-one solutions.

Can application optimisation solve app issues?

Although compression techniques can provide performance increases in terms of the WAN, optimising apps to run over your organisations WAN's actually offers benefits far beyond simple compression. Application optimisation can often be provided by the same hardware used for compression, but there is a key difference: Application optimisation requires just one device next to the app server. Because the application optimisation directly affects web traffic, optimisation benefits all app users, not just WAN users.

Application Optimisation	Benefit
Better use of browser objects such as JavaScript	Many application developers have the Javascript browser re-downloaded and other browser objects, such as style sheets, each time a different page is referenced. Application optimization tools can re-write pages when required to make sure that these large objects are cached in the browser. Reordering objects can also make pages render faster, giving a better user experience.
Compression and optimisation of content and images	Web browsers internally support compression without requiring any add-on software; most web servers don't bother to compress objects. They simply compress as and when required. This helps speed to access and reduces network load.
HyperText Transfer Protocol (HTTP) Extensions and support for emerging standards such as the SPDY protocol	HTTP, the protocol underlying Hypertext Markup Language, has always been known to be inefficient. Acceleration hardware can help to interweave connections and increase the speed to access over high- latency, low-bandwidth network connections. SSL offload to the optimization-acceleration device can also speed loaded app servers.

Traditionally, application optimization was the realm of a family of products called application delivery controllers or ADC (formerly known as load balancers). But network product makers have migrated these techniques into other devices as well.

Traffic priority and bandwidth management

Voice and video applications require a constant and predictable bandwidth among simultaneous users. Other apps, such as email and web-based programs, tend to be more up and down in their bandwidth requirements.

Wanstor's suggested techniques to provide bi-directional traffic management:

Management Technique	Result
Transmission Control Protocol modification as and when required	By changing TCP window sizes and delaying TCP acknowledgments, individual applications can be better managed and controlled by IT teams.
Application intelligence for User Datagram Protocol applications	UDP-based apps, such as voice and video, are not easily flow-controlled the way TCP apps can. By understanding more about the internals of a UDP app, WAN optimization devices can perform call admission control
Subdividing apps	Some applications mix both delay-sensitive and bulk traffic over the same connection. WAN optimisation devices may be able to break out multiple types of traffic and give different priorities to each type based on deep knowledge of the internal functions of the application.
App identification	Differentiating between business and recreational apps (such as collaborating via SharePoint versus video streaming from YouTube) goes deeper than looking at port numbers. By directly identifying actual applications, WAN optimisation devices can provide granular insight and then limit or guarantee bandwidth as required to meet project objectives.
Time-of-day awareness	Although many data centres run 24/7 x 365 many offices, shops and restaurants are only open for a proportion of the day. This provides the opportunity to use bandwidth (most people buy usage on a per 24hr basis) differently during opening hours. At Wanstor we suggest maintenance activities such as Log transfers, backups, software updates and other maintenance should be pushed to outside core operational hours and will benefit from different bandwidth management rules. By moving a lot of your traffic to off-peak times IT teams can also realise some cost savings as well.

In a fully managed hub-and-spoke network, quality of service (QoS) mechanisms can be used to guarantee particular bandwidth and prioritisation for each application. However as the way we shop, eat and play has changed so have the networks which support businesses and organisations which enable humans to undertake their day to day activities.

Multiple data centres, branch-to-branch communication and the use of generally unmanaged circuits (such as Internet, wireless and shared services) have reduced the ability of simple QoS mechanisms to guarantee acceptable app performance. WAN optimisation and acceleration projects also now require management of bandwidth between sites. Simple mechanisms, such as those found in common edge firewalls with unified threat management (UTM), are not sufficient for the complex requirements of a mix of applications and topology.

Bandwidth management can be particularly trying because true bandwidth management works well only in the outgoing direction for each site. Once the packets have come into a site, they've already consumed bandwidth and pushed out other apps that might have been more important.

Simply dropping packets that exceed predefined limits won't work in most situations. WAN optimisation and acceleration vendors have come up with a variety of techniques to providesophisticated bidirectional traffic management.

Use standard based tools to provide better network visibility

Most WAN optimisation techniques try to improve service with limited resources by controlling use of certain resources. But a significant step toward any WAN optimisation and acceleration project depends on gaining network visibility. At Wanstor we believe it is a “must have” that the network management team can answer questions around the applications using their network such as:

+ What applications are being used?
+ Who is running them and when?
+ How much bandwidth do they use (individually and collectively)?
+ What types of errors are occurring?
+ What response times are users experiencing?
+ Which systems are the top talkers and which are the top listeners?

The old reporting categories must be modified because visibility in current WAN environments involves far more than merely tracking IP addresses and ports. True network visibility extends up the stack to identifying real people and real apps.

Without strong visibility into the network, no WAN optimization and acceleration project can be successful. Control of the unknown simply leads to frustration and confusion, while good visibility into network and app use can also provide metrics to measure overall project or programme success.

Many devices (including switches, routers, firewalls, WAN optimization controllers (WOCs) and application delivery controllers) will send IPFIX and NetFlow data to a management system. Where no IPFIX data is available, both open-source and commercial hardware and software IPFIX and NetFlow exporters are available to give visibility into unencrypted network traffic.

The benefit of choosing IPFIX and NetFlow is that it represents a standard approach, which means that an organisation will be able to gain visibility into different components mixed and matched on its network.

Link balancing and dynamic routing improves network reliability

Although service-level agreements (SLAs) can set expectations, network managers must prepare for the inevitable link downtime that any WAN will experience. When business-critical apps are used over the network, most organisations choose to use dual links into each of their sites to minimize blockages created by traffic peaks or network problems.

Simply having multiple links doesn't ensure high availability, as some mechanism must be in place to use the links. If VPN tunnels are in place, some organisations use dynamic routing protocols such as the Open Shortest Path First (OSPF) protocol to make use of dual links. Having two links on at all times always prompts a return on investment question: How can we use both links and still get the most network for the pounds invested?

WAN optimization and acceleration vendors have introduced a variety of techniques to balance traffic over multiple network links, with varying levels of success. Because TCP/IP networks have their own routing protocols, attempts to force traffic to take a particular route or to signal a route to upstream devices (such as Multi-Protocol Label Switching or MPLS routers) are often complicated and create brittle networks.

While the idea of using as much of two circuits as possible is attractive from budget and theoretical perspective, network managers should very carefully evaluate any vendor proposal to perform outbound load balancing or dynamic link selection.

Experiences with this type of load balancing have not been positive for all businesses. In some cases, these types of technologies have required very specific network configurations for correct operation, and may end up creating more problems than they solve.

Does load balancing improve application reliability?

Although application reliability is not necessarily a WAN-specific concern, the importance of enterprise applications emphasizes the need for more sophisticated types of load balancing and high availability strategies that stretch across data centres.

Traditional load balancing uses a Layer 2 or Layer 3 device as the front-end to a series of systems offering an identical service. As requests come in to the load balancer, it makes a decision based on a predetermined algorithm and passes the request on to whichever system is selected.

The load balancer then manages state information so that further requests from the same client are all directed to the same system. The algorithm chosen can be as simple as a round-robin process or it can be more sophisticated, taking into account CPU utilization, response time and other factors. Originally, the goal of most load balancers was scalability - the ability to handle a greater load than any single server could manage. Over time, the goal has changed. Now, the low cost of server hardware has led many organisations to use load balancers simply for reliability. With two (or more) servers available, uptime can be extended and maintenance windows shortened, even if the load can reside entirely on a single server.

Although people have been talking about global server load balancing for a decade or more, network managers should be aware of one important fact: Global server load balancing has not a solved problem. Because of the way that the Internet, Domain Name System servers and web browsers work, there is no guaranteed reliable approach to providing high availability across multiple data centres.

It should be noted that several techniques have been tried, including DNS-based load balancing and Border Gateway Protocol (BGP) load balancing, but no approach works 100 percent of the time in 100 percent of the possible failure cases. Indeed, the numbers aren't even close to 100 percent. For every global load-balancing technique discussed, there are many potential places where load balancing will not deliver the desired results. Load balancing is usually provided by dedicated software or hardware, such as a multilayer switch or a DNS server. Many experts consider the distinction between hardware and software load balancers to be no longer meaningful.

Wanstor's suggested techniques to provide bi-directional traffic management:

Technology	Most commonly found in	But also available in
Data compression and reduction	WAN optimization controllers (WOCs); discrete hardware appliances or software-based virtual appliances	Some functionality may be available in web security gateways, but pure data compression and reduction are not often found in other product spaces.
Application optimization	Application delivery controllers (ADCs), load balancers	WOCs often include some application optimization features. Web application firewalls are a separate niche.
Traffic prioritization and bandwidth management	Quality of Service (QoS) and visibility products	WOCs often include traffic prioritization; UTM products and next generation firewalls generally include basic bandwidth management and prioritsation.
Routing and link balancing	Branch and edge firewalls or combination router–virtual private network (VPN) devices	Stand-alone edge routers all generally have this capability, but the location of the router outside of the firewall (which prevents it from seeing into encrypted VPN tunnels) pushes this feature into whatever device handles VPNs for the branch.
Security features; use and misuse controls	UTM products and next-generation firewalls	Web security gateways and proxy servers may include limited web focused features. Standalone IPSes are rarely used in the branch when UTM or next-generation firewalls are available.

Integration of security tools

WAN optimization is usually considered a largely technical exercise, the goal of which is to get more value out of each pound invested for connectivity. Many network managers now take a more holistic view of network use, and look to security-focused products to help them control overall use of both enterprise and Internet apps. Because most WANs already have a firewall device at the border of each remote site, these devices may be called upon to provide more than simple firewall and VPN services.

Security device manufacturers are bringing many branch management features to their edge devices, including URL and content filtering, app identification and control, bandwidth management, intrusion prevention and antimalware. At Wanstor we believe Network Managers should consider including the capabilities of branch firewall devices in their overall network optimization plan for several reasons.

First, these devices are typically already in use, so activating additional capabilities may be as simple as a few mouse clicks or a low-cost subscription add-ons. Branch firewalls are key parts of the WAN. What's more, changes to traffic profiles or traffic types will also affect the operation and capabilities of the firewalls.

How Wanstor can help Network Managers optimise their WAN

In this blog post we have offered some suggestions around common WAN challenges which Network Managers can use to help improve their WAN performance. Wanstor also offers a range of network optimisation solutions to help:

+ Reduce application latency to remote end-users
+ Create multiple pathways to ensure application availability
+ Centralize the network environment
+ Decrease operating and management costs
+ Maximize bandwidth utilization
+ Postpone the need to upgrade WAN bandwidth
+ Improve disaster recovery position by speeding backup and data replication over the WAN

For more information about Wanstor networking services, please click here

124-126 Borough High Street | London | SE1 1LB | info@wanstor.com

wanstor