GIGAOM RESEARCH

The promise of next-generation WAN optimization

Enrico Signoretti

September 9, 2014

This report is underwritten by Bridgeworks.

TABLE OF CONTENTS

Executive summary

Bandwidth, throughput, and latency aren't issues when you are within the boundaries of a data center, but things drastically change when you have to move data over a distance. Applications are designed to process data and provide results as fast as possible, because users and business processes now require instant access to resources of all kinds. This is not easy to accomplish when data is physically far from where it is needed.

In the past decade, with the exponential growth of internet, remote connectivity, and, later, large quantities of data, lack of bandwidth has become a major issue. A first generation of wide area network (WAN) optimizing solutions appeared in the market with the intent of overcoming the constraints of limited bandwidth connectivity. Sophisticated data-reduction techniques like compression, deduplication, traffic shaping, caching, proxying, and so on were integrated to minimize traffic between data centers and branch offices or for DC-to-DC communication. WAN optimization can contribute significantly in improving the quality and the quantity of services delivered to branch offices, replicate storage at longer distances for disaster recovery (DR) or business continuity (BC), reduce WAN costs, and improve mobile connectivity.

Recently things have changed significantly. Traditional WAN optimization was mainly conceived for solving lack of bandwidth in a time when legacy protocols were designed for local area network (LAN) connectivity. Data was neither compressed nor encrypted, and computers were unable to manage huge amounts of complex data. Now things are the other way around: High-bandwidth links (10Gbs or more) are considerably cheaper than in the past, new protocols are emerging, data is often compressed and encrypted at the source, and even mobile devices can concurrently manage multiple huge data streams. Traditional WAN optimization was simply not designed to efficiently manage these new requirements . Efficiency, utilization, and latency are the real issues now.

Next-generation WAN optimization, designed with a radically new philosophy, has the right characteristics to offer unprecedented scalability, better latency management, and uncompromised link utilization.

One new approach, rooted in a deep knowledge of storage DNA, looks at the problem in a radically different way compared to what we've seen before. It addresses the problem by keeping in mind modern types of data (compressed and encrypted) and focusing on mitigating latency issues while maximizing

efficiency and predictability at scale. The result is overall TCO improvement, better latency management, and an outstanding utilization rate of high-bandwidth links.

Key highlights from this report include:

- Traditional WAN optimization is not sufficient for dealing with high-performance WAN connectivity and modern types of data. As the amount of stored data grows, widely used compression and encryption techniques offset traditional WAN optimization efficiencies.
- Efficiency — both in utilization rate and latency mitigation — is now key. Next-generation WAN optimization is a new concept designed with scalability, efficiency, and TCO in mind.
- Next-gen WAN optimization provides benefits similar to traditional WAN optimization but with improved TCO, freedom to scale and the ability to prepare the infrastructure for new needs (such as cloud, object storage, and mobile).

Primary use cases for WAN optimization

Techniques for increasing the efficiency of WAN connections are useful where limits and constraints of bandwidth and latency exist. WAN optimization can be used on private links or the public internet. Depending on the particular implementation, it can yield various benefits to many types of data movements and applications. Examples of possible applications of WAN optimization include:

- Communications between data center and remote offices. Connecting remote offices can be costly, and in many countries, it is not unusual to find poor, oversubscribed public connectivity infrastructures with limited bandwidth and high latency.

- Storage replication. Consistent storage data replication can be tricky. Block and file storage protocols work in different ways, and, furthermore, most of them are lossless protocols that can't be easily encapsulated into lossy protocols like TCP/IP. For storage, latency is one of the major issues, and predictability is another important aspect of the communication.

- Disaster recovery and business continuity: When it comes to DR and BC, connection and distance between data centers are some of the most important factors for maintaining an acceptable RPO (recovery point objective).

- Mobile connectivity. Users ask for instant access to enterprise data, and now they have access to fast networks on their end. But latency and network unpredictability could make it unreliable.

- Cloud. Especially when the enterprise thinks about hybrid cloud implementations, WAN connectivity is a fundamental layer of the architecture. Once again, bandwidth, latency, and predictability of the WAN link are the parameters to look at for achieving the best user experience and delivering reliable services.

- Migrations. Any infrastructure IT life cycle endures migrations from time to time. The larger the infrastructure, the more frequent the migrations. When migrations also involve moving services from one site to another, you always need to provide a reliable WAN link to optimize data movements between sites to minimize downtimes.

- Content delivery. Video and photos are now permanent features in our personal and professional lives. They are bigger than other file types and use massive storage capacity. It's usually easier, primarily for mobile devices, to transfer media files multiple times than to store them locally.

- Content-delivery networks. Relevance of data streaming is dramatically growing, and end users always look for the best user experience. Efficient and optimized data movements from central repositories to distribution networks and end points are fundamental to pursue high-quality streaming and the best user experience with larger media files.

- Satellite communications. Academic research, oil and gas, and many other civil and military applications produce massive amounts of data all around the world, even in the middle of nowhere. And it has to be moved and processed at the fastest speed to become effective. WAN optimization can significantly improve the efficiency of cumbersome and expensive satellite TCP data links.

These are only a few examples. WAN optimization use cases are countless in real-world scenarios, and it is helpful every time you need to efficiently move data over a WAN link.

Business benefits of WAN optimization

The impact and, consequently, the value of WAN optimization is particularly evident when performance and savings are the two flip sides of one coin. Better efficiency and higher utilization are synonyms for savings, while predictability and optimization lead to improved performance.

Doing more with less is the easiest way to describe WAN optimization, which means lowering WAN costs and saving money. In this case, traditional WAN optimization techniques enable the acquisition of less capable and cheaper links (i.e., less bandwidth or higher latency). In fact, traditional WAN optimization, which is particularly focused on bandwidth through traffic shaping, caching, and compression techniques, can be useful in those cases where bandwidth is scarce.

On the other hand, WAN optimization saves money on the network cost itself (reduce WAN cost), and other important productivity benefits result. For example, when speed is the most important factor for making business decisions, you can't rely on an unpredictable connection. In most scenarios, application performance and predictability are all related to latency and how it is managed. Even though latency is not avoidable (especially at long distances and on multi-hop VPN-based TCP/IP connections), next-generation WAN optimization techniques show the ability to mitigate its effects.

More generally, depending on the implementation, WAN optimization can heavily contribute to improved performance and lower costs in many critical scenarios. Possible applications range from improved flexibility of DR and BC to improved compliance (i.e., reducing remote backup windows or backup vaulting) thanks to better bandwidth and latency management. At higher levels, business applications can really benefit from a faster access to data, especially in sensitive environments like banking and financial markets.

WAN optimization, thanks to better utilization and predictability added to standard WAN connections, has a positive side effect in the capacity-planning process as well: It can help delay the acquisition of higher-performance connections while providing room for data growth.

Traditional WAN optimization: It's all about perception

Traditional WAN optimization mainly focuses on a series of techniques aimed at reducing the impact of data transfers in low-bandwidth scenarios by simply reducing the amount of transferred data.

This approach has two direct consequences:

- It optimizes the use of low-performance connections, giving the impression of greater bandwidth availability. Caching and proxying in conjunction with compression and deduplication reduce the data footprint during movements.
- The end user, thanks to some specific traffic and protocol optimizations, has the perception of lesser latency at the user-interface level. For example, though SMB is a chatty protocol, with the right traffic inspection and reorganization, browsing a remote complex directory tree could be much faster and smoother.

Most sophisticated implementations also have other features for error checking and correction as well as packet shaping or connection management (such as limiting the number of concurrent connection and the per-user data rates). The final objective is always the same: reduce data transmission and retransmission.

Traditional WAN optimization partially addresses latency issues by running some infrastructure services directly on the appliance (e.g., small virtual machines running an active directory, DNS, or DHCP) or by tweaking TCP/IP parameters (e.g., window-size scaling, selective acknowledgements, and so on). The undeniable drawback here is that some of these capabilities can compromise the TCO due to the greater complexity of the infrastructure to manage.

Even though usually the results are evidently positive and the end user immediately perceives the difference before and after the introduction of traditional WAN optimization, original design limitations remain that could strongly compromise its real effectiveness in modern environments.

New challenges need a paradigm shift

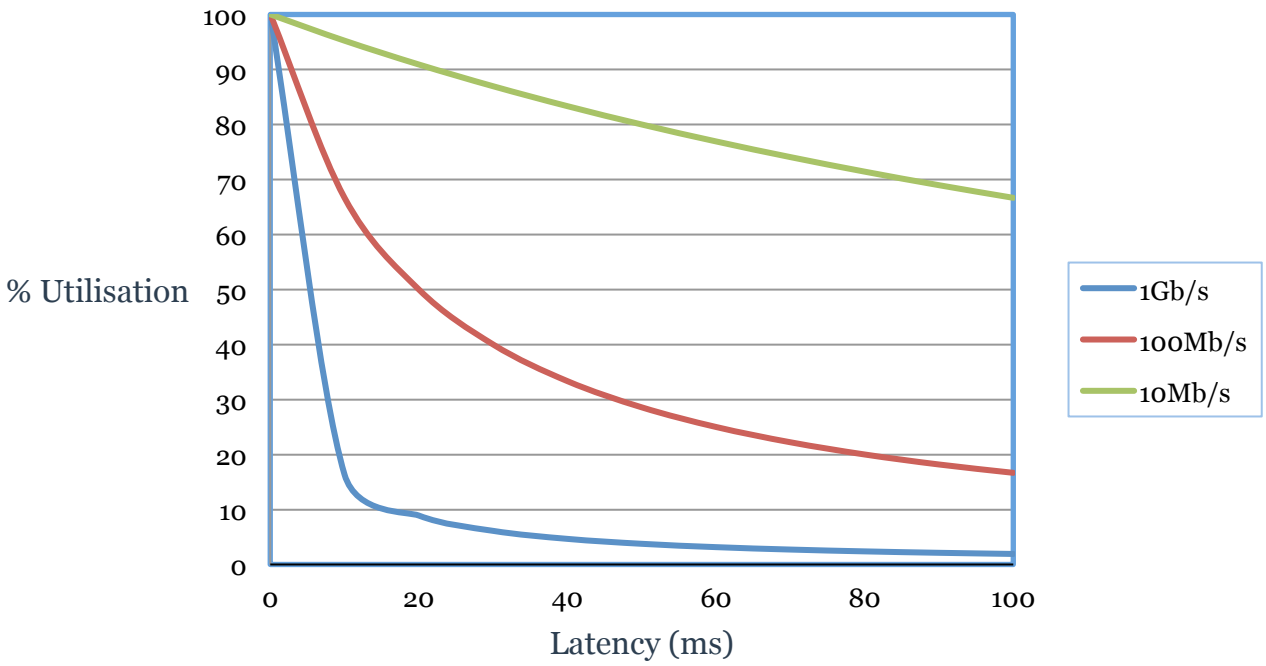
All things about storage and data are changing dramatically quickly. WAN optimization is no exception.

Enterprises and service providers now have access to high-bandwidth WAN links. Especially in metropolitan areas, it's not difficult to see 1 Gbit/s or even 10 Gbit/s options (with 25 Gbit/s around the corner) connecting data centers or remote offices. At the same time, higher bandwidth doesn't mean lesser latency.

Contrary to general thinking, the higher the bandwidth, the more evident the latency problem becomes, especially when TCP/IP is involved.

The real efficiency of these huge data links is compromised by how TCP/IP works. One of the key ways it optimizes the data flow across the network, without overrunning the receiving end, is by the number of packets it sends in a group to the other end. This is known as the receive window size (RWS). Once TCP/IP has sent the group of packets, it will wait until it receives an acknowledgement (ACK) packet from the other end before sending the next group of packets. While this is not a problem when latency is low, it will totally characterize communication in the opposite case. High latency can definitely compromise network performance at levels such that a 1 Gbit/s link could resemble 100 Mbit/s — or worse.

Latency versus utilization



Source: Gigaom Research

A second problem arises from the kind of data we move today. Compression and deduplication are no longer effective because many new file formats and data streams are compressed and encrypted at the source. Modern computers as well as smartphones or tablets have enough CPU power to decompress and decrypt data, but on the flip side, local storage capacities are limited due to the use of flash memory (vastly used to improve power consumption and speed up data access). In any case, it's not only a technical issue: Data encryption is now a basic security request in any enterprise environment, especially considering privacy concerns or where data travels over long distances. Moreover, some protocols, like server message block (SMB), for example, are now more efficient than ever and can be harder to further optimize.

Last but not least, traditional WAN-optimization techniques rely on large amounts of CPU, RAM, and local storage. Compression and deduplication use loads of CPU while RAM and storage are fundamental for heavy caching, proxying, and storing data as well as metadata and hash tables. The main practical problems of this approach are that CPU-bound efficiency limits scalability and has higher acquisition

costs (TCA). In fact, finding traditional reasonably priced solutions capable of working at a speed higher than a few Gbit/s is difficult.

Looking at next-generation WAN optimization

To obtain scalability, mitigate latency, and improve bandwidth utilization on high-performance connections, WAN optimization has to be redesigned from the ground up.

Moving from data optimization (basically data reduction) to latency optimization is not easy. Especially when looking at storage protocols, understanding the effects of latency on communication is important. Receive window size (RWS) is one important aspect to consider, but understanding the nature and managing the behavior of each single protocol is important as well. Storage block protocols, for example, are usually based on SCSI (often encapsulated in other protocols and then re-encapsulated in TCP/IP packets), which is complex and difficult to tune. Each of these protocols has its own peculiarities and should be treated for what it is. Other storage protocols are emerging too: Object storage, for example, doesn't use SCSI to access and replicate data between sites, while all data chunks are usually encrypted at the source. On the other hand, legacy protocols like FTP are still used in many data centers all around the world and have their own unique characteristics too.

In these cases, trying to apply compression and deduplication techniques could be useless and counterproductive if the primary goal is predictability.

Caching, when properly implemented, remains a key factor to minimize latency and boost overall efficiency.

- It doesn't use a large amount of CPU and can heavily contribute to faster data access in many scenarios.
- It doesn't have to be big, but it does have to be managed by clever algorithms and synced between the sites.
- In complex topologies, involving many different WAN optimization appliances and end point devices, the cache should be organized to provide the right quantity of data to feed various requests as a function of the end points' capabilities, thus avoiding data starvation, data bunching, or timeouts.

Scalability, one of the most important characteristics of modern IT infrastructures, is often underestimated until it's too late. When it comes to networking equipment and one single link to optimize, it is even more complicated, and scale-out architectures are not applicable. Backend architecture efficiency is a must, and avoiding CPU-bound techniques is the second step toward providing the right level of vertical scaling.

Now that 1 Gbit/s is quite common, higher-bandwidth links are becoming cheaper, data is exponentially growing, and private and hybrid clouds are gaining traction, it's more important than ever to look at architecture designs offering the best potential in terms of efficiency and scalability.

Implementing next-generation WAN optimization

Protocol understanding is a main point for optimizing the communication over a WAN link. Because storage data comes in many different block sizes, splitting data blocks into the most efficient size for WAN characteristics is important to boost the utilization. Equally important is efficiently reconstituting the data on the other side of the WAN link. Designing a totally different kind of WAN optimizer requires a deep understanding of storage protocol, focusing mainly on latency-effect mitigation, and a knowledge of latency's role in long-distance data movements (due to the way TCP/IP works). The result should be a mechanism that can transparently reorganize this type of communication and avoid the constraints imposed by RWS.

In practice, consider appliances on each side of a link. After sending a first group of packets, instead of waiting for the ACK signal, open a new virtual connection and send another group of packets to the other end of the wire. This operation is repeated many times, and it is constantly tuned until the appliances strike the right balance between the number of virtual connections and packets sent to fully utilize the WAN link. Physical latency is unaffected, but line utilization goes up to 95 percent or more. An accompanying software employs different algorithms that constantly monitor, review, and adjust the set of network parameters and virtual connections to ensure optimum performance at all times. A smart caching mechanism coordinates and syncs caches between appliances. It leverages the same intelligence of the networking layer, and its cleverness comes from the ability to understand the right quantity of data to cache as a function of a device's ability to send and ingest data over the involved protocols.

A solution that can take full advantage of these mechanisms must have the ability to rebuild the original data block stream even when packets arrive in disarray or have to be retransmitted. It should also support storage protocols that cover all block and bulk file storage (iSCSI, SCSI, SAS, FC, and FTP) as well as major vendor replication protocols, including REST, which can be helpful for multi-site object and cloud storage deployments or hybrid cloud implementations. This kind of solution can easily handle multiple 10 Gbit/s links and support various types of topologies for different deployment scenarios.

Key factors for lowering TCO and improve ROI would be tremendous ease of use, highly automated functionalities, and transparency of the solution.

Key takeaways

Traditional WAN optimization is no longer sufficient to cope with the challenges imposed by new high-performance WAN connectivity and modern types of data. While stored data is growing exponentially, compression and encryption techniques, widely used to save space and improve security at the source, offset traditional WAN-optimization efficiencies.

Today, in terms of utilization rate and latency mitigation, efficiency is the most important aspect. Instead of focusing on bandwidth through data reduction, next-generation WAN optimization has now found a radically new and better way to achieve this goal: A new concept designed from the ground up with scalability, efficiency, and TCO in mind.

Next-gen WAN optimization brings similar, but improved, benefits of traditional WAN optimization where it wasn't possible in the past. The most evident enhancements for the modern enterprise are improved TCO, freedom to scale, and readying infrastructure for new needs such as cloud, object storage, and mobile.

The characteristics of next-gen WAN optimization promise a positive impact on businesses, from many points of view:

- Lower WAN costs, thanks to higher efficiency and utilization rates
- Better business decisions, thanks to improved application and performance predictability as well as the ability to move more data more quickly
- More flexible DR/BC scenarios, thanks to better RPOs on farther remote sites, due to the improved efficiency of WAN links
- An improved level of compliance with shorter backup windows, thanks to higher utilization rates of the links
- Ease of adoption of new cloud and mobile technologies, thanks to the new supported protocols

Next-generation WAN optimization at its best focuses on link efficiency and utilization, ability to scale, transparency, and ease of use, which enables the enterprise to think about WAN connectivity in a totally different way, combining savings with the freedom to earnestly look forward to modern high-speed and heavily distributed infrastructures.

About Enrico Signoretti

Enrico Signoretti is an independent consultant, trusted adviser, and blogger who has been immersed into IT environments for over 20 years. His career began with Assembler in the second half of the 1980s before he moved on to UNIX platforms until now in the "Cloudland." He is constantly keeping an eye on how the market evolves and is continuously looking for new ideas and innovative solutions. He's a fond sailor and unsuccessful fisherman.

About Gigaom Research

Gigaom Research gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports, and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, Gigaom Research addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: research.gigaom.com.