Closing the WAN Intelligence Gap

Ashton, Metzler & Associates

Leverage Technology & Talent for Success

August 2006

Closing the WAN Intelligence Gap

Table of Contents

1.0 Introduction and Objective

The primary goal of any IT organization is to ensure that the applications that the company uses to run its business can be accessed securely and that the response time and availability of those applications are acceptable. Throughout this white paper, that goal will be referred to as the *IT Goal.*

While achieving the IT Goal has never been easy, there are some fundamental shifts under way in both the technology and business environments that are making achieving that goal significantly more difficult. This white paper will describe these shifts and will illustrate the demands that these shifts are making on the Wide Area Network (WAN). This white paper will also describe the fact that the intelligence that has been deployed in the WAN over the last several years is not sufficient to support these new demands. Throughout this white paper, the gap between the demands that applications are placing on the WAN and the ability of the WAN to support those demands will be referred to as the WAN intelligence gap. Having identified the WAN intelligence gap, this white paper will conclude by outlining an approach to optimizing the performance of applications that will enable IT organizations to close this gap.

In order to include insight from IT organizations in this white paper, Ashton, Metzler & Associates interviewed two IT professionals. One of these interviewees is a technology manager at a high-tech company. The other interviewee is a global IT architect for a medical organization. These interviewees will be referred to as the *Technology Manager* and the *IT Architect.*

2.0 The Shifting Business Environment

One of the key shifts in the business environment is the movement of employees out of a headquarters site and into branch offices. To put this in perspective, as recently as a few years ago, the vast majority of a company's employees resided in a headquarters facility and accessed corporate applications locally. Currently, the vast majority of a company's employees reside in branch offices and access applications remotely. The Technology Manager put this into perspective when he stated, "I have meetings with people in China and we are all working on the same document, and it is almost like we are all in the same room."

Another key shift in the business environment is the growing emphasis that companies are placing on security and compliance. One manifestation of this shift is the increased use of secure protocols such as HTTPS. Another manifestation is the ongoing consolidation of servers out of branch offices and into one or more centralized data centers. To understand how this consolidation is driven by compliance, consider the Sarbanes-Oxley Act (SOX). Like most of the recent spate of regulations, SOX requires that companies place a greater emphasis on assuring the accuracy, security and confidentiality of data. It is very difficult to do this when there are multiple copies of the company's data on servers in branch offices.

A related shift in the business environment is the movement on the part of many medium- and large-sized organizations to reduce the number of data centers that they support. For example, HP recently announced that it was going to reduce the number of data centers that it supports from eighty-five down to six.[1]

[1] "Hewlett-Packard picks Austin for two data centers," http://www.statesman.com/business/content/business/stories/other/05/18hp.html

The result of these trends is that instead of accessing an application locally over a high-speed, low-latency LAN, a majority of a company's employees now access applications using secure protocols such as HTTPS over a relatively low-speed, high-latency WAN. In addition, as companies consolidate their data centers, they increase the distance between their remote employees and their consolidated data centers. This has the effect of further increasing the WAN latency for these remote employees.

Both the increased WAN latency faced by the majority of a company's employees and the use of secure protocols such as HTTPS make the challenge of achieving the IT Goal notably more difficult. That is because increased WAN latency typically results in degraded application performance. In addition, the use of secure protocols means the data stream is now encrypted and cannot be processed by many of the traditional application optimization appliances.

3.0 The Shifting Technology Environment

As mentioned in the introduction, the primary goal of any IT organization is to ensure that the applications that the company uses to run its business can be accessed securely and that the response time and availability of those applications are acceptable. Achieving that goal requires an understanding of trends in the development of enterprise applications as well as trends in the deployment of WANs. The next two sections of this white paper will discuss trends in the development of enterprise applications, and subsequent sections will discuss trends in the deployment of WANs.

3.1 Application Complexity

The era of mainframe computing was characterized by the deployment of monolithic applications. The phrase *monolithic application* refers to an application in which all the relevant functionality (i.e., the user interface, the application logic and the database) resides on a single computer.

Most companies have moved away from deploying monolithic applications and toward a form of distributed computing that is often referred to as *n-tier applications.* N-tier applications are applications in which the functionality that had at one time been provided by a monolithic application is now decomposed into multiple tiers. Since these tiers are implemented on separate systems, n-tier applications are more impacted by WAN performance than are monolithic applications.

For example, the typical three-tier application is composed of a Web browser, a Web server and a database server. The information flow in a three-tier application travels in a linear fashion from the Web browser to the Web server to the database and then back again, over the Internet, using standard protocols such as HTTP or HTTPS.

The movement to a service-oriented architecture (SOA) based on the use of Web services-based applications represents the next step in the development of distributed computing. This approach to distributed computing has been adopted by the majority of enterprise IT organizations as well as software vendors such as Microsoft, SAP and Oracle.

Just as n-tier applications are more impacted by WAN performance than are monolithic applications, Web services–based applications are more impacted by WAN performance than are n-tier applications. To understand why that is the case, consider the three-tier application architecture that was previously discussed. In a three-tier application, the Web server and the database typically reside in the same data center. As a result, the impact of the WAN is constrained to a single traffic flow, that being the flow between the Web browser and the Web server.

In a Web services-based application, the Web services that compose the application often run on servers that are housed within multiple data centers. As a result of housing the Web services in multiple data centers, the WAN impacts multiple traffic flows and hence has a greater overall impact on the performance of a Web services–based application than it does on the performance of an n-tier application.

3.2 The Application Development Environment

The IT Architect commented that although application architectures have evolved from being monolithic to being highly distributed, most software is written as if all the software were running in a single data center. He added that HTTP is a very chatty protocol and that if an application is written to use HTTP, "the performance is just terrible."

The IT Architect also pointed out that in the typical application development environment, the focus is on delivering the promised functionality on time and with relatively few bugs. In most cases, there is little if any focus during the design and development of an application on how well that application will run over a WAN. He added that the worst architectures come from some of the software vendors and that "there is no rhyme or reason to what some of them do."

This lack of focus on application performance often results in what is referred to as *chatty applications.* A chatty application requires several round-trips between the client and server to complete a single transaction. This lack of focus on how well applications will run over a WAN takes on growing importance because, as mentioned in the previous section, application architectures are evolving in such a way that the WAN is having an ever-increasing impact on application performance.

The IT Architect stated that major software vendors such as Oracle and SAP have made a significant investment in SOA and composite applications. He added that these applications will work only if the user interface is based on a portal framework that provides access to all types of applications. The use of portals to access multiple applications will be discussed in the next section of this white paper.

4.0 The Use of Enhanced Portals

Historically, the primary use of a portal has been to allow users access to static information. However, software vendors such as SAP, Oracle, Microsoft and Siebel have begun to make enhanced use of portals in order to also provide transactional interfaces to back-end systems and to automate complex business processes.

A good example of the enhanced use of portals is the Microsoft Office SharePoint Sever 2007 product. According to Microsoft, "Office SharePoint Server 2007 is an integrated suite of server applications that improves organizational effectiveness by providing comprehensive control over electronic content, accelerating shared business processes and facilitating better-informed decisions and information-sharing across boundaries."[2]

Figure 1 *(page 5)* depicts some of the applications that Microsoft is combining on top of Office SharePoint Server 2007.

[2]Microsoft Office SharePoint Server 2007 Evaluation Guide, February 2006

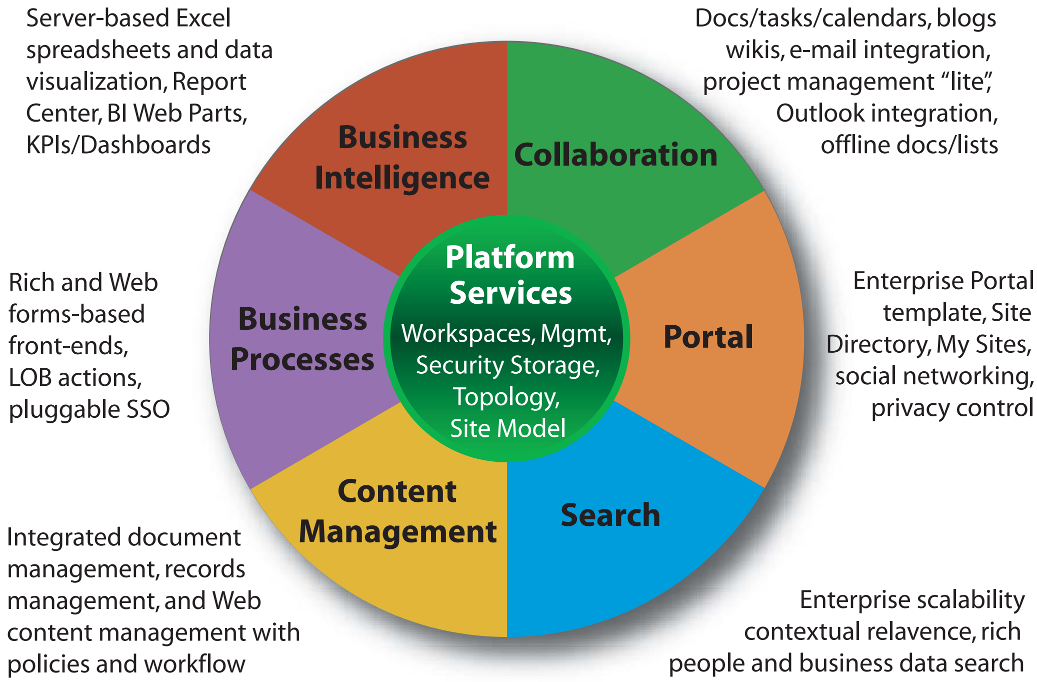

Figure 1. Microsoft® SharePoint® 2007

Microsoft states that SharePoint 2007 was designed to utilize an SOA and support interoperability standards such as XML and SOAP. Microsoft also states that as a result of these design decisions, it is easy to integrate SharePoint 2007 with existing business processes and applications. While SharePoint 2007 was designed to be easily integrated with existing business processes and applications, it was not designed in such a way that it is easy to improve the performance of the application. For example, most of the content is marked as private and noncacheable. In addition, in conjunction with SharePoint 2007, new versions of 2007 Microsoft Office System applications such as Word, PowerPoint and Excel will use a compressed XML document format termed Open XML. These documents cannot be further compressed or differenced by conventional means.

The IT Architect commented that in order to accelerate applications such as SharePoint, you must be aware of the application characteristics. He added that if he wants to open a Word document using Remote Procedure Call (RPC) over HTTP,[3] there is nothing he can do about either RPC or HTTP. What he can do is implement techniques such as object caching and byte-level differencing.

The Technology Manager agreed on the need to understand application characteristics. He stated that instead of throwing bandwidth at a performance problem, a more effective way to increase application performance is to reduce application turns. He also stated, "There are some critical things that you can only do at the application level. For example, certain mechanisms or protocols used in the application can only be optimized with a thorough knowledge of those mechanisms and protocols."

[3]It is common to refer to protocols such as HTTP as an application. Throughout this white paper HTTP will be referred to as a protocol and the term application will refer to enterprise applications such as Microsoft® SharePoint® 2007 or Oracle Financials. Analogously, throughout this white paper, the phrase application layer acceleration will refer to acceleration techniques that are applied to enterprise applications such as Microsoft® SharePoint® 2007 or Oracle Financials.

A thorough knowledge of an application's semantics and protocols is beneficial in accessing complex collaboration applications such as Microsoft's SharePoint® Services. SharePoint supports a variety of protocols for requesting information over a WAN, which gives clients a choice of applications from which they can access remote files. Figure 2 illustrates three ways that a client can access a SharePoint Services file. Depending on the application that is requesting the document, the client can use HTTP via a Web browser, RPC via MS Word, or Web-based Distributed Authoring and Versioning (WebDAV) via Windows Explorer.

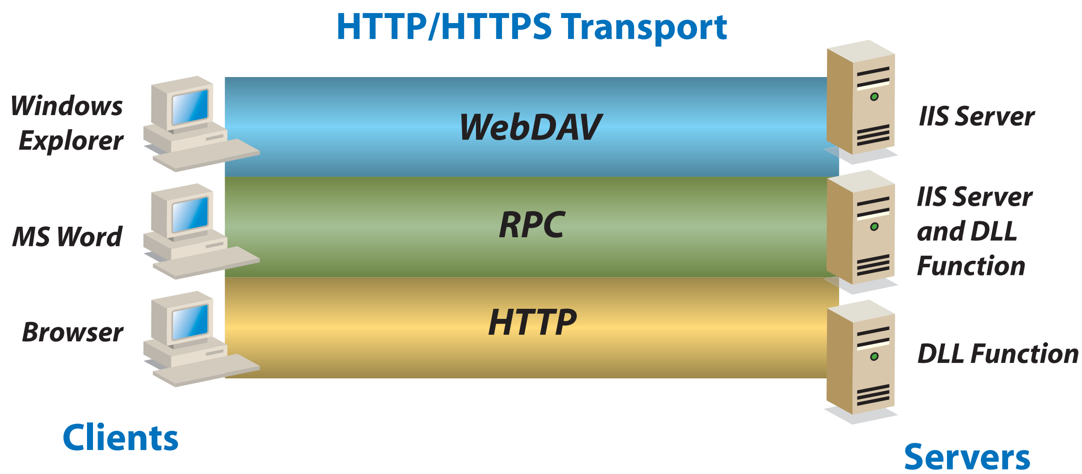

Figure 2. Three Ways to Access a Microsoft® SharePoint® Services File over a WAN

By understanding these protocols and, more specifically, how applications use them to request and receive information, you have more control over how to more efficiently access information. In addition to understanding the protocols that enable access to applications, it is equally important to know what type of information you are accessing.

The Technology Manager added, "One of the things I am working on is to accelerate dynamic page loads by reducing turns." He added that the trend in applications is toward dynamic pages and that the word *dynamic* is key. He stated that "you need a good understanding of the application to understand that there are certain changes in a dynamic page that make it different from a static page and that you need the ability to pick out the changes and keep the static pieces the same."

The acceleration of applications based on an understanding of the characteristics of the application will be discussed in greater detail in the next two sections of this white paper.

5.0 The WAN Intelligence Gap

Over the last few years, applications have become notably more demanding. As was previously discussed, one of the factors that are causing applications to become more demanding is that both organizations and applications are becoming increasingly distributed. The ongoing distribution of people and applications is one of the primary factors driving a rapid increase in WAN traffic. For example, one recent study[4] indicated that WAN traffic volumes are doubling every nine months.

As was also previously discussed, applications are becoming increasingly complex in terms of the enhanced use of portals, the wide range of application objects that need to be supported combined with the variety of ways that these objects can be accessed, the requirement to use secure protocols and the ongoing use of chatty protocols. In addition to the use of chatty protocols, applications themselves are also becoming increasingly chatty.

Over the last few years, some additional functionality has been added to the WAN with the intention of improving application performance. For example, QoS functionality has been added to the WAN in the form of both more intelligent queuing in enterprise routers and the deployment of MPLS-based services on the part of service providers. In addition, the last few years have seen the deployment into the WAN of techniques such as compression, caching and protocol acceleration. However, these techniques have been deployed at the lower layers of the OSI protocol stack. What we have not seen over the last few years is the deployment into the WAN of innovative functionality that allows the WAN to perform application-layer acceleration.

As depicted in Figure 3, the demands of a company's applications are increasing at a significantly faster rate than is the intelligence in the WAN to support those applications. The difference between the burgeoning demands of enterprise applications and the ability of the WAN to support those demands is what was referred to in the introduction as the WAN intelligence gap.

The Technology Manager commented on the WAN intelligence gap when he stated that "organizations have been distributed and applications have responded to that, but the WAN has not." He also stated that the limitations of the WAN are preventing organizations from taking full advantage of the new portal-based applications.

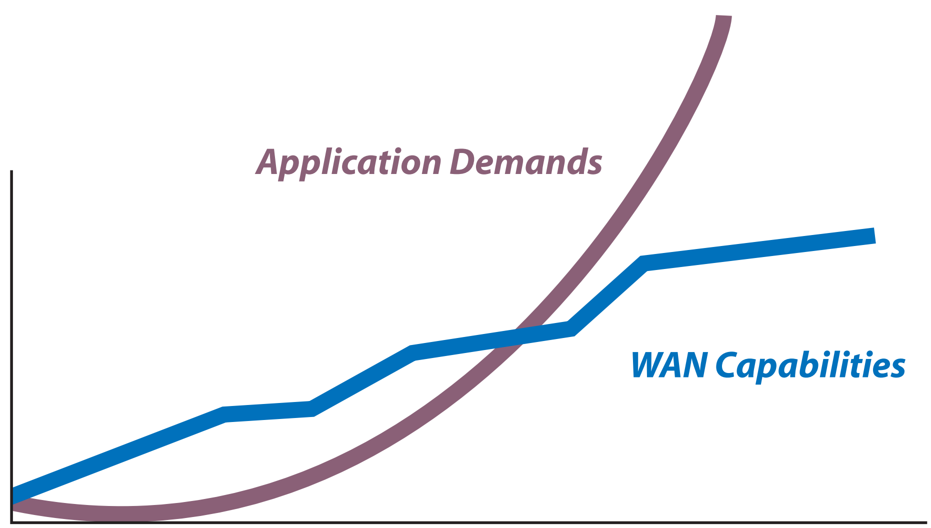

Figure 3. WAN Intelligence Gap

[4]"Migrating the Cisco WAN to the Cisco 7600 Series Router," Cisco IT@Work Case Study

6.0 The New Layer of WAN Intelligence

As described previously, ongoing shifts in the business and technology environments are placing increasingly stringent demands on the WAN, which has led to the WAN intelligence gap. The characteristics of the WAN that contribute to this gap include insufficient bandwidth, high latency, packet loss and congestion. Table 1 contains a listing of some of the techniques that organizations have begun to deploy in order to reduce the WAN intelligence gap. A brief description of these techniques is contained in the glossary *(page 12)*.

WAN Characteristic	WAN Optimization Techniques
Insufficient Bandwidth	Data Reduction: • Data Compression • Differencing
High Latency	Session Optimization: • Increase TCP Window Size Mitigate Round-trip Time • Request Prediction • Response Spoofing
Packet Loss	Congestion Control Forward Error Correction (FEC)
Network Contention	Quality of Service (QoS)

Table 1: Techniques to Improve Application Performance

The techniques listed in Table 1 can be applied at a variety of layers of the OSI model. For example, it is natural to implement technologies such as FEC and QoS at the packet layer (Layer 3), while it is natural to implement congestion control and session optimization at the transport layer (Layer 4).

The two classes of optimization techniques that have the biggest impact on application performance, data reduction and the mitigation of round-trip delay, can be applied at multiple levels of the OSI model. Some of the existing approaches to improving application performance implement these techniques at the packet and session layers (Layer 5). However, a more innovative and impactful approach to closing the WAN intelligence gap is to understand how an application handles its data objects and protocol semantics, and then use this knowledge to apply these techniques at the application layer.

In order to understand the value of applying data reduction and the mitigation of round-trip delay at the application layer, it is important to realize that a fundamental characteristic of all networked applications is that they move data objects from one point to another, typically over a WAN. For example, in order for an ERP application such as SAP to present a screen to a user, a form has to be transmitted over a network. Analogously, when a collaboration portal opens a spreadsheet, a file is transmitted over a network.

Part of the advantage of performing data reduction techniques such as compression and differencing at the application layer is that these techniques can be applied directly to an entire application object. This approach is more efficient than if the same techniques are applied at Layer 3 on a packet-by-packet basis.

To understand how this works, assume that a user in a branch office is running an application that opens a large file that contains the company's current inventory. That file would be compressed and sent to the branch office for use by the employee and also stored in the branch office for future use. Further assume that at a later time another employee needs to access a current version of the file. By working at the application level, it is possible to compare the entire current version of the file with the entire version that was last sent to the branch office. As part of this comparison, the differences between the current version of the file and the version that was last sent to the branch office are identified. These differences are then compressed and sent to the branch office, where they are used to create a current version of the inventory file.

Understanding an application's data objects as well as the application's semantics enables an application optimization solution to recognize application data and anticipate its use from previous requests. This capability allows the solution to make requests of a distant server prior to those requests being made by the client (request prediction). In this way, the application optimization solution is in a position to respond locally when the client does make those requests (request spoofing).

The accuracy at which a solution can perform entire object differencing or object request prediction depends on the ability of the solution to identify an object. However, as shown in Figure 2, identifying an object is complex in part because there are multiple ways that an object can be referenced. As a result, the ability to successfully apply object differencing or object request prediction requires a deep understanding of the application objects so that the application objects can be

- identified, which requires an understanding of how the application is naming the object;
- framed, which requires determining the beginning and ending of the object as it is being transmitted;
- typed, which requires determining the type of object; e.g., open XML.

The use of entire object differencing and object request prediction can effectively mitigate round-trip delays and dramatically reduce the amount of data transmitted over the network. The result of application-level acceleration, as opposed to packet-level or session-level acceleration, is a significant increase in effective application throughput and a dramatic reduction in application response time. As mentioned above, accurate object identification and efficient differencing can only be accomplished with application-level knowledge.

One of the challenges that IT organizations face when evaluating the viability of deploying any application performance enhancement technique is to quantify in advance the impact that the technique will have on application throughput and application response time. While third-party tests are helpful, it is not possible for a third-party test to be set up in such a way as to reflect the environment found inside every organization. Hence, in order to quantify the impact of an application performance-enhancement technique, IT organizations need to test that technique in their environment. This can be done either by setting up a lab environment that mirrors the company's production environment or by implementing the technique on a segment of the company's network.

7.0 Closing the WAN Intelligence Gap

A vendor that is taking the lead relative to application-level acceleration is Certeon, Inc. This section of the white paper contains information provided by Certeon.

Certeon's S-Series™ Application Acceleration Appliances speed mission-critical applications and minimize network traffic, while maintaining end-to-end security. Certeon introduced its S-Series Application Acceleration Appliances in February 2006 and is the only vendor that addresses the impact of the WAN upon application performance at the application-data object level. The S-Series is uniquely able to perform application object-level differencing through its embedded Application Acceleration Blueprints™ without requiring any modification to the client, server or application software. An Application Acceleration Blueprint includes a description of the ways that a specific application identifies and encapsulates data objects that it transacts over the network. Having an Application Acceleration Blueprint improves the accuracy of matching prior data objects and thus significantly improves data reduction going over a WAN.

Equally important as reducing the amount of data to improve application performance in high-latency WAN environments is reducing the number of round-trip wait times to enable faster application responsiveness. Once again, Certeon's Application Acceleration Blueprints can perform a valuable service here. By incorporating logic describing how key message exchanges work for a particular application, a Blueprint, working in conjunction with the Certeon S-Series, can effectively eliminate round-trip wait times.

Certeon S-Series' appliances support Blueprints for generic HTTP applications as well as Microsoft Office and SharePoint and Oracle® E-Business Suite applications. For more information about Certeon's solutions for application acceleration over a WAN, go to **www.certeon.com**.

8.0 Summary and Call to Action

IT organizations have the goal of ensuring that the applications the company uses to run its business can be accessed securely and that the response time and availability of those applications are acceptable. In order to achieve that goal given the ongoing technology and business trends that were discussed in this white paper, IT organizations need to add application-level intelligence to the WAN.

In order to create a strategy for adding application-level intelligence to the WAN, there are a number of questions that IT organizations need to ask themselves about their environments. These questions include the following:

I. What are the major line of business applications that are running on your network currently? How will that change over the next two years?

II. Is your company migrating toward the use of a Web-based portal for collaboration? For the delivery of information?

III. Is your company migrating toward the use of a Web-based portal for its line of business applications?

IV. How many remote offices does your company have? Where are they located, and what are their bandwidth and latency characteristics?

V. How much use does your company make of SSL? How is that usage changing?

There are also a number of questions that IT organizations need to ask their potential vendors to ensure that their solutions can process traffic at the applications layer and hence effectively close the WAN intelligence gap. These questions include the following:

I. What techniques does the vendor have to perform acceleration at the true application level, e.g., SharePoint or SAP?

II. How are the vendor's acceleration techniques impacted by the steps that your company has taken to ensure security?

III. What type of response times can the vendor provide to your applications running in your environment?

IV. How does your vendor mitigate the effect of chattiness? Of round-trip times?

V. How does a vendor's product reduce the amount of data that traverses the WAN?

VI. In your environment, what improvement can the vendor make to the performance of your applications with application acceleration turned off? With it turned on?

Glossary

Compression: The role of compression is to reduce the size of a file of data prior to transmitting that file over a WAN.

Congestion Control: The goal of congestion control is to ensure that the sending device does not transmit more data than the network can accommodate. To achieve this goal, the TCP congestion control mechanisms are based on a parameter referred to as the congestion window. TCP has multiple mechanisms to determine the congestion window.

Differencing: The goal of differencing is to avoid sending an entire file from origin to destination. In particular, the goal of differencing is to send only the changes that have been made to the file since the last time it was sent.

Forward Error Correction (FEC): FEC is typically used at the physical layer (Layer 1) of the OSI stack. FEC can also be applied at the packet layer whereby an extra packet is transmitted for every n packets that are sent. This extra packet is used to recover from an error and hence avoid having to retransmit packets.

Increase the TCP Window Size: The TCP window size refers to the number of packets that can be sent without receiving an acknowledgment. The goal of increasing the TCP window size is to eliminate times when there is both data to be sent and WAN capacity to support that transmission, but the data is not sent due to the limitations of the window size.

Quality of Service (QoS): QoS refers to the ability of the network to use functionality such as queuing in order to provide preferential treatment to certain classes of traffic, such as voice traffic.

Request Prediction: By understanding the application semantics, it is often possible to anticipate a request that will be made in the near future. Making this request in advance of it being needed eliminates virtually all of the delay when the request is actually made.

Request Spoofing: This refers to situations in which a client makes a request of a distant server, but the request is responded to locally.