# Supporting Server Consolidation Takes More than WAFS

Ashton, Metzler & Associates

Leverage Technology & Talent for Success

October 2005

# 1. Introduction

A few years ago, the conventional wisdom was that branch offices were heading towards obsolescence. In most companies today, however, the vast majority of employees are found in branch offices, giving strong evidence to the fact that branch offices are gaining in importance.

At the same time that branch offices have shown themselves to be a key component of business, they have also been undergoing significant changes from an IT perspective. For example, the vast majority of companies have either already begun the process, or are considering implementing a process of consolidating servers out of branch offices and into one or more centralized data centers. Server consolidation is driven by tactical factors, such as the desire to reduce cost and complexity, as well as by strategic factors, such as the need to comply with governmental regulations.

While consolidating servers makes sense from a business perspective, it introduces two types of issues that the network organization must resolve. One issue that has received a lot of attention is that many applications are written in a way that they perform well when they are run over a Local Area Network (LAN). However, these same applications often perform poorly when they are run over the Wide Area Network (WAN) that connects the company's branch offices to the corporate data center. This is particularly true of applications such as Microsoft Word, Excel or PowerPoint that use Common Internet File System (CIFS), which is a notably chatty protocol. This issue will be referred to in this white paper as The Applications Related Issue.

One of the interesting aspects of The Applications Related Issue is that it is artificial and potentially transitory. The Applications Related Issue is artificial in that it is the result of using the CIFS protocol to solve a problem that it was not designed to solve – file access over a WAN. The Applications Related Issue is potentially transitory because Microsoft has recently issued its next Windows Server release. According to industry sources, this software, which is designated 'R2', will eliminate many of the file access problems.

The second type of issue that results from server consolidation has not received anything close to the same level of attention. This issue will be referred to in this paper as The Network Related Issue. The Network Related Issue refers to the fact that independent of the chatty nature of CIFS, the applications that use CIFS require the WAN to provide them with the same level of traffic management that it offers to other applications such as Voice over IP (VoIP) and SAP.

Unlike The Applications Related Issue, The Network Related Issue is neither artificial nor transitory. The Network Related issue is a result of the requirement to ensure the performance of a company's key applications. This requirement is not going away and is particularly acute on branch office networks that tend to be comprised of relatively low bandwidth links.

Given its momentum in the marketplace, virtually all network organizations must create a plan to support server consolidation. The goal of this paper is to expose the reader to the key application and network issues that result from server consolidation. The intent is that this exposure will position the reader to create a plan to resolve these issues in a manner that is effective today and which also accommodates known technology evolution.

## 2. The Evolving Branch Office

Figure 1 depicts the approach that many companies took as they evolved the IT infrastructure within their branch offices. In particular, in an effort to improve the performance of the IT infrastructure, most companies distributed servers to branch offices. In many cases servers are deployed in branch offices in an ad hoc fashion and usually without any ability for centralized management or monitoring. It is also common for companies to deploy servers that do not conform to company standards.

**Distributed Servers**
**Figure 1**

Companies that implement the server architecture that is depicted in Figure 1 incur significant costs. For example, there is considerable cost associated with the servers themselves. In addition to the cost of the servers, there is the cost of the licenses for the software that is running on the servers, the cost of the real estate, as well as the cost of administering and maintaining the servers. Note that these costs still exist if the servers are consolidated into a centralized data center. However, there is an economy of scale that is associated with a centralized data center that is not present in individual branches.

Companies that implement the server architecture that is depicted in Figure 1 also incur a level of complexity that makes it very difficult for them to have effective control over the IT infrastructure. As a result of this lack of control, it is extremely difficult for companies with distributed servers to implement effective security or business continuity procedures. In addition, the last few years has seen a significant increase in terms of the governmental regulations to which companies must comply. These regulations include:

The Sarbanes-Oxley Act
This act requires management to make a written assertion stating their responsibility for establishing and maintaining an adequate control structure and procedures for financial reporting.

HIPAA
HIPAA (the Health Insurance Portability and Accountability Act) requires companies in the health care industry to provide administrative simplification, security and privacy.

The Gramm-Leach-Bliley Act
This act requires companies to give consumers privacy notices that explain the institution's information-sharing practices and to give consumers the right to limit some of the sharing of its information.

Each of these acts requires that companies put 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. Consolidating the servers into centralized data centers makes these tasks notably easier to accomplish.

In order to both reduce cost and complexity, many companies are moving to the architecture depicted in Figure 2. In this architecture, most if not all servers are consolidated into one or more centralized data centers.

**Centralized Servers**
**Figure 2**

Microsoft is an example of a company that has recently gone through an initiative to consolidate servers and data centers [1]. The goal of this initiative was to improve the operational efficiency of the IT organization while simultaneously reduce the Total Cost of Ownership (TCO). According to Microsoft, this initiative resulted in a 40% reduction in spending and a 25% increase in Service Level Agreement (SLA) performance.

## 3. The Applications Related Issue

As previously mentioned, many companies currently regard their branch offices as a key business resource. As a result, most companies have made an increasing number of applications available to employees at branch offices.

Kubernan[2] recently surveyed several hundred IT professionals to determine which applications were either currently running in their company's branch offices or would be within a year. The responses to that survey are contained in Table 1.

| Application | Percentage Expected |
|---|---|
| Email | 99% |
| Backup/Storage | 87% |
| Voice over IP | 79% |
| CRM | 63% |
| Real Time Video | 56% |
| Enterprise Resource Planning | 54% |
| Sales Force Automation | 52% |
| Citrix | 50% |
| On-Demand Video Streaming | 49% |
| Supply Chain Management | 42% |

Source: Kubernan

**Anticipated Branch Office Applications**
**Table 1**

As Table 1 demonstrates, there is a wide variety of applications that branch office employees need to access. However, only one of the applications listed in Table 1 (email) typically relies on the CIFS protocol. In particular, the most common use of CIFS is made by Microsoft applications, such as Microsoft Outlook, Word and

[1] Servers and Data Center Consolidation: Microsoft IT Enhances Cost Savings, Availability, and Performance, June 2004
[2] Kubernan is an analyst and consulting joint venture of Steve Taylor and Jim Metzler

PowerPoint.  Applications such as VoIP, enterprise resource planning or customer relationship management typically do not use CIFS.

CIFS works by sending packets from the client to the server in order to request some kind of service, such as opening, closing or reading a file.  The server processes these packets and checks to see if the client has the appropriate file permissions.  If the client has the proper permissions, the server then executes the request and sends one or more packets back to the client.

When run over a LAN, these service request packets introduce negligible latency. However, as was mentioned in the introduction to this white paper, companies that are consolidating their servers into a centralized data center often end up running CIFS over the WAN.  When run over the WAN, these service request packets add latency that is potentially noticeable to the end user.

A more important factor that influences the user's experience is that CIFS decomposes all files into smaller blocks prior to transmitting them.  For example, assume that a client was attempting to open up a two-megabyte file on a remote server.  CIFS would decompose that file into tens, or possibly hundreds of small data blocks.  The server sends each of these data blocks to the client where it is verified and an acknowledgement is sent back to the server.   The server must wait for an acknowledgement prior to sending the next data block.  As a result, it can take several seconds for the user to be able to open up the file.

Because each data block must be acknowledged prior to sending the next data block, the CIFS protocol is reminiscent of the bisync protocol that was introduced by IBM in 1964.  While this type of protocol can work effectively in a LAN environment, it introduces unacceptable delay when run over a WAN.  As a result, no modern WAN protocol requires that each data block be acknowledged prior to transmitting the subsequent data block.

Given both the severity and the prevalence of the performance problems created by running CIFS over the WAN, the IT industry has developed a solution that is referred to as Wide Area File Services (WAFS).  The goal of a WAFS solution is to make CIFS run as well over a WAN as it does over a LAN.

One of the limitations of a typical WAFS solution is that it only addresses The Applications Related Issue and not The Network Related Issue.  As a result, if a company deploys a typical WAFS solution and does not also implement traffic management, it is highly likely that branch office employees who attempt to access a file in a centralized site will still experience unacceptable performance. This concept will be demonstrated later in this white paper.

As mentioned, Microsoft recently issued R2, the new Windows Server release. Microsoft has stated [3] that R2 represents the first in a wave of upcoming branch office technologies from Microsoft and industry partners.

Some of the relevant features that will be included in R2 include:

File Differencing
R2 includes Remote Differential Compression (RDC). RDC is a compression technology that replicates only the changes that are needed in order to ensure global file consistency.

Robust File Replication
The purpose of a Distributed File System (DFS) is to unite files on different computers into a single name space. R2 will contain a completely rewritten replication engine for the DFS. The new DFS is notably more robust and scalable. The new DFS utilizes RDC to increase the efficiency of the WAN. In addition, if a WAN connection were to fail, data can be stored and forwarded when the WAN becomes available.

R2 will also include some enhanced management tools such as:

- The Microsoft Management Console has been expanded to include an enterprise-wide administration framework for managing file and print services

- The enhanced DFS Namespaces user interface allows for easier management of file system roots within a network infrastructure, presenting shared folders to users as a grouping called a *Namespace*

- The replication of branch office data to a central server can now be automated

R2 has some significant advantages versus a traditional WAFS solution. One of these advantages is that unlike the traditional WAFS solutions, deploying R2 does not require that IT organizations deploy an additional appliance in each branch office. In addition, in many cases Microsoft users will be able to acquire R2 for free and hence completely avoid the cost of acquiring an appliance.

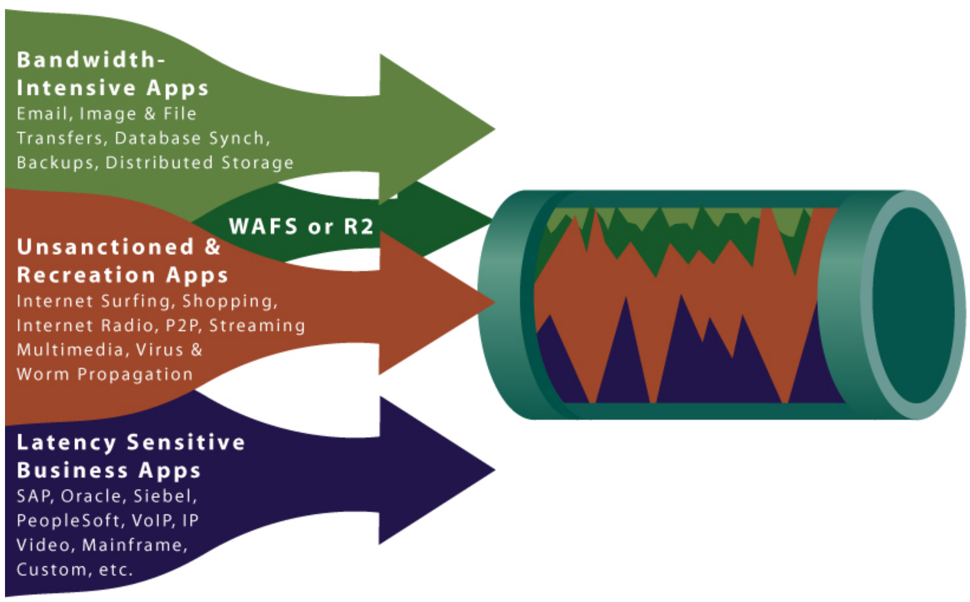
R2 also provides a higher level of security than do most WAFS solutions. For example, in the current versions of CIFS there is not any effective internal security for the session once the user has authenticated. As a result, when companies run CIFS over the WAN this traffic is vulnerable to a man-in-the-middle intercept.

[3] Microsoft Windows Server 2003 R2 Beta 2, Reviewers Guide, May 2005

## 4. The Network Related Issue

As previously mentioned, one of the limitations of a typical WAFS solution is that it only addresses The Applications Related Issue and not The Network Related Issue. In order to successfully support server consolidations, network organizations need a solution that addresses both issues.

It is important to realize that The Network Related Issue is not unique to CIFS traffic. Traffic management is required in any situation (Figure 3) in which bandwidth is scarce and there are one or more delay sensitive, business critical applications.




**Variety of applications battle for bandwidth**
**Figure 3**

One example of a latency-sensitive, business-critical application is VoIP. Over the last few years the majority of companies have made at least some deployment of VoIP. One of the features that distinguish VoIP from a more typical data application is the rigorous demands that voice places on the underlying IP network. For example, the ITU (International Telecommunication Union) recommends that the end-to-end delay associated with a voice call not exceed 150 ms. Experience has shown that it is possible to exceed that goal by a small amount; however, if the delay becomes too large, the quality of the voice call degrades noticeably.

Another example of a latency-sensitive, business-critical application is SAP. Several of the SAP modules are notably delay sensitive. An example of this is the Sales and Distribution (SD) module of SAP that is used for sales order entry. If the SD component is running slowly a company can compute the lost productivity of the company's sales organization as they waste time waiting for the SD module to respond. In addition, if the SD module times out, this can irritate the customer to the point where they hang up, taking their business elsewhere.

In order to successfully support myriad applications on a branch office network, network organizations need to implement traffic management. The focus of the organization's traffic management processes must be the company's applications, and not just the megabytes of traffic traversing the network. Some of the key steps in a traffic management process include:

- Discovering the Application

  Application discovery has to occur at Layer 7. In particular, information gathered at Layer 4 or lower allows a network manager to assign a lower priority to their Web traffic than to other WAN traffic. However, without information gathered at Layer 7, a network manager is not able manage the company's application to the degree that they could perform a task such as assigning a higher priority to some Web traffic than to other Web traffic.

- Profiling the Application

  Once the application has been discovered, it is necessary to identify how long the application traffic stays in various portions of the network. It is also necessary to determine the bandwidth requirement for that application and how it varies by location and by user.

- Quantifying the Impact of the Application

  As is described in detail in the next section of this white paper, since many applications share that same WAN circuit, these applications will tend to interfere with each other. In this step of the process the degree to which a given application interferes with other applications is identified.

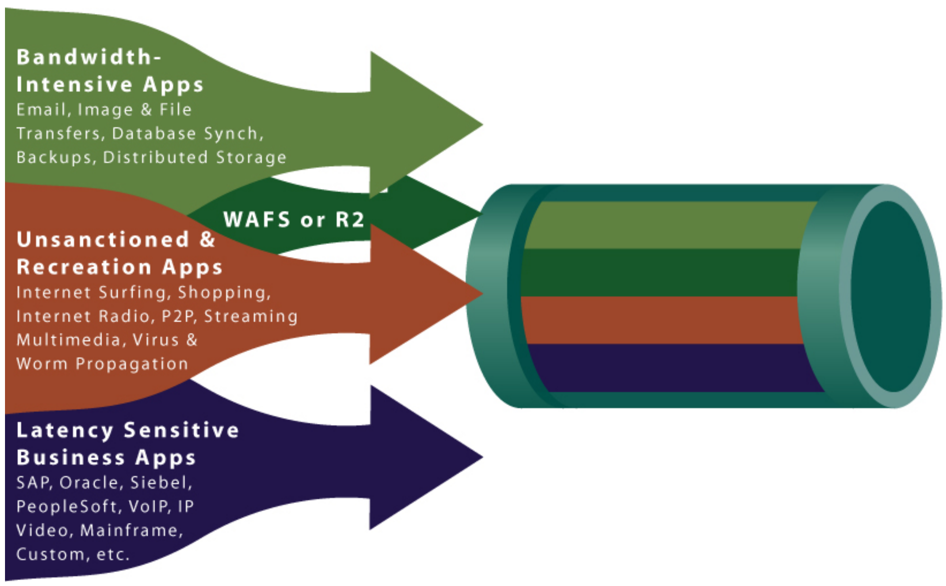
- Assigning Appropriate Bandwidth

  Once the bandwidth requirements have been determined and the degree to which a given application interferes with other applications is identified, it is now possible to assign bandwidth to an application. In some cases, this will be done in a way to ensure that the application performs well. In other cases, this will be done primarily to ensure that the application does not interfere with the performance of other applications.

- Ongoing Monitoring of the Network

Applications and networks are dynamic in nature. Because of that, network organizations need to continually monitor the network to identify significant changes such as the change in usage of an existing application or the usage, either sanctioned or otherwise, of a new application.

The desired outcome of implementing traffic management is that myriad application coexist and application SLAs are met (Figure 4).




**Impact of Traffic Management**
**Figure 4**

## 5. A Network Assessment for Server Consolidation

This section of the white paper demonstrates what happens to the performance of those applications if a company were to run CIFS traffic on a branch office network and not implement traffic management.

In order to accomplish this goal, this section will analyze a hypothetical company that will be referred to as Acme. Acme has consolidated its servers into a central data center and has deployed a WAFS appliance in each branch in order to improve file access for Microsoft Office applications. There is a T1 link into each of Acme's branch offices, and each of these branch offices is running the same set of applications.

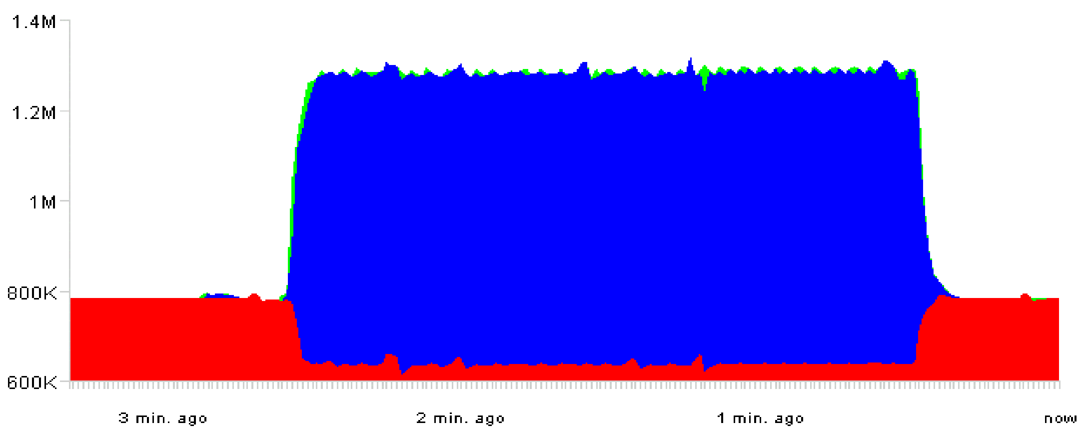
For the sake of simplicity, this analysis will only focus on two applications at a time, even though most companies run tens of applications over their branch office networks. In addition, it is important to note that even though this analysis will focus on only one of Acme's branch offices, the results of this analysis would apply to all of Acme's branch offices.

This analysis will present three scenarios. In each scenario there is CIFS traffic that is generated by a user in a branch office initiating a file replication. This CIFS traffic shares the T1 circuit into Acme's branch offices with several other applications and Acme has not applied traffic management on this circuit. Each scenario will demonstrate that in the absence of traffic management, having CIFS traffic share a circuit with just one other application results in significantly degraded performance. That performance will degrade steadily as additional applications share the same circuit.

Scenario 1: CIFS Traffic Negatively Impacting a Videoconference

> In this scenario, the CIFS traffic shares the T1 circuit with a 3-way videoconference which is comprised of two data streams – each running at 384 Kbps. In Figure 5, the red represents the videoconferencing traffic and the blue represents the CIFS traffic.

As is shown in Figure 5, the videoconference is in progress and obtaining all of the bandwidth that it needs when the file replication is initiated. The file replication lasts for approximately two minutes during which time the quality of the videoconference is greatly degraded. This will most likely lead to user complaints.

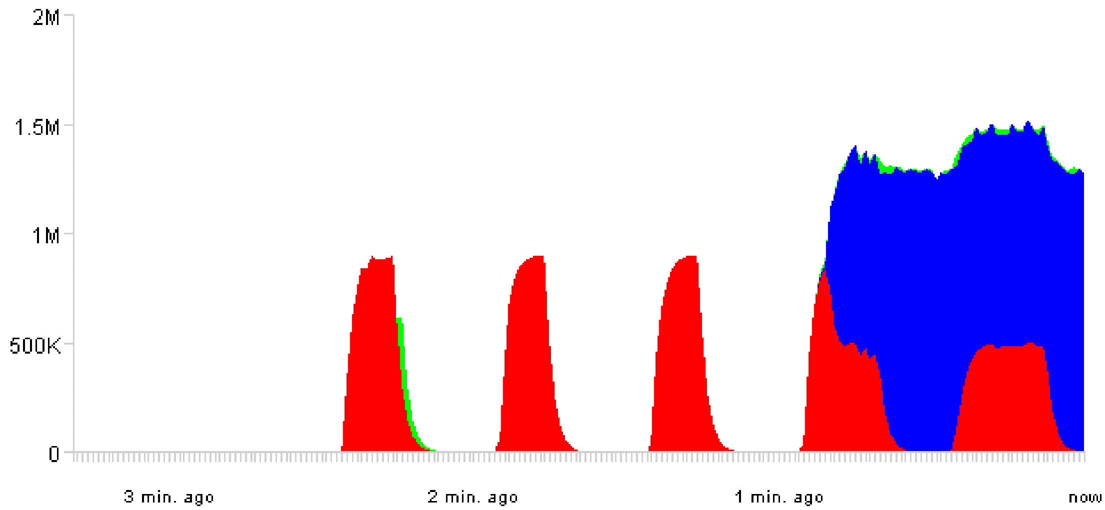



**CIFS Traffic and Videoconferencing**
**Figure 5**

Scenario 2: CIFS Traffic Negatively Impacting Business Transactions

> In this scenario, the CIFS traffic shares the T1 circuit with a business application such as sales order processing. In Figure 6, the red represents the traffic generated by the business transactions and the blue represents the CIFS traffic.

As is shown in Figure 6, in the absence of the CIFS traffic, each business transaction takes approximately ten seconds to complete. However, while the file replication is taking place, the business transactions take approximately twice as long to complete. This results in lost productivity for the employees that are entering the sales orders. This is also likely to result in complaints to Acme's IT organization from the sales organization.



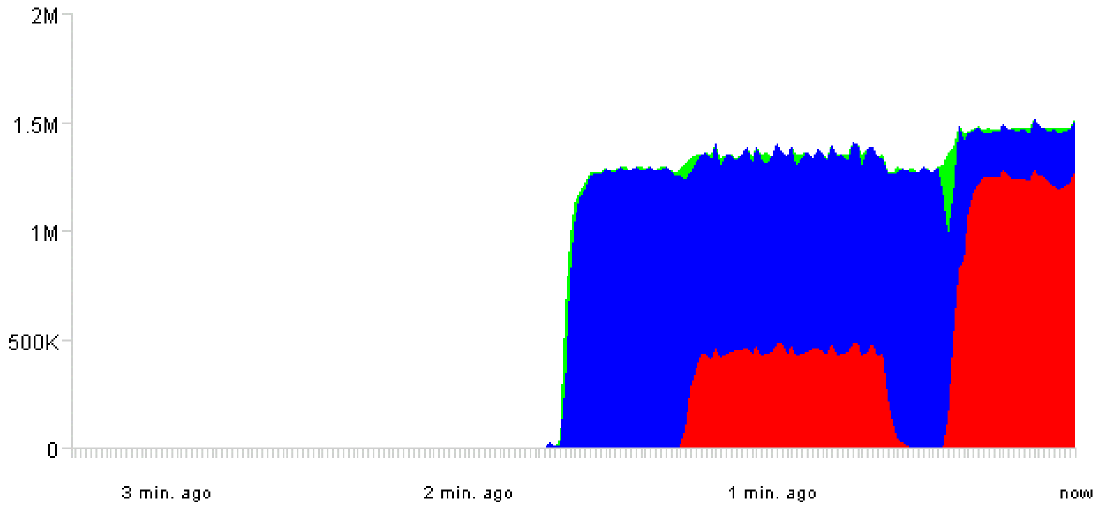

**CIFS Traffic and Business Transactions**
**Figure 6**

<u>Scenario 3: Web Traffic Negatively Impacting CIFS Traffic</u>

> In this scenario the CIFS traffic shares the T1 circuit with Web traffic. In Figure 7 the red represents the Web traffic and the blue represents the CIFS traffic.

As is shown in Figure 7, when there is no other traffic on the T1 circuit, the file replication consumes all of the available bandwidth. However, when Web traffic is also attempting to transit the T1 circuit, the bandwidth available to the file replication changes notably. The first instance of Web traffic corresponds to a single user access the Web. In this case, the bandwidth available to the file replication is reduced by 40%, and so the amount of time it takes to replicate the file is increased by approximately 65%. This type of increase may or may not result in complaints from frustrated users.

The second instance of Web traffic corresponds to eight users accessing the Web. In this case, the bandwidth available to the file replication is reduced by 80%. As a result, it would take approximately five times as long to replicate the file. This type of increase would definitely result in complaints from frustrated users.




**CIFS and Web Traffic**
**Figure 7**

## 6. Summary

Driven both by the need to reduce cost and gain more control over the IT infrastructure, the majority of enterprises have begun to consolidate their servers into central data centers. However, the initial implementations of server consolidation demonstrated that applications such as Microsoft Office that rely on the CIFS protocol perform poorly in this type of environment.

To compensate for this poor performance, the IT industry has developed a solution that is referred to as WAFS. WAFS attempts to overcome some of the aspects of the CIFS protocol that make it inappropriate to run over a wide area network.

However, Microsoft is well aware of the limitations of CIFS. According to a variety of industry sources, Microsoft's R2 will eliminate many of the file access problems. IT professionals who need to support server consolidation initiatives need to closely follow the deployment of R2 to determine how much of the file access problem R2 actually solves.

However, whether an IT organization deploys a WAFS appliance or relies on R2, this only addresses a part of the file access problem. As was demonstrated in section 5 of this white paper, without traffic management functionality, CIFS traffic can adversely affect delay-sensitive, business-critical applications such as videoconferencing. Alternatively, without traffic management functionality, applications that are less timely and business critical can adversely impact CIFS traffic.

In order to successfully support CIFS traffic, IT organizations must deploy a traffic management process that focuses on application performance. The first step in the process is to be able to discover the applications. This requires that the network be monitored at Layer 7.

Other steps in the traffic management process include determining the bandwidth requirement for each application as well as how that application interferes with other applications that transit the same WAN circuit. This allows a network organization to assign the appropriate amount of bandwidth to each application.

The final step in the traffic management process is to continually monitor the network to identify significant changes such as the change in usage of an existing application or the usage, either sanctioned or otherwise, of a new application.