# Optimizing Performance For Your Branch Offices

Jim Metzler

## How do you choose the right technology for your WAN?

The goal of branch office optimization solutions is to improve the performance of applications delivered from the datacenter to the branch office or directly to the end user. If organizations can improve the performance of these applications enough, it might well be possible to centralize some or all of the existing branch office IT infrastructure; e.g., file servers, email servers, SMS servers and local tape backup equipment.

Myriad techniques comprise branch office optimization solutions. Table 1 lists some of these techniques and indicates how organizations can use each of these techniques to overcome some characteristic of the WAN that impairs application performance.

Some of the principal WAN optimization techniques include:

- **Caching**—This refers to keeping a local copy of information, with the goal of either avoiding or minimizing the number of times that information must be accessed from a remote site. If caching is done at the object level (e.g., file, Web object, or email attachment), there is always a risk that the local copy is not identical to the actual object being requested from the central server.
- **Compression**—The role of compression is to reduce the size of a file 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; aka, De-duplication**—The goal of differencing is to avoid sending an entire file or data stream from origin to destination. In particular, the goal of differencing is to send only the changes that have been made to the file or data stream 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 network layer (Layer 3), whereby an extra packet is transmitted for every *n* packets sent. This extra packet contains redundant data and is used to recover from an error and hence avoid having to retransmit packets.
- **Quality of Service (QOS)**—QOS refers to the ability of the network to use functionality such as queuing to provide preferential treatment to certain classes of traffic, such as voice. Some QOS solutions allocate bandwidth, while others prioritize the packets in the queue. Some solutions do both independently.
- **TCP Acceleration**—TCP acceleration involves a range of techniques, including simple steps such as increasing the TCP window size. The TCP window size refers to the number of packets that can be sent without receiving an acknowledgment. Other techniques include connection pooling, limited and fast re-transmits, and support for high-speed TCP. The goal of these techniques is to make TCP perform better in a wide range of high-latency environments.
- **HTTP Acceleration**—This involves techniques such as request predication, HTTP pipelining (whereby multiple HTTP requests are written out to a single socket without waiting for the corresponding responses), as well as the caching of

*Dr. Jim Metzler has been an application developer, marketing manager, product manager, engineering manager, IT manager, research director and consultant. Along with Steven Taylor, Jim is a founding member of Kubernan, an analyst firm. He can be reached at jim@Kubernan.com.*

**TABLE 1 Techniques To Improve Application Performance**

| WAN Characteristic | WAN Optimization Techniques |
|---|---|
| Insufficient Bandwidth | Data Reduction:<br>Data Compression<br>Differencing (aka, de-duplication)<br>Caching |
| High Latency | Protocol Acceleration:<br>TCP<br>HTTP<br>CIFS<br>NFS<br>MAPI<br><br>Mitigate Round-trip Time<br>Request Prediction<br>Response Spoofing |
| Packet Loss | Congestion Control<br><br>Forward Error Correction (FEC) |
| Network Contention | Quality of Service (QOS) |

static Web pages to improve the performance of HTTP.

■ **Request Prediction**—By understanding the semantics of specific protocols or applications, it is often possible to anticipate a request a user will make in the near future. Making this request in advance of it being needed eliminates virtually all the delay when the user actually makes the request.

Many applications or application protocols have a wide range of request types that reflect different user actions or use cases. It is important to understand what a vendor means when it says it has a certain application level optimization. For example, in the CIFS (Windows file sharing) protocol, the simplest interactions that can be optimized involve drag and drop. But many other interactions are more complex. Not all vendors support the entire range of CIFS optimizations.

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

In some cases, one of the techniques listed above might be the entire solution. For example, a company that ships large files between the U.S. and India could decide to put an appliance that only does compression on each end of that link.

In many cases, however, techniques like the ones listed above are combined into a broader-based solution. For example, companies that consolidate servers by moving them out of branch offices and into a centralized datacenter end up running CIFS over their WAN. Since CIFS is a "chatty" protocol, this can result in poor performance. To compensate, many vendors have deployed multi-function WAN optimization solutions. These solutions typically implement myriad technologies, such as CIFS and NFS (Network File System) acceleration, caching, compression, differencing, request prediction and spoofing, together with security functionality, such as encryption.

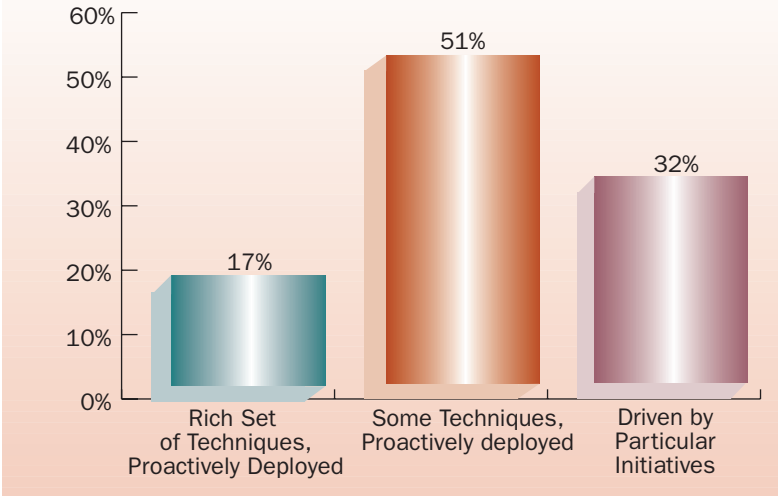
### Tactical vs. Strategic Solutions

To put the question of tactical vs. strategic in context, a company and its IT department must identify the problem they are trying to solve, answering questions such as: Is the problem just the performance of this one application as used just by employees in (for example) the Pac Rim? If that is the problem statement, then the company is looking for a very tactical solution.

However, the company might decide that the problem it wants to solve is how to guarantee the performance of *all* of its critical applications for *all* of its employees under as wide a range of circumstances as possible. In this case, the company needs a strategic solution.

Historically, branch office optimization solutions have been implemented in a tactical fashion. That means that companies have deployed the least amount of equipment possible to solve a specific problem. Kubernan recently asked several hundred IT professionals about the tactical vs. strategic nature of how they use these techniques. Their answers, which Figure 1 shows, indicate the deployment of these techniques is becoming a little more strategic.

**FIGURE 1 Approach To Network And Application Optimization**




One respondent whose perspective supports this position is a COO in the electronics industry. This executive noted that his company's initial deployment of network and application optimization techniques was to solve a particular problem. However, he also stated that his company is "absolutely becoming more proactive moving forward with deploying these techniques," indicating a shift to a more strategic view.

Similarly, a network architect for a motion picture company commented that his organization has been looking at these technologies for a number of years, but has only deployed products to solve some specific problems, such as moving extremely large files over long distances. He noted that his organization now wants to deploy products proactively to solve a broader range of issues relative to application performance. "Even a well written application does not run well over long distances," the Motion Picture Architect said. "In order to run well, the application needs to be very thin, and it is very difficult to write a full featured application that is very thin."

### Current Deployments

Table 2 (p. 54) depicts the deployment status of some of the primary branch office optimization techniques.

The data in Table 2 show that IT organizations plan to deploy a wide range of branch office optimization techniques.

A Team Leader who was part of the survey base pointed out that his company has historically made significant use of satellite links and, as a result, they have been deploying some form of optimization appliance for about 10 years. One of

**It's important to understand how cost changes as the solution scales**

the primary differences he sees in the current generation of optimization appliances is that they provide a broader range of functionality than previous generations of these appliances did.

An Engineering CIO stated that his organization originally deployed a WAFS solution to alleviate redundant file copies. He said he has been pleasantly surprised by the additional benefits of using the solution. His organization also plans on doing more backup of files over the network, and he expects the WAFS solution they have already deployed will assist with this.

The points that the Engineering CIO raised go back to the previous discussion of a tactical vs. a strategic solution. We find that most IT organizations that deploy a network and application optimization solution do so tactically, and later expand the use of that solution to be more strategic.

Therefore, when choosing a network and application optimization solution, it is important to ensure that the solution can scale to provide additional functionality beyond what is initially required.

### Selection Criteria

Below is a set of criteria that IT organizations can use to select a branch office optimization solution.

■ **Performance**—Third party tests of a solution can be helpful. It is critical, however, to quantify the kind of performance gains that the solution will provide in the particular environment where it will be installed. As part of this quantification, it is important to identify if the performance degrades either if additional functionality within the solution is activated, or if the solution is deployed more broadly across the organization.

■ **Transparency**—It should be possible to deploy the solution and not have anything such as routing, security or QOS break. The solution should also be transparent relative to both the existing server configurations and the existing authentication, authorization and accounting (AAA) systems. In addition, the solution should not make troubleshooting any more difficult.

■ **Solution Architecture**—If the organization intends for the solution to be able to support additional optimization functionality over time, it is important to determine if the hardware and software architecture can support new functionality without an unacceptable loss of performance.

■ **OSI Layer**—Organizations can apply many of these techniques at various layers of the OSI model. They can apply compression, for example, at the packet layer. The advantage of applying compression at this layer is that it supports all transport protocols and all applications. The disadvantage is that it cannot directly address any issues that occur higher in the stack.

Alternatively, having an understanding of the semantics of the application means that compression can also be applied to the application; e.g., SAP or Oracle. Applying compression, or other techniques such as request prediction, in this manner has the potential to be more effective.

■ **Capability to Perform Application Monitoring**—Some solutions provide the ability to monitor the end-to-end response time of an *n*-tier application or the Mean Opinion Score for VOIP traffic. Ideally, these solutions also provide the capability to isolate the source of performance problems.

Alternatively, some solutions work well with third-party monitoring tools, meaning in part that the solution does not negate the ability of these third party tools to see critical data.

■ **Scalability**—One aspect of scalability is the size of the WAN link that can be terminated on the appliance. More important is how much throughput the box can actually support with the relevant and desired optimization functionality turned on.

Other aspects of scalability include how many simultaneous TCP connections the appliance can support as well as how many branches or users a vendor's complete solution can support.

Downward scalability is also important. Downward scalability refers to the ability of the vendor to offer cost-effective products for small branches or even individual laptops.

■ **Cost-effectiveness**—This criterion is related to scalability. In particular, it is important to understand what the initial solution costs. It is also important to understand how the cost of the solution changes as the scope and scale of the deployment increase.

■ **Application Sub-classification**—An application such as Citrix or SAP is composed of multi-

**TABLE 2 Deployment Of Branch Office Optimization Techniques**

| | No Plans to Deploy | Have Not Deployed, But Plan to Deploy | Deployed in Test Mode | Limited Production Deployment | Broadly Deployed |
|---|---|---|---|---|---|
| Compression | 27% | 22% | 9% | 25% | 17% |
| Caching | 31% | 19% | 11% | 19% | 20% |
| HTTP acceleration | 37% | 20% | 8% | 18% | 17% |
| TCP acceleration | 37% | 24% | 13% | 17% | 9% |
| Wide Area File Services (WAFS) | 51% | 23% | 9% | 11% | 5% |

ple modules with varying characteristics. Some branch office optimization solutions can classify at the individual module level, while others can only classify at the application level.

■ **Module vs. Application Optimization**—In line with the previous criterion, some branch office optimization solutions treat each module of an application in the same fashion. Other solutions treat modules based both on the criticality and characteristics of that module. For example, some solutions apply the same optimization techniques to all of SAP, while other solutions would apply different techniques to the individual SAP modules based on factors such as their business importance and latency sensitivity.

■ **Disk vs. RAM**—Advanced compression solutions can be either disk- or RAM-based. Disk-based systems typically can store as much as 1,000 times the volume of patterns in their dictionaries as compared with RAM-based systems, and those dictionaries can persist across power failures. The data, however, is slower to access than it would be with the typical RAM-based implementations, although the performance gains of a disk-based system are likely to more than compensate for this extra delay.

While disks are more cost-effective than a RAM-based solution on a per-byte basis, given the size of these systems, they do add to the overall cost and introduce additional points of failure. Standard techniques such as RAID can mitigate the risk associated with these points of failure.

■ **Protocol Support**—Some solutions are specifically designed to support a given protocol (e.g., UDP, TCP, HTTP, CIFS, MAPI) while other solutions support that protocol generically. In either case, the critical issue is how much of an improvement in the performance of that protocol the solution can cause in the type of environment in which the solution will be deployed.

It is also important to understand if the solution makes any modifications to the protocol that could cause unwanted side effects, such as breaking access control lists (ACLs).

■ **Security**—The solution must not break the current security environment, such as breaking firewall ACLs by hiding TCP header information. In addition, the solution itself must not create any additional security vulnerabilities.

■ **Ease of Deployment and Management**—This is crucial because few branch offices have local IT staff, and so it is important that unskilled personnel can install the solution. In addition, the greater the number of appliances deployed, the more important it is that they be easy to configure and manage.

It's also important to consider which other systems will have to be touched to implement the branch office optimization solution. Some solutions, especially cache-based or WAFS solutions, require that every file server be accessed during implementation.

■ **Change Management**—Since most networks experience periodic changes such as the addition of new sites or new applications, it is important that the branch office optimization solution can adapt to these changes easily. It is preferable if the solution can adjust to these changes automatically.

■ **Support of Meshed Traffic**—A number of factors are causing a shift in the flow of WAN traffic away from a simple hub-and-spoke pattern and to more of a meshed flow. If a company is making this transition, it is important that the branch office optimization solution can support meshed traffic flows and can support a range of features such as asymmetric routing.

■ **Support for Real-Time Traffic**—Also, many companies have deployed real-time applications. For these firms, it is important that the branch office solution can support real-time traffic.

Some real-time traffic like VOIP and live video can't be accelerated because they are already highly compressed. Header compression might be helpful for VOIP traffic, and most real-time traffic will benefit from QOS.

■ **Link by link vs. a Global Solution**—The distinction between these two approaches is the amount of automated and intelligent coordination that exists between the individual components of the solution.

In a link-by-link solution, the coordination between the component parts is manual. A global solution implements extensive automation, with real-time measurements of application performance communicated among the system components. The goal of the global approach is to enable the solution to make continuous, dynamic adjustments to the application's performance.

**Real-time traffic like VOIP offers its own challenges**

## Conclusion

As recently as a few years ago, IT organizations were not concerned about managing application delivery. Now this is top-of-mind for virtually all IT organizations. However, in spite of the importance of application delivery, few IT organizations perform this difficult function very well.

One of the major components of application delivery is the ability to manage the performance of applications delivered from the datacenter to the branch office or directly to the end user. Solutions to this problem, often referred to as branch office optimization solutions, are available from numerous vendors. Due to both their complexity and importance, when IT organizations evaluate these solutions, they should take a structured approach. One part of that approach is to use evaluation criteria such as the ones described in this article.

As mentioned, branch office optimization solutions are just one component of application delivery. Subsequent articles will detail other components, including management and control□