



**Missing Link**   
**Security Services**  
Mark Bouchard, Founder

## The Case and Criteria for Combining Application Acceleration and Security

### About the author

---

Mark Bouchard, CISSP, is the founder of Missing Link Security Services LLC, a consulting firm specializing in information security and risk management strategies. A former META Group analyst, Mark has assessed and projected the business and technology trends pertaining to a wide range of networking and information security topics for over 10 years. He is passionate about helping enterprises address their information security challenges. During his career he has assisted hundreds of organizations worldwide with strategic and tactical initiatives alike, from the development of multi-year strategies and overall architectures to the justification, selection, acquisition, implementation and operation of security and privacy solutions.



## Table of Contents

---

Introduction.....	3
The Need for Application Acceleration .....	3
The Changing Nature of “The Business” .....	3
The Changing Nature of Applications.....	4
Table 1: Drivers for Application Acceleration.....	4
The Need for Application Security .....	4
Backhauling Breaking Down.....	5
The Compliance Factor.....	5
Table 2: Drivers for (Branch Office) Application Security .....	6
The Benefits of a Combined Solution .....	6
Putting Organizational Issues and Politics Aside .....	7
The Criteria for a Combined Solution .....	7
Key Functional Criteria for Application Acceleration.....	7
Key Functional Criteria for Application Security .....	8
Bringing it All Together.....	8
Conclusion .....	9



## Introduction

Several high-impact business and technology trends are driving the need for today's organizations to find ways to enhance the performance and security of their applications. For instance:

- To remain competitive, IT must help streamline the business by simultaneously ensuring superior responsiveness of applications and agility of the underlying infrastructure while still managing to rein in costs;
- A burgeoning set of regulatory requirements and an ongoing shift in hacker motivation continue to raise the bar in terms of what it takes to achieve an "appropriate and acceptable level of security"; and,
- At the same time, the volume, nature, and origin of application traffic is undergoing dramatic changes.

This paper will first elaborate on these trends, further characterizing the various challenges driving the need for greater degrees of application performance and security. Subsequently, it will establish the benefits of having a single, combined product that addresses both sets of objectives – acceleration and security – before identifying the key criteria for selecting a corresponding solution.

## The Need for Application Acceleration

### The Changing Nature of "The Business"

The need to enhance application performance, especially over WAN connections, is due in large part to the changing nature of today's businesses. To begin with, the background pressure being applied by executive management to "accelerate the business" is nearly universal. Processes that are central to top-line growth, such as product development, supply chain management, and sales, must be facilitated in a manner that enables them to be accomplished faster. Consequently, the applications that enable these processes must go faster too.

Furthermore, the associated systems must be more agile. Sufficient spare capacity and flexibility must be present to better accommodate increasingly frequent and invariably time-sensitive changes to business direction, plans, and objectives. Today it may be all about manufacturing widgets and selling them via a direct sales force. But tomorrow it could be all about engaging 17 different business partners in a Web2.0-based collaborative effort to provide do-it-yourself kits for constructing space-station modules.

And let's not forget the bottom line either. Cost control is certainly a big issue for most companies. So anything that IT can do to enhance user productivity and optimize resource utilization will certainly be welcome too. Indeed, anticipated cost savings is one of the primary drivers behind another closely related trend, namely the re-location and consolidation of computing resources from distributed sites to centralized data centers. A second driver in this case is that centralization also helps reduce the level of effort required to establish compliance with prevailing regulations. In any event, from the perspective of application performance, the impact is that users located in remote offices must now access more of their applications over wide area connections. This approach, of course, is inherently slower than having access to them locally.

Compounding matters even further is the simultaneous de-centralization of users. This is occurring as organizations steadily embrace the concept of the extended enterprise and encourage greater degrees of user mobility. The motivation ties back to the desire to streamline the business both by enhancing user productivity and by facilitating the essential activities of key constituencies. The result is a steadily growing need to provide access and computing services to all types of users (e.g., employees, the general public, customers, and partners) from all types of locations (e.g., remote offices, home, public kiosks, partner sites). And the upshot is that whereas the impact of resource centralization might otherwise be considered manageable, because it only affects a relatively small percentage of an organization's users, that percentage is actually growing quite rapidly. In fact, current estimates are that from 50% to 90% of an organization's users are accessing their applications remotely, at least part of the time.

So what? Prevailing business conditions are forcing an increasing percentage of application sessions over organizations' WAN connections. But as any network engineer knows, even long-distance links have relatively reasonable latencies associated with them. Bandwidth requirements aside, the issue then is generally not the network. Rather the bigger problem is the inherent nature of many applications and their underlying protocols.



Simply put, they were not designed to operate beyond the LAN. Specifically, they exhibit a characteristic referred to as “chattiness” where it takes numerous back and forth exchanges between client and server to complete a single, user-level transaction. This characteristic is not a problem for very low latency environments such as LANs. But for most other scenarios, the cumulative effect will range from being mildly annoying to potentially rendering applications completely unusable.

## The Changing Nature of Applications

What we have so far then is that application performance is being adversely impacted by both the changing nature of “the business” as well as the inherent nature of legacy applications. As if these issues were not enough, however, organizations must also contend with the changing nature of applications.

A predictable outcome of the desire to accelerate the business is more automation of critical and/or time-consuming business processes. This means more applications and, consequently, more traffic on the organization’s network. Of course, growing traffic volumes are also being fueled by the proliferation of user-centric applications, many of which are non-essential – or worse. Rich internet applications and social networking sites (e.g., MySpace, Facebook, YouTube) not only generate a significant network load but also have the potential to expose organizations to various liabilities. Leakage of sensitive information and the risk of other types of security breaches are just two possibilities. In any event, the greater issue is the growing contention for a finite resource – WAN bandwidth – and the potential for the performance of more important applications to suffer due to the presence and detrimental behavior of those that are clearly less important.

Two other changes are also worthy to note. First, there is a steadily increasing reliance on web applications and, therefore, underlying web protocols. This is significant because these are some of the most egregious offenders in terms of being “chatty”. Secondly, applications are generally becoming more complex. The current and emerging generations of applications are far more likely to enable real-time interaction and incorporate rich content in the form of graphics, video, and even voice. The result is that bandwidth requirements are rising at the same time that sensitivity to latency is becoming a common condition.

Table 1: Drivers for Application Acceleration

<b>Business Factors</b>	<b>Application Factors</b>
<ul style="list-style-type: none"> <li>• There is a prevailing need to accelerate the business.</li> <li>• Computing resources are being centralized (e.g., data center consolidation).</li> <li>• Users are being de-centralized (e.g., due to mobility and globalization).</li> </ul>	<ul style="list-style-type: none"> <li>• Many legacy applications were not designed for the WAN.</li> <li>• Traffic volumes are steadily rising.</li> <li>• There is a continuing proliferation of chatty/inefficient web protocols.</li> <li>• New applications are complex and increasingly sensitive to latency.</li> </ul>

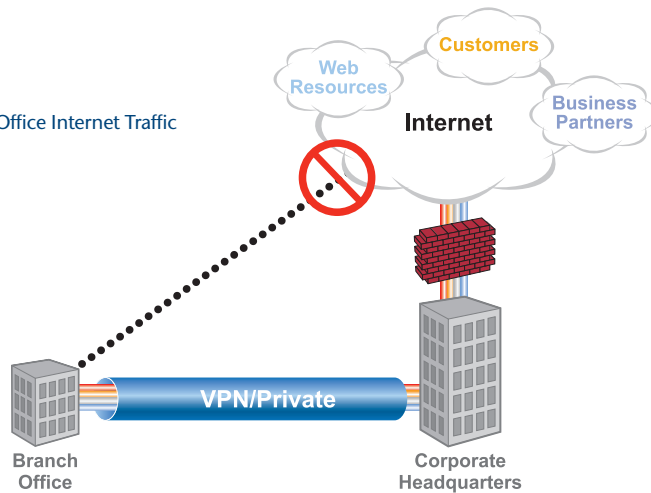
## The Need for Application Security

The need to improve application performance over the WAN is clearly not the only challenge facing today’s organizations. Security is another significant issue for practically everyone.

In this context, the primary location requiring attention is the remote/branch office. Larger, regional facilities and central sites typically have a core set of countermeasures that is already deployed. At a minimum these include firewalls, antivirus software, and intrusion prevention technology. Ideally they also include web security gateways, email protection systems, web application firewalls, and a full complement of server and desktop-based security tools. In contrast, most branch offices are only lightly secured and often rely on the countermeasures deployed at the central site for protection. For instance, branch office traffic that is destined to/from the Internet will be routed over the WAN connection between the two sites and will be secured by the countermeasures that comprise the central site’s Internet gateway. Depicted in figure 1, this configuration is commonly referred to as backhauling. It is motivated by the desire to reduce costs associated with security infrastructure and personnel, as well as to help ensure that policies are enforced consistently.



Figure 1: Backhauling Branch Office Internet Traffic



## Backhauling Breaking Down

Unfortunately, for the majority of organizations, the practice of backhauling is steadily breaking down. One minor problem in this regard is connection creep. This refers to the all-too-common scenario where semi-autonomous branch personnel take it upon themselves to establish local, direct connections to the Internet. This may help them meet their supposedly “unique needs”, including increased performance, but it inevitably puts the entire organization at greater risk.

A more significant shortcoming ties back to the changing nature of both businesses and applications. Not only is the amount of application traffic using web protocols and technologies on the rise, but a growing percentage of it is also destined to or originating from locations on the Internet. Of particular note is the rising use of software as a service, or SaaS. For smaller companies SaaS may simply entail the use of web mail or Google Apps. For others, Salesforce.com, WebEx, and managed security services are familiar examples. In addition, the slow yet steady adoption of Service Oriented Architectures further contributes to this trend. Web services and associated technologies drive the volume of Internet traffic by fostering a scenario where the individual functions and components that comprise an application can reside virtually anywhere and be owned by virtually anyone.

In any event, the point is that backhauling (a) unnecessarily subjects all of this traffic to a latency penalty, and (b) simultaneously puts it in contention with any other traffic traversing the WAN link, thereby jeopardizing the performance of the entire lot. In contrast, by deploying full-featured, centrally managed web security gateways in their branch offices, organizations could safely allow direct-to-Internet communications, and application performance would be improved across the board.

## The Compliance Factor

Regulatory requirements are clearly another significant factor driving the need for greater security. The Payment Card Industry Data Security Standard (PCI-DSS), Sarbanes-Oxley Act (SOX), and various disclosure laws (e.g., SB1386) strongly encourage enterprises to implement comprehensive information security programs and, in general, to follow associated best practices. Enhancing security at remote sites is a derivative implication, one that requires the consideration, if not implementation, of the following measures:

- Threat inspection and filtering capabilities, to help prevent local infections (e.g., viruses, worms, etc) from spreading to other connected locations;
- Encryption capabilities, to ensure the confidentiality of site-to-site communications (potentially even for types of WAN connections that are supposedly private); and,
- User-level control capabilities, to help implement the principle of least privileges and thereby reduce (a) the exposure of users to threats, and (b) the exposure of sensitive data by users.

Notably, as best practices these measures are appropriate even in the absence of any specific regulatory pressure. This is particularly true given the current state of the security landscape. Even greater levels of attack activity and threat sophistication have become status quo as hackers now focus on monetizing their efforts instead of just building reputations.



Speaking of increasing sophistication, the migration of threats ‘up the computing stack’, from the network layer to the application layer, is another relevant issue. This is a natural consequence of hackers wanting to evade currently deployed countermeasures, the vast majority of which are network-layer centric. Not surprisingly, the implication is that organizations need to implement countermeasures with visibility, inspection, and control capabilities above the network layer. In other words, they require tools that secure services, applications, data, and users, as opposed to just ports and protocols.

Table 2: Drivers for (Branch Office) Application Security

***Backhauling is breaking down***

***Achieving regulatory compliance requires adhering to best practices:***

- Better containment should be established.
- Inter-office communications should be secured.
- User access/activities should be granularly controlled.

***The threat landscape is evolving.***

## The Benefits of a Combined Solution

Thus far it has been established that today’s organizations have a substantial need for both application acceleration and application security. But how should they go about meeting these needs? One way would be to deploy multiple devices, with one or more used to fulfill each set of objectives. However, a far superior approach would be to deploy a single device that incorporates both sets of functionality (i.e., acceleration and security). Indeed, the benefits of such a combined solution include the following:

- **It eliminates potential conflicts.** For example, if acceleration functions are executed prior to security functions, then alteration of the traffic could compromise the ability to properly conduct inspections and enforce policies. Alternately, executing security controls first introduces the possibility of encryption completely negating the effectiveness of the acceleration mechanisms. It’s only when all of the functions are combined in an integrated manner that they can always be assured of being invoked in the best possible order.
- **It consolidates infrastructure.** Having fewer devices to deploy, operate, and maintain reduces complexity, capital expenditures, and operational effort.
- **It simplifies and unifies policy management.** Policies for both functions can be set in the same way and at the same time, promoting greater efficiency and further reducing the likelihood of conflicts or omissions.
- **It eliminates the latency penalty associated with duplicative processing.** Packet handling, even at the network and transport layers, takes time. The need to conduct higher-layer operations, such as L7 protocol parsing, session reconstruction, and application-specific inspections and optimizations, introduces considerably more delay. So why make matters even worse by requiring much of this processing to be done multiple times?
- **It enables acceleration of encrypted traffic.** The average amount of inter-office traffic that is encrypted ranges from 30% for typical organizations to as high as 75-100% for those in regulated industries. And these percentages are trending upward. A solution that lacks the ability to accelerate encrypted traffic, therefore, will ultimately not be much of a solution!
- **It compensates for the latency introduced by inspection and filtering processes.** Simply put, security inspections and policy enforcement take time to accomplish. Running acceleration techniques in conjunction with them should more than offset the associated latency penalty, ensuring that peace and happiness is maintained throughout user land.
- **It ensures that “junk” traffic is not needlessly accelerated.** By applying threat inspection and granular user-level controls first, inappropriate/unnecessary/unsafe traffic can be eliminated from the WAN pipe altogether. Load on the acceleration engine(s) will also be reduced. In this way, organizations can effectively “stop the bad, while accelerating just the good.”



Admittedly, the latter two items could also be achieved with a multi-device approach. However, a combined solution still has the advantage of reducing the effort required to coordinate policies and associated configuration settings.

## Putting Organizational Issues and Politics Aside

For some enterprises, one potential obstacle to using a combined solution is their organizational dynamics. Networking objectives are handled by the networking team and security objectives are designed, funded, and managed by the security team. Basic coordination between the teams is persistently a “work in progress”. Thus, full-blown collaboration and combination of initiatives is even more unlikely. In this case, however, there are incentives to try to overcome these predispositions.

For networking professionals, the primary advantage of a combined solution is that it maximizes the performance gains that can be achieved. In particular, it ensures that encrypted traffic can still be accelerated and it eliminates the need for duplicative processing. As a result, an application acceleration project that otherwise might only be marginally effective can now be turned into an unmitigated success. Of course, the combined solution also scores points based on its ability to simultaneously reduce network complexity, as well as capital and operational expenditures.

The benefit for security professionals, interestingly, is essentially the same; although the motivation is somewhat different. Ensuring that performance gains are maximized reduces the potential that latency introduced by essential security functionality will be blamed for impeding, or “decelerating”, the business.

## The Criteria for a Combined Solution

The case for a solution that combines acceleration and security is certainly compelling. However, making a great deal of sense is really only half of the equation. To be feasible, such an approach must be supported by the availability of a product that effectively addresses both sets of objectives.

### Key Functional Criteria for Application Acceleration

Suitability in terms of application acceleration requires that a solution (a) provides coverage for a wide range of applications, and (b) incorporates multiple techniques for optimizing application traffic.

At a minimum, support should be included for those applications that exhibit one or more of the following characteristics: they are critical to the business, they represent a significant percentage of network traffic, or they clearly suffer from poor performance. In this regard, web applications and associated technologies are not only notoriously poor performers but are also becoming increasingly prevalent. In addition, so-called “killer applications” for branch office users include file handling and email. Beyond this core set though, employees ultimately access a wide variety of applications. These range from home-grown utilities and complex business software to increasingly latency sensitive and bandwidth hungry applications such as VoIP, video, and interactive, multimedia solutions – all of which could benefit from measures to improve performance.

As for traffic optimization, it is important to recognize that the goal is actually twofold. In addition to reducing latency, it also makes sense to reduce bandwidth consumption. This could further improve application performance and, at the very least, will help control the recurring monthly fees owed to the organization’s network service providers. In general, optimization techniques can be broken into three categories.

**Compression and Caching Techniques.** These reduce the amount of data transiting the connection in the first place, freeing up bandwidth and alleviating any congestion that may be occurring. Basic, common compression algorithms tend to have limited applicability (e.g., GZIP for files and HTTP). Achieving broader coverage and significantly greater gains depends on recognizing, caching, and sending short labels for repeated sequences of all types of IP traffic – a technique that is often referred to as byte-level caching. The sequences can be maintained either in memory (for up to 10x reduction) or on disk (for up to 100x reduction). The difference from traditional caching technology is that the labels correspond to blocks of data, as opposed to individual objects or entire files. If a file is modified, then all that needs to be re-sent are the blocks of data that were changed.

**Acceleration Techniques.** These reduce the impact of poor application behavior, such as the “chattiness” that results in a single transaction requiring multiple back-and-forth exchanges over the connection. Basic capability in this area involves enhancing the operation of TCP, for example by adjusting receive-side windows



to more efficiently use available bandwidth and implementing forward-error correction to help reduce the inherent chattiness of this protocol. However, it is essential to also include protocol and application-specific optimizations, at least for the most common or egregious services, such as web/HTTP, email/MAPI, and file services/CIFS. In general, this is accomplished via object caching and finding ways to reduce the number of back-and-forth exchanges normally executed by these services.

**Control Techniques.** These effectively reduce end-to-end latency. Basic capability in this area involves traditional quality-of-service and bandwidth management mechanisms to help delay-sensitive traffic get through the network quicker. In general, this type of traffic prioritization will have the greatest benefit when bandwidth is constrained or when there are devices along the end-to-end path that are over-loaded, perhaps due to traffic spikes or over-subscription conditions. A second control technique is path optimization. This involves dynamically routing sessions to alternate data centers, or even just along secondary portions of the end-to-end path, based on real-time evaluation of operating conditions. Finally, another potential option is the ability to pre-position popular content at the remote end.

Significantly, an ideal solution should also include an inherent level of intelligence. This is necessary to ensure that specific optimization techniques are only used if they will truly provide a benefit for the given type of traffic being processed and the given network configuration. For instance, in some scenarios engaging basic compression algorithms can incur a latency penalty that is greater than the benefit that will be derived.

## Key Functional Criteria for Application Security

The requirements pertaining to application security are similar to those for application acceleration in one very important way: comprehensiveness is the key to success. The point here is that a full set of functionality must be provided both to secure inter-office communications and to safely enable direct-to-Internet connections (in lieu of backhauling). Specific security capabilities for enterprises to look for include the following:

- **Granular user and application control.** This entails being able to authenticate users and subsequently being able to control their activities, including blocking them, based on role-based authorization. Furthermore control should be to the level of individual applications (e.g., IM, P2P, and streaming media) and not just the ports and protocols being utilized.
- **Advanced URL filtering.** Being able to control access to web sites based on a periodically updated URL database should ideally be augmented by the ability to conduct real-time classification of new sites.
- **Advanced threat filtering.** Multi-protocol protection should be provided for viruses and worms, as well as other types of threats (e.g., spyware). In addition, screening should be accomplished not only for known threats using signatures, but also for unknown ones using heuristic and behavior-based detection mechanisms.
- **SSL inspection and encryption.** Based on configured policies, it should be possible to decrypt selected traffic for the purposes of inspection and optimization, and then re-encrypt it before sending it on its way. Similarly, it should be possible to selectively encrypt inter-office traffic streams on a per-session or per-application basis to enable greater assurance of confidentiality when appropriate.
- **Proxy architecture.** This is not a strict necessity, per se. However, products based on full proxies do have a significant advantage over those with other types of architectures. Specifically, proxies excel in terms of both the depth of inspections that can be performed and the granularity of associated filtering and blocking capabilities.

## Bringing it All Together

Having a full set of both application acceleration and security capabilities is absolutely necessary, but not quite sufficient. Many of the aforementioned benefits of a combined solution will not be realized if the individual capabilities are merely co-located on a single device. In this regard, additional essential characteristics that should be exhibited included the following:

- **Management that is centralized, unified, and simplified.** Respectively, this entails the ability to manage multiple devices at once, the presence of a single management system that covers all of the incorporated capabilities, and having a high degree of intuitiveness and ease of use that is pervasive across the full set of



lifecycle management functions (i.e., configuration, monitoring, troubleshooting, and reporting).

- **True integration.** If individual capabilities are co-located but not actually integrated, then previously cited processing and management efficiencies will be greatly diminished.
- **Transparency and compatibility.** In general, the security and acceleration operations conducted on communications traffic should be transparent. In other words they should not change the traffic (e.g., the packet headers) to the point that other network and systems management tools are rendered ineffective. In terms of compatibility, there should be multiple options for key services, such as user authentication, authorization, and logging, as well as network interfaces. This way organizations can leverage their existing infrastructure solutions and avoid having to implement an array of new systems or technologies.

Finally, it should be clear from this discussion that the source of a combined application acceleration and security solution is also important. Selecting a vendor with long-standing experience and a proven track record in both disciplines should help ensure that component capabilities are each best-in-class and that they have been brought together in the best manner possible.

## Conclusion

---

Prevailing trends entailing the centralization of computing resources, de-centralization of users, complexity of applications, and need for greater levels of information protection leave today's organizations with little choice. IT must find ways to enhance the performance and security of its applications, particularly for users located in remote/branch offices. In this regard, a single product that addresses both sets of objectives has several advantages over the alternative of deploying multiple, single-purposes devices. Not only does a combined solution help reduce costs and network complexity, but it also ensures that performance and security gains are maximized (e.g., by being able to accelerate encrypted traffic and eliminate duplicative processing).