# Traditional Application & Service Delivery Challenges

Ashton, Metzler & Associates

Leverage Technology & Talent for Success

## Introduction

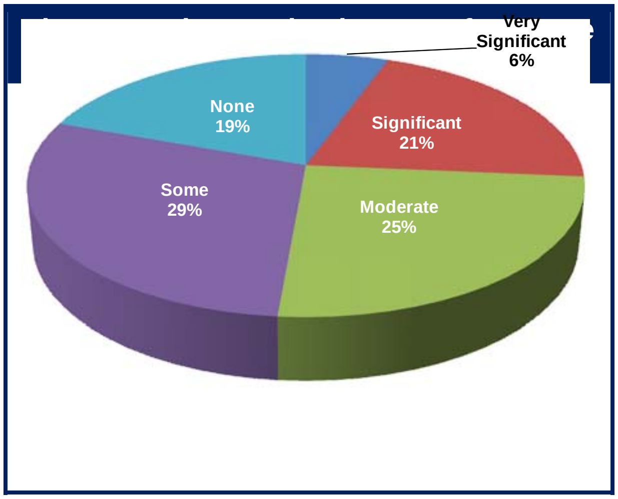
The goal of this document is to describe some of the traditional (a.k.a., first generation) challenges that make it difficult for IT organizations to ensure acceptable application and service delivery.

In 2012 a survey was given to the subscribers of Webtorials. Throughout this document, the IT professionals who responded to the surveys will be referred to as The Survey Respondents.

## Limited Focus of Application Development

The Survey Respondents were asked "When your IT organization is in the process of either developing or acquiring an application, how much attention does it pay to how well that application will perform over the WAN?" Their answers are shown in Figure 1.

As is often the case with surveys, the data in Figure 1 presents a classic good news – bad news situation. The good news is that the data in Figure 1 indicates that just over a quarter of IT organizations place a significant or very significant emphasis on how an application performs over the WAN during application development or acquisition. The bad news is that almost three quarters of IT organizations don't.



***The vast majority of IT organizations don't have any insight into the performance of an application until after the application is fully developed and deployed.***

## Network Latency

Network latency refers to the time it takes for data to go from the sender to the receiver and back. Since the speed of data flow is basically constant[1], WAN latency is directly proportional to the distance between the sender and the receiver. Table 1 contains representative values for network latency; both for a LAN as well as for a private WAN[2].

[1] There are slight variations in the speed of data flow in copper vs. the speed of date flow in fiber optics.

[2] The phrase *private WAN* refers to services such as Frame Relay and MPLS that are intended primarily to interconnect the sites within a given enterprise.

Table 1: Network Latency Values

| Network Type | Typical Latency |
|---|---:|
| LAN | 5 ms |
| East coast of the US to the West coast of the US | 80 ms – 100 ms |
| International WAN Link | 100 ms – 450 ms |
| Satellite Link | Over 500 ms |

As described by Moore's Law of Internet Latency[3], Internet latency is typically greater than the latency in a private WAN. That law references the business model used by the Internet and it states, “As long as Internet users do not pay for the absolute (integrated over time) amount of data bandwidth which they consume (bytes per month), Internet service quality (latency) will continue to be variable and often poor.”

## Availability

Despite the Internet's original intent to provide communication even during a catastrophic event, application availability over the Internet is somewhat problematic. The Internet is not a single network, but rather millions of networks interconnected to appear as a single network. The individual networks that compose the Internet exchange information between each other that describes what IP address ranges they contain (a.k.a., routes). Within a single network - called a routing domain - a specialized networking protocol is used to communicate IP address ranges to all the routers within the individual network. Routing protocols within a network can detect a network link failure and update the routing table on all routers within a few seconds when properly designed. For the exchange of information between networks - called inter-domain routing - a special routing protocol, the Border Gateway Protocol (BGP), is used. The size and complexity of the Internet as well as the inherent characteristics of BGP mean that a failed network link and the resulting routing path change may take several minutes before all routing tables are updated. In contrast, traditional voice circuits take milliseconds to reroute voice calls when a network link fails.

The impact of a network link failure and the time it takes for the Internet to update its routing table and to find an alternative path varies according to type of application involved. For a simple web application, a brief outage may go unnoticed if users are not loading the web page during the outage. For real-time applications like VoIP or IP Video, an outage of several seconds may cause interrupted calls and video sessions. In addition, there are two primary types of communication over the Internet: TCP and UDP. With TCP communication, lost packets are retransmitted until the connection times out. With UDP communication, there is no built-in mechanism to retransmit lost data and UDP applications tend to fail rather than recover from brief outages.

## Bandwidth Constraints

Unlike the situation within a LAN, within a WAN there are monthly recurring charges that are generally proportional to the amount of bandwidth that is provisioned. For example, the cost of

[3] http://www.tinyvital.com/Misc/Latency.htm

T1/E1 access to an MPLS network varies from roughly $450/Mbps/month to roughly $1,000/Mbps/month. In similar fashion, the cost for T1/E1 access to a Tier 1 ISP varies from roughly $300/Mbps/month to roughly $600/Mbps/month. The variation in cost is largely a function of geography. WAN costs tend to be the lowest in the United States and the highest in the Asia-Pacific region.

To exemplify how the monthly recurring cost of a WAN leads to bandwidth constraints, consider a hypothetical company that has fifty offices and each one has on average a 2 Mbps WAN connection that costs on average $1,000/month. Over three years, the cost of WAN connectivity would be $1,800,000. Assume that in order to support an increase in traffic, the company wanted to double the size of the WAN connectivity at each of its offices. In most cases there wouldn't be any technical impediments to doubling the bandwidth. There would, however, be financial impediments. On the assumption that doubling the bandwidth would double the monthly cost of the bandwidth, it would cost the company over a three-year time frame an additional $1,800,000 to double the bandwidth. Because of the high costs, very few if any companies provision either their private WAN or their Internet access to support peak loads. As such, virtually all WANs, both private WANs and the Internet, exhibit bandwidth constraints which result in packet loss.

## Packet Loss

Packet loss can occur in either a private WAN or the Internet, but it is more likely to occur in the Internet. Part of the reason for that was previously mentioned - the Internet is a *network of networks* that consists of millions of private and public, academic, business, and government networks of local to global scope. Another part of the reason for why there is more packet loss in the Internet than there is in a private WAN is the previously mentioned Internet business model. One of the affects of that business model is that there tend to be availability and performance bottlenecks at the peering points.

If packet loss occurs, TCP will re-transmit packets. In addition, the TCP slow start algorithm (see below) assumes that the loss is due to congestion and takes steps to reduce the offered load on the network. Both of the actions have the affect of reducing throughput on the WAN.

## Characteristics of TCP

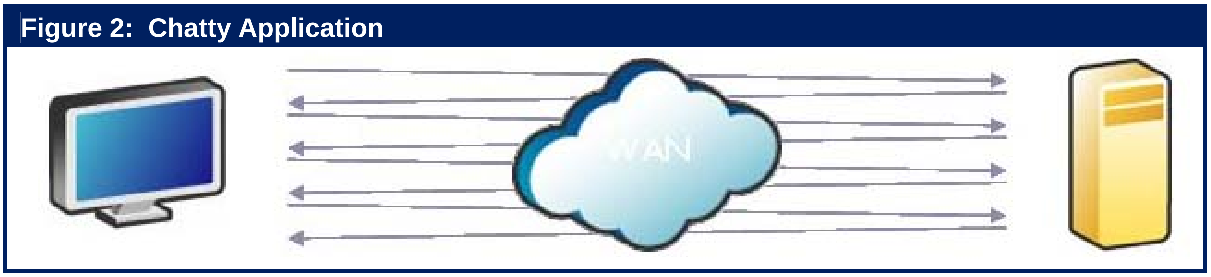
TCP is the most commonly used transport protocol and it causes missing packet(s) to be re-transmitted based on TCP's retransmission timeout parameter. This parameter controls how long the transmitting device waits for an acknowledgement from the receiving device before assuming that the packets were lost and need to be retransmitted. If this parameter is set too high, it introduces needless delay as the transmitting device sits idle waiting for the timeout to occur. Conversely, if the parameter is set too low, it can increase the congestion that was the likely cause of the timeout occurring.

Another TCP parameter that impacts performance is the TCP slow start algorithm. The slow start algorithm is part of the TCP congestion control strategy and it calls for the initial data transfer between two communicating devices to be severely constrained. The algorithm calls for the data transfer rate to increase if there are no problems with the communications. In addition to the initial communications between two devices, the slow start algorithm is also applied in those situations in which a packet is dropped.

## Chatty Protocols and Applications

The lack of emphasis on an application's performance over the WAN during application development is one of the factors that can result in the deployment of chatty applications[4] as illustrated in Figure 2.

Figure 2: Chatty Application



To exemplify the impact of a chatty protocol or application, let's assume that a given transaction requires 200 application turns. Further assume that the latency on the LAN on which the application was developed was 5 milliseconds, but that the round trip delay of the WAN on which the application will be deployed is 100 milliseconds. For simplicity, the delay associated with the data transfer will be ignored and only the delay associated with the application turns will be calculated. In this case, the delay over the LAN is 1 second, which is generally not noticeable. However, the delay over the WAN is 20 seconds, which is very noticeable.

The preceding example also demonstrates the relationship between network delay and application delay.

***A relatively small increase in network delay can result a significant increase in application delay.***

The Survey Respondents were asked how important it is for their IT organization over the next year to get better at optimizing the performance of chatty protocols such as CIFS. Their responses and the responses of last year’s survey respondents are shown in Table 2.

Table 2: Importance of Optimizing Chatty Protocols

| Level of Importance | 2011 Responses | 2012 Responses |
|---|---|---|
| Extremely | 12% | 6% |
| Very | 27% | 21% |
| Moderately | 33% | 33% |
| Slightly | 18% | 24% |
| Not at all | 10% | 16% |

Optimizing chatty protocols such as CIFS was one of the primary challenges that gave rise to the first generation of WAN optimization products in the 2007 time frame. The data in Table 2 indicates that optimizing chatty protocols has become somewhat less important to IT organizations. That said, the data in Table 2 indicates that for 60% of The Survey Respondents

[4] Similar to a chatty protocol, a chatty application requires hundreds of round trips to complete a transaction.

it is at least moderately important for their organization to get better at optimizing these protocols over the next year.

## Myriad Application Types

The typical enterprise relies on hundreds of applications of different types, including applications that are business critical, enable other business functions, support communications and collaboration, are IT infrastructure-related (i.e., DNS, DHCP) or are recreational and/or malicious.  In addition, an increasing amount of traffic results from social media.

Because they make different demands on the network, another way to classify applications is whether the application is real time, transactional or data transfer in orientation.  For maximum benefit, this information must be combined with the business criticality of the application.  For example, live Internet radio is real time but in virtually all cases it is not critical to the organization’s success.

## Webification of Applications

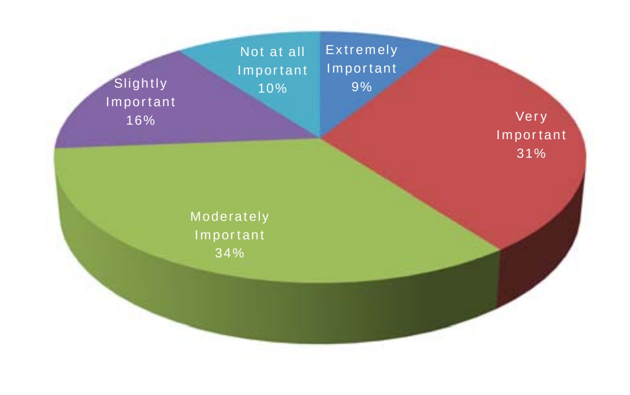
The phrase Webification of Applications refers to the growing movement to implement Web-based user interfaces and to utilize Web-specific protocols such as HTTP.  Web-based applications is a mainstream model of computing in which an application is accessed over the Internet or an Intranet and the user interface is a browser.  Web-based applications are popular in part due to the ubiquity of web browsers.  Another reason for the popularity of Web-based applications is that in contrast to traditional client/server applications, an upgrade to the server-side code does not require that changes be made to each client.  Browser functionality is a key enabler that allows businesses to adopt BYOD (Bring Your Own Device), and hence avoid the capital investment it takes to refresh end user devices, but still have the requisite functionality to enable users to successfully access applications.

There are, however, multiple challenges associated with this class of application.  The security challenges associated with this class of application was highlighted in IBM’s X-Force 2010 Trend and Risk Report[5].  That report stated that, “Web applications accounted for nearly half of vulnerabilities disclosed in 2010 -- Web applications continued to be the category of software affected by the largest number of vulnerability disclosures, representing 49 percent in 2010. The majority represented cross site scripting and SQL injection issues.”

There are also performance challenges that are somewhat unique to this class of application. For example, unlike CIFS, HTTP is not a chatty protocol.  However, HTTP is used to download web pages and it is common for a web page to have fifty or more objects, each of which requires multiple round trips in order to be transferred.  Hence, although HTTP is not chatty, downloading a web page may require hundreds of round trips.

[5] http://www-07.ibm.com/businesscenter/au/services/smbservices/include/images/Secure_mobility.pdf

Figure 3: Importance of Optimizing Protocols Other than TCP



The Survey Respondents were asked how important it was over the next year for their IT organization to get better at optimizing protocols other than TCP; e.g., HTTP and MAPI. Their answers, which are shown in Figure 3, demonstrate that the webification of applications and the number of round trips associated with downloading a web page is a traditional application delivery challenge that is still of interest to the vast majority of IT organizations.

An extension of the traditional problems associated with the webification of applications is that many organizations currently support Web-based applications that are accessed by customers. In many cases, customers abandon the application, and the company loses revenue, if the application performs badly. Unfortunately, according to market research[6], these Web-based applications have become increasingly complex. One result of that research is depicted in Table 3. As shown in that table, the number of hosts for a given user transaction varies around the world, but is typically in the range of six to ten.

Table 3: The Number of Hosts for a Web-Based Transaction

| Measurement City | Number of Hosts per User Transaction |
|---|---|
| **Hong Kong** | 6.12 |
| **Beijing** | 8.69 |
| **London** | 7.80 |
| **Frankfurt** | 7.04 |
| **Helsinki** | 8.58 |
| **Paris** | 7.08 |
| **New York** | 10.52 |

Typically several of the hosts that support a given Web-based transaction reside in disparate data centers. As a result, the negative impact of the WAN (i.e., variable delay, jitter and packet loss) impacts the Web-based transaction multiple times. The same research referenced above also indicated that whether or not IT organizations are aware of it, public cloud computing is having an impact on how they do business. In particular, that research showed that well over a third of Web-based transactions include at least one object hosted on Amazon EC2.

***Web-based applications present a growing number of management, security and performance challenges.***

[6] Steve Tack, Compuware, Interop Vegas, May 2011

## Server Consolidation

The majority of companies consolidated servers out of branch offices and into centralized data centers. This consolidation usually reduced cost and enabled IT organizations to have better control over the company's data.

***While server consolidation produces many benefits, it can also produce some significant performance issues.***

Server consolidation typically results in a chatty protocol such as Common Internet File System (CIFS), which was designed to run over the LAN, running over the WAN.

## Data Center Consolidation

In addition to consolidating servers, over the last few years many companies also reduced the number of data centers they support worldwide. This increases the distance between remote users and the applications they need to access.

***One of the effects of data center consolidation is that it results in additional WAN latency for remote users.***

The reason why the preceding conclusion is so important is because, as previously discussed, even a small increase in network delay can result in a significant increase in application delay.

## Server Overload

A server farm is a group of servers that are networked together with the goal of meeting requirements that are beyond the capability of a single server. One of the challenges associated with implementing a server farm is to ensure that a request for service is delivered to the most appropriate server. There are many ways to define what the phrase *most appropriate server* means. Certainly the server has to be available. Ideally, the most appropriate server is the server that is processing the lightest load of any member of the server farm.

In addition to the situation in which there are more requests for service than can be handled by a single server, another way that a server can become overloaded is by having to process computationally intense protocols such as SSL.

## Distributed Employees

The 80/20 rule in place until a few years ago stated that 80% of a company's employees were in a headquarters facility and accessed an application over a high-speed, low latency LAN. The new 80/20 rule states that 80% of a company's employees access applications over a relatively low-speed, high latency WAN.

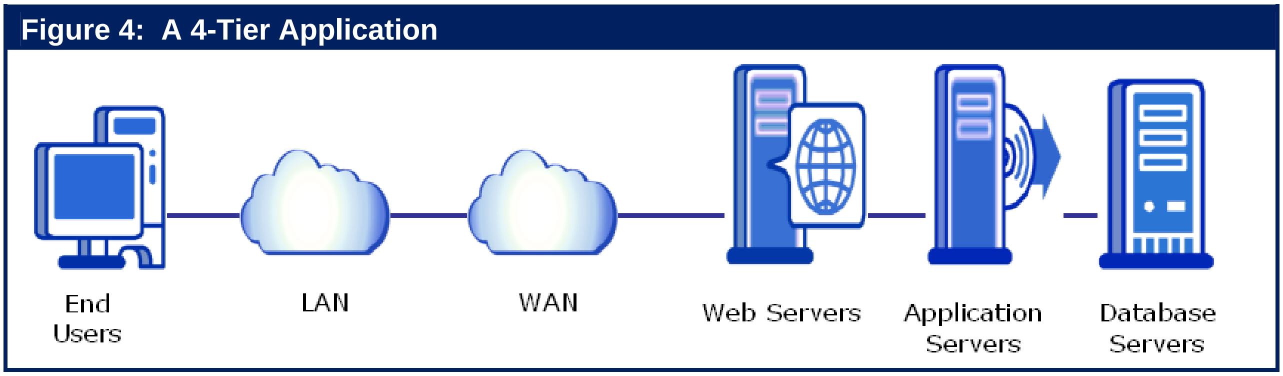
***In the vast majority of situations, when people access an application they are accessing it over the WAN instead of the LAN.***

The preceding discussion of chatty protocols and applications exemplifies one of the challenges associated with accessing an application over a WAN. As that discussion showed, there are protocols and applications that perform in acceptable fashion when run over a LAN but which perform unacceptably when run over a WAN – particularly if the WAN exhibits even moderate levels of latency.  The impact of that challenge is exacerbated by the fact that applications are typically developed over a LAN and as previously documented, during the application development process most IT organizations pay little if any attention to how well an application will run over the WAN.

## Distributed Applications

Most IT organizations have deployed a form of distributed computing often referred to as *n-tier applications*.  The browser on the user's device is typically one component of an n-tier application.  The typical 4-tier application (Figure 4) is also comprised of a Web tier, an application tier and a data base tier which are implemented on a Web server(s), an application server(s) and a database server(s).   Until recently, few, if any, of the servers were virtualized.

**Figure 4:  A 4-Tier Application**



Distributed applications increase the management complexity in part because each tier of the application is implemented on a separate system from which management data must be gathered.  The added complexity also comes from the fact that the networks that support these applications are comprised of a variety of switches, routers, access points, WAN optimization controllers, application delivery controllers, firewalls, intrusion detection systems and intrusion protection systems from which management data must also be gathered.

As recently as a few years ago, few, if any, of the servers in the typical n-tier application were virtualized.  However, in the current environment it is becoming increasingly common to have these servers be virtualized and that further complicates the task of ensuring acceptable application and service delivery.

## Complexity

The overall complexity of both private WANs and the Internet tends to increase the impact of the previously described application delivery challenges.  For example, as the number of links that the data has to transit between origin and destination increases, so does the delay.  As delay increases, the negative impact of a chatty protocol or application is magnified.

It is not, however, just the number of links and the complex topologies that complicate application delivery, it is also the use of complex protocols such as TCP and BGP.  The Internet uses BGP to determine the routes from one subtending network to another.  When choosing a route, BGP strives to minimize the number of hops between the origin and the destination.  BGP doesn't, however, strive to choose a route with the optimal performance characteristics; i.e., lowest delay, lowest packet loss. Given the complex, dynamic nature of the Internet, a given network or a particular peering point router can go through periods where it exhibits severe delay and/or packet loss. As a result, the route that has the fewest hops is not necessarily the route that has the best performance.

As noted in the preceding paragraph, the traditional distributed application environment is complex in part because there are so many components in the end-to-end flow of a transaction. If any of the components are not available, or are not performing well, the performance of the overall application or service is impacted. In some instances, each component of the application architecture is performing well, but due to the sheer number of components the overall delay builds up to a point where some function, such as a database query, fails. Some of the implications of this complexity on performance management are that:

***As the complexity of the environment increases, the number of sources of delay increases and the probability of application degradation increases in a non-linear fashion.***

***As the complexity increases the amount of time it takes to find the root cause of degraded application performance increases.***

***As the complexity increases, so does the vulnerability to security attacks.***

## Expanding Scope of Business Critical Applications

A decade or so ago when business critical applications were deployed, the scope of the application was intra-company; e.g., all of the users of the application were employees of the company. In the current environment, virtually all organizations use their applications and networks to interact with myriad people outside of the company including suppliers, business partners and customers. This presents some significant challenges for IT organizations, as they are typically held responsible for the performance and management of these applications even though in many cases they don't have access to the enabling IT infrastructure that they would have if the application was entirely intra-company.

## Increased Regulations

Most governments and regulators are aware of the growing criticality of IT in general, and of the increased importance of the Internet in particular. This awareness has led to increased legislation and regulation as governments attempt to exert control over how businesses operate. These regulations range from law enforcement (i.e., Communications Assistance for Law Enforcement Act – CALEA) to privacy (i.e., Health Insurance Privacy and Accountability Act – HIPPA) to fraud protection (i.e., Payment Card Industry Data Security Standard – PCI DSS) and to hundreds of other regulations. Legislative processes, however, operate considerably slower than the technology industry does and so regulations are often out of date with, or inappropriate for the current technology. Another application and service delivery challenge is that with a growing body of laws and regulations at both the national and local level, it is often difficult for an IT organization to know if they are in compliance with regulations.

## Security Vulnerabilities

Security vulnerabilities can be classified as both a first and a second-generation application and service delivery challenge. The distinction between a first and a second-generation security challenge is based on factors such as who is doing the attack, what are they attacking, what tools and techniques are they using and what is their motivation.

For example, until recently the majority of security attacks were caused by individual hackers, such as Kevin Mitnick, who served five years in prison in the late 1990s for computer- and communications-related hacking crimes. The goal of this class of hacker is usually to gain notoriety for themselves and they often relied on low-technology techniques such as dumpster diving.

However, over the last few years a new class of hacker has emerged and this new class of hacker has the ability in the current environment to rent a botnet or to develop their own R&D lab. This new class includes crime families and hactivists such as Anonymous. In addition, some national governments now look to arm themselves with Cyber Warfare units and achieve their political aims via virtual rather physical means. Examples include China's alleged attack on Google and the Stuxnet attack on Iran's nuclear program.

In addition to the types of attacks mentioned in the preceding paragraph, one of the ways that the sophistication of the new generation of attackers has been manifested is just the sheer scale of the attacks. As recently as a decade ago, the peak rate of Distributed Denial of Service (DDoS) attacks was roughly 500 Mbps. In the current environment, the peak rate is more than 50 Gbps. This means that over the last decade the peak rate of a DDoS attack has increased by at least a factor of one hundred. Another example of the sophistication of the current generation of hacker is the growing number of attacks based on SQL injection. In this type of attack, malicious code is inserted into strings that are later passed to an instance of SQL Server for parsing and execution. The primary form of SQL injection consists of direct insertion of code into user-input variables that are concatenated with SQL commands and executed.

In March 2012, IBM published its annual X-Force 2011 Trend and Risk Report[7]. That report highlighted the fact that new technologies such as mobile and cloud computing continue to create challenges for enterprise security. Some of the key observations made in that report are:

- **Mobile Devices**
  The report stated that in 2011 there was a 19 percent increase over 2010 in the number of exploits publicly released that can be used to target mobile devices such as those that are associated with the movement to Bring your Own Device (BYOD) to work. The report added that there are many mobile devices in consumers' hands that have unpatched vulnerabilities to publicly released exploits, creating an opportunity for attackers.

- **Social Media**
  With the widespread adoption of social media platforms and social technologies, this area has become a target of attacker activity. The IBM report commented on a surge in phishing emails impersonating social media sites and added that the amount of information people are offering in social networks about their personal and professional lives has begun to play

[7] X-Force 2011 Trend and Risk Report

a role in pre-attack intelligence gathering for the infiltration of public and private sector computing networks.

- **Cloud Computing**
  According to the IBM report, in 2011, there were many high profile cloud breaches affecting well-known organizations and large populations of their customers. IBM recommended that IT security staff should carefully consider which workloads are sent to third-party cloud providers and what should be kept in-house due to the sensitivity of data. The IBM X-Force report also noted that the most effective means for managing security in the cloud may be through Service Level Agreements (SLAs) and that IT organizations should pay careful consideration to ownership, access management, governance and termination when crafting SLAs.

The Blue Coat Systems 2012 Web Security Report[8], focused on a number of topics including malnets and social networking. A malware network, or malnet, gathers users, most frequently when they are visiting trusted sites and routes them to malware. According to the Blue Coat Report, "In 2011, malnets emerged as the next evolution in the threat landscape. These infrastructures last beyond any one attack, allowing cybercriminals to quickly adapt to new vulnerabilities and repeatedly launch malware attacks. By exploiting popular places on the Internet, such as search engines, social networking and email, malnets have become very adept at infecting many users with little added investment."

The report noted the increasing importance of social networking and stated that, "Since 2009, social networking has increasingly eclipsed web-based email as a method of communications." The report added that, "Now, social networking is moving into a new phase in which an individual site is a self-contained web environment for many users – effectively an Internet within an Internet." For example, according to the Blue Coat report 95% of content types that are found on the Internet are also found within social networking sites. The five most requested subcategories of content that were requested from social networking sites, and the percentage of times that they were requested are shown in Table 4.

**Table 4: Most Requested Content from Social Media Sites**

| Subcategory of Content | Percentage of Times it was Requested |
|---|---|
| Games | 37.9% |
| Society/Daily Living | 23.8% |
| Personal Pages/Blogs | 6.4% |
| Pornography | 4.9% |
| Entertainment | 4.2% |

Part of the challenge that is associated with social network sites being so complex is that IT organizations can not just look at a social media site as one category and either allow or deny access to it. Because these sites contain a variety of classes of content, IT organizations need the granular visibility and control to respond differently to requests at the same social media site for different types of content.

[8] http://www.bluecoat.com/sites/default/files/documents/files/BC_2012_Security_Report-v1i-optimized.pdf