A SUMMARY VERSION

The 2009 Handbook of Application Delivery

A Guide to Decision Making in Challenging Economic Times

By Dr. Jim Metzler

PLATINUM SPONSORS

www.webtorials.com

Introduction

This document highlights the key concepts that are contained in the 2009 Application Delivery Handbook. Throughout this document, the full version of the handbook will be referred to as The Handbook and the summary will be referred to as the handbook summary. Readers seeking more detailed information on any of the topics covered in the handbook summary should download The Handbook[1].

Throughout the handbook summary, the phrase *application delivery* will refer to the task of ensuring that the applications that an enterprise uses:

- Can be effectively managed
- Exhibit acceptable performance
- Incorporate appropriate levels of security
- Are cost effective

Over the last few years application delivery has become a priority for virtually all IT organizations. However, while many IT organizations have become better at application delivery, the majority of IT organizations struggle with this highly complex task. One of the primary goals of The Handbook is to help IT organizations become better with application delivery. One of the ways that The Handbook summary achieves that goal is by creating and analyzing a framework that IT organizations can customize for use in their environment. The four primary components of the framework are:

- Planning
- Network and application optimization
- Management
- Control

Another way that The Handbook achieves that goal is by identifying criteria that IT organizations can use when evaluating alternative solutions.

The Worldwide Economic Environment

In 2009 the economy is politely referred to as being *challenging*. While this challenging economic environment will certainly put pressure on IT budgets, IT organizations need to continue to invest in application delivery solutions. That follows in large part because IT organizations have made, and continue to make, significant investments in enterprise applications to support key business processes. IT organizations need to protect these investments, and the business processes that they enable, by ensuring that these applications can be effectively delivered. It also follows because many of the optimization techniques described in chapter 6 of The Handbook will reduce cost and many of the techniques described in the rest of The Handbook also contribute to the IT organization's ability to manage cost. For example, the majority of IT organizations are deploying virtualized servers as a way to reduce cost. As discussed in chapter 4 of The Handbook, these initiatives will not be successful if IT organizations do not overcome the management challenges that are associated with virtualized servers.

[1] The 2009 Application Delivery Handbook is available at http://webtorials.com/abstracts/2009-Application-Delivery-Handbook.htm

Application Delivery Challenges

One of the key challenges associated with ensuring acceptable application delivery is that in the vast majority of instances it is the end user, and not the IT organization, that first notices application degradation. Making matters worse, when the IT organization is made aware that the performance of an application is degrading, the organization is often unsure of the cause in part because any component of IT could be the cause of that degradation.

In addition, in most instances:

- IT organizations tend to plan and manage each component of IT in isolation from each other.
- In many instance, each component of the IT infrastructure can perform well, yet the overall performance of the application is unacceptable.
- White papers and other documents that are intended to help IT organizations get better at application delivery tend to focus on just one component of IT. As such, these documents provide little guidance relative to the end-to-end issues caused by complex applications running over a complex infrastructure.

The Handbook is designed to provide the end-to-end guidance that the typical white paper cannot provide. An unfortunate side affect of providing that guidance is that The Handbook is lengthy which is why this summary handbook was created.

As noted, one of the key challenges associated with ensuring acceptable application performance is that the individual groups within the IT organization function in a siloed or stove-piped fashion. That means that these groups typically do not share:

- Terminology[2]
- Goals
- Tools
- Processes

While they can make some progress on their own, there is a limit to how much success the rank and file of the IT organization can make relative to eliminating organizational and technological silos. This creates a CIO mandate whereby the CIO must drive a transformation to where the IT infrastructure organization focuses not on individual technologies, but on a systematic approach to application delivery. For example, the role of the network operations center (NOC) has changed significantly. In most cases, this change has occurred without the support of senior IT management. In large part due to the lack of that support, roughly a quarter of NOCs do not satisfy the current expectations that are placed on the NOC. Using the NOC as just one example of the CIO mandate, what is needed is for senior IT managers to drive the evolution to an Integrated Operations Center (IOC) that has effective tools and processes to support all components of IT, both individually and as an end-to-end system.

The complexity of application delivery is driven in part by the evolving applications environment. Some of the components of that environment that drive complexity include:

- An application development process that largely ignores the impact of the WAN.
- The webification of applications that introduces protocols that are chatty and dense.

[2] The word *service* is a good example of this lack of common terminology. The various subgroups within an IT organization typically use the word service to refer to different things.

- Server consolidation that results in more users accessing applications over a WAN.
- Data center consolidation that results in the WAN link between the user and the application being lengthy and hence exhibiting high levels of latency and packet loss.
- The decentralization of employees that results in even more employees accessing applications over the WAN.
- The use of Software as a Service that results in less IT control over the management and performance of applications.
- The deployment of distributed applications, which introduces additional sources of delay.
- The use of Web 2.0 applications and application development techniques both of which result in less IT control over the management and performance of applications.

Another factor that increases the complexity of application delivery is virtualization. There can be compelling reasons to virtualize servers, desktops and storage. However, each of these initiatives present distinct management and/or optimization challenges.

Planning

There are a number of planning functions that are critical to successful application delivery. One of these functions is the use of WAN emulation to enable software engineers to develop applications that perform well over a WAN. Another key function is baselining. Baselining provides a reference from which service quality and application delivery effectiveness can be measured. It does so by quantifying the key characteristics (e.g., response time, utilization and delay) of applications and various IT resources including servers, WAN links and routers.

An important task for all IT organizations is to integrate planning and operations. One of the reasons to integrate planning and operations is that it results in the reduction in the number of management tools that must be acquired and supported. This reduces cost, which is particularly important in this challenging economic environment. Another reason to integrate planning and operations is because it also increases the communications within the IT organization. This follows because fewer tools result in less disagreement over the health of the IT infrastructure and the applications that use that infrastructure. One of the technologies that can be used to better integrate planning and operations is route analytics.

Network and Application Optimization

The phrase *network and application optimization* refers to an extensive set of techniques that organizations have deployed in an attempt to optimize the performance of networks and applications as part of assuring acceptable application performance. The primary role these techniques play is to:

- Reduce the amount of data sent over the WAN.
- Ensure that the WAN link is never idle if there is data to send.
- Reduce the number of round trips (a.k.a., transport layer or application turns) necessary for a given transaction.
- Mitigate the inefficiencies of protocols and applications.
- Offload computationally intensive tasks from client systems and servers

Some of the basic tasks of network and application optimization can be gained by deploying devices that function within the packet delivery network. By *packet delivery* network is meant the packet payload and the transport, network and data link layers of the Internet protocol suite. However, more sophisticated techniques require an application delivery network (ADN). ADN solutions leverage functionality that resides higher in the OSI protocol stack and can improve the effectiveness of application delivery based on the ability of these solutions to recognize application layer signatures and to then differentiate among the various applications that share and contend for common transport resources. Some of the primary ADN characteristics include optimization, management and control.

There are two principal categories of network and application optimization products. One category is typically referred to as a WAN Optimization Controller (WOC). WOCs are often referred to as *symmetric solutions* because they typically require an appliance in both the data center as well as the branch office. In most cases, WOCs are implemented as an appliance. As is described below, however, some vendors have implemented virtualized WOCs. This class of solution is often referred to as a s*oftware only solution* or as a *soft WOC*.

The goal of a WOC is to improve the performance of applications delivered from the data center to the branch office or directly to the end user over networks such as Frame Relay, ATM or MPLS. WOCs implement a wide variety of technologies, including caching, compression, congestion control, forward error correction, protocol acceleration, as well as request prediction and spoofing.

The second category of network and application optimization products is typically referred to as Application Delivery Controllers (ADCs). ADCs began as simple layer 4 load balancers but now provide a wide range of functionality including SSL offload, application firewall, global traffic distribution, rate shaping, DDoS/DoS protection, asymmetrical application acceleration and response time monitoring.

Just as devices such as servers can be virtualized, so can appliances such as WOCs. A *Virtual Appliance* is based on network appliance software, together with its operating system, running in a virtual machine in a virtualized server. Virtual appliances can include WOCs, firewalls, and performance monitoring solutions among others. A virtual appliance offers the potential to alleviate some of the management burdens in branch offices because most of the provisioning, software updates, configuration, and other management tasks can be automated and centralized at the data center.

Another form of virtualization is clustering. For example, it is possible to cluster a number of ADCs and have the cluster perform as a single ADC. Another option is to implement a cluster of physical appliances with an ADC providing the load balancing across the individual appliance platforms.

Several types of appliances such as ADCs can support yet another form of virtualization, where the system's hardware platform supports a number of independent software partitions. A partitioned appliance can be configured to dedicate a separate partition to each application or service being delivered. This allows the configuration of each partition to be optimized for the specific type of application traffic being processed.

Managed Service Providers (MSP)

Managed Service Providers (MSPs) are not new. The last few years, however, have seen the development of a new class of MSP – the Application Delivery MSP (ADMSP). Two of the many benefits of using an ADMSP are the ability to leverage both the ADMSP's expertise and their technology.

There are two primary categories of managed application delivery services provided by ADMSPs: site-based services and Internet-based services. Site-based services are comprised of managed WOCs and/or ADCs installed at participat-

ing enterprise sites. The application optimization service may be offered as an optional add-on to a WAN service or as a standalone service that can run over WAN services provided by a third party. Where the application delivery service is bundled with a managed router and WAN service, both the WOC and the WAN router would be deployed and managed by the same MSP.

Whether implemented in a do-it-yourself (DIY) manner or via site-based services, the traditional classes of application delivery solutions (ADC, WOC, soft WOC) were designed to address application performance issues at both the client and server endpoints. These solutions make the assumption that performance characteristics within the WAN itself are not optimizable because they are determined by the relatively static service parameters controlled by the WAN service provider. This assumption is reasonable in the case of private WAN services. However, this assumption does not apply to enterprise application traffic that transits the Internet because there are significant opportunities to optimize performance within the Internet itself based on Application Delivery Services (ADSs).

An ADS is an Internet-based services that focuses on the acceleration of the increasing number of applications that traverse the Internet. Ensuring acceptable application performance over the Internet is difficult because the Internet is a network of networks and the only service providers that get paid to carry Internet traffic are the providers of the first and last mile services. All of the service providers that carry traffic between the first and last mile do so without compensation. One of the affects of this business model is that there tends to be availability and performance bottlenecks at the peering points. Another affect is that since there is not a single, end-to-end provider, service level agreements (SLAs) for the availability and performance of the Internet are not available.

An ADS leverages service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. All client requests to the application's origin server in the data center are redirected via DNS to an ADS server in a nearby point of presence (PoP). This edge server then optimizes the traffic flow to the ADS server closest to the data center's origin server.

Management

Part of the challenge facing IT organizations, and another reason for the mandate to have CIOs drive a transformation of the IT organization, is the organizational dynamic that exists inside of many IT organizations. Part of that organizational dynamic is that less than half of IT organizations indicate that there is a cooperative relationship between their application development organization and their network organization. Another part of the dynamic is that ineffective management processes are one of the biggest impediments to effective application delivery.

There are a number of management tasks that are essential to successful application delivery. As noted, one of the key challenges associated with ensuring acceptable application delivery is that in the vast majority of instances it is the end user, and not the IT organization, the first notices application degradation. As such, the ability to have end-to-end visibility is a minimum management requirement. In this context, *end-to-end visibility* refers to the ability of the IT organization to examine every component of IT that impacts communications once users hit ENTER or click the mouse button when they receive a response from an application. IT organizations that are looking to deploy a tool to provide end-to-end visibility should evaluate these solutions using a broad set of selection criteria.

The port 80 black hole creates a management and a control challenge. Port 80 is the port that servers listen to while expecting to receive data from Web clients. As a result, a firewall can't block port 80 without eliminating much of the traffic on which a business may depend. Taking advantage of this fact, many applications will port-hop to port 80 when their normally assigned ports are blocked by a firewall. This behavior creates what is referred to as the *port 80 black hole*. Well-known applications that do port hoping include AOL's instant messaging (AIM), Skype and applications based on

the Financial Information eXchange (FIX) protocol. Just looking at these three examples, the port 80 black hole creates issues relative to:

- Security – AIM can carry viruses and worms
- Compliance – In some instances, regulations require that IMs must be archived
- Legal – The file sharing enabled by Skype can be against the law
- Performance – FIX based applications can be very time sensitive

The traditional approach that most IT organizations have taken relative to network and application alarming is to set static threshold alarms. One of the problems with static threshold alarms is that most IT organizations set them at a high value. As a result, most IT organizations miss the majority of alarms. An alternative approach is referred to as proactive alarms or analytics. The goal of analytics is to automatically identify and report on possible problems in real time so that organizations can eliminate the problems before they impact users. One key concept of proactive alarming is that it takes the concepts of baselining and applies these concepts to real-time operations. One of the key selection criteria for an analytics solution is that the solution needs to be able to baseline the network to identify normal patterns and then identify in real time a variety of types of changes in network traffic.

As mentioned, one of the technologies that can be used to better integrate planning and operations is route analytics. From an ongoing management and operations perspective, the goal of route analytics is to provide visibility, analysis and diagnosis of the issues that occur at the routing layer. A route analytics solution achieves this goal by providing an understanding of precisely how IP networks deliver application traffic. This requires the creation and maintenance of a map of network-wide routes and of all of the IP traffic flows that traverse these routes. This in turn means that a route analytics solution must be able to record every change in the traffic paths as controlled and notified by IP routing protocols. One of the key selection criteria that an IT organization should look at when selecting a route analytics solution is the breadth of routing protocol coverage that the solution provides.

Application Performance Management

Application performance management (APM) is a relatively new management discipline. The newness of APM is attested to by the fact that ITIL has yet to create a framework for APM. Successful APM requires a holistic approach based on integrated management of both the application itself as well as the end-to-end IT infrastructure. This approach must focus on the experience of the end user of the application or service and must address most, if not all, of the following aspects of management:

- Adoption of service level agreements (SLAs) for at least a handful of key applications and services.
- End-to-end monitoring of all end user transactions.
- Automatic discovery of all the elements in the IT infrastructure that support the key applications and services. These are referred to as The Key Elements.
- Establishing performance objectives for The Key Elements.
- Proactive and predictive monitoring of The Key Elements.
- Prioritizing outages and other incidents based on potential business impact.

- Triage and root cause analysis applied at both the application and infrastructure levels.
- Automated generation of performance dashboards and historical reports to allow both IT and business managers to gain insight into SLA compliance and performance trends.

Automation

The automation of management tasks is a critical topic for multiple reasons. One reason is that the majority of IT capital and personnel resources are consumed maintaining the status quo and this percentage increases every year as more functionality is added to the IT infrastructure. The second reason is that performing repetitive, time-consuming tasks is error prone. Automation has the potential to reduce the amount of resources consumed by management tasks and simultaneously to improve the quality associated with those tasks.

Some of the tasks that it makes the most sense to automate include:

- Configuration Management
- Event Management
- Service Level Management
- Security Management

Control

To effectively control both how applications perform, as well as who has access to which applications, IT organizations must be able to utilize the following functionality:

- **Route optimization**
  The goal of route optimization is to make intelligent decisions relative to how traffic is routed through an IP network. Route optimization is an integral part of an Internet-based service from an ADMSP. Route optimization can also be applied by IT organizations whenever there is the possibility of multiple paths from between end points.
- **SSL VPN Gateways**
  One of the purposes of an SSL VPN gateway is to communicate directly with both the user's browser and the target applications and enable communications between the two. Another purpose of the SSL VPN gateway is to control both access and actions based on the user and the endpoint device. IT organizations should evaluate SSL VPN gateways based on a variety of selection criteria.
- **Traffic Management and QoS**
  Traffic Management refers to the ability of the network to provide preferential treatment to certain classes of traffic. It is required in those situations in which bandwidth is scarce, and where there are one or more delay-sensitive, business-critical applications.
- **Next Generation WAN Firewall**
  In order to overcome challenges such as the previously mentioned Port 80 Black Hole, a new generation of WAN firewall is required. IT organizations should evaluate next generation firewalls based on a variety of selection criteria.

The Road Ahead

As previously noted, there are a number of factors the make ensuring acceptable application delivery difficult; e.g., the deployment of chatty protocols, the centralization of IT resources, the use of new highly-distributed, application architectures. These factors are not going away anytime soon. In addition, there are a number of emerging challenges that will further complicate the task of ensuring acceptable application delivery. It is not possible to state with certainty which challenges will impact a particular IT organization twelve to eighteen months from now. It is, however, possible to state with certainty that all IT organizations will be impacted by change in that time frame.

One example of the type of changes that are likely to impact most IT organization is that in the not too distant future it will be common for a user in a branch office to utilize a virtualized desktop. That user will likely access the branch office router over a virtual LAN (VLAN). However, in addition to routing, that branch office router may also host virtual machines that support a variety of applications and/or Web services. In addition, since the deployment of WAN Optimization Controllers (WOCs) is increasing, in the near future it will be much more likely than it is today that the data flow within the branch office transits a WOC before it hits the WAN. However, this may not be the traditional WOC. For example, in addition to providing standard WOC functions such as caching, compression and protocol acceleration, this WOC will also host virtual machines that provide network services such as DNS and DHCP. Given the ever-increasing concern about security, in the near future it will be even more likely than it is today that there will also be a firewall in the branch office. This may be a traditional firewall, or firewall software running on a virtualized machine, possibly inside either the router or the WOC.

The data flow next transits a WAN link that today is almost always a terrestrial link. However, for both backup and performance reasons, we will see the deployment of 3G links that have delay characteristics that will further exacerbate the WAN performance issues. Upon entering the data center, the traffic hits a virtualized ADC. After transiting the ADC, the next step for the traffic is to transit a number of virtualized web servers, application servers and database server, each of which may or may not be isolated from each other by virtualized firewalls. When the application requires data it gets it from a pool of virtualized storage.

Virtualization, however, is not the only emerging technology that will complicate application delivery. New application architectures such as a Service-Oriented Architecture (SOA) with Web services and Web 2.0 will also complicate application delivery. In a Web services-based application, the Web services that comprise the application typically run on servers that are housed within multiple data centers. In many instances, at least some of these Web services reside in data centers owned by a company's partners, customers and suppliers. As a result, the IT organization has little insight into, or control over, what is happening in those data centers. In addition, the WAN impacts multiple traffic flows and hence has a greater overall impact on the performance of a Web services-based application that it does on the performance of an n-tier application.

Many IT professionals associate the phrase Web 2.0 with social networking sites such as MySpace. While that is reasonable, one of the most concrete aspects of Web 2.0 is not what it does, but the fact that Web 2.0 applications are typically constructed by aggregating other applications together. This has become such a common concept that a new term, mashup, has been coined to describe it. According to Wikipedia, a mashup is a web application that combines data from more than one source into a single integrated tool.

Mashups are powerful but challenging. When you have an application that calls on another application that is designed, controlled and operated by another organization, you have given up virtually all visibility and control over that piece of your overall application. If there is an availability or performance problem, in many cases you have little recourse other than to wait for the problem to go away.

Summary

Over the last few years, ensuring acceptable application delivery difficult has become a top of mind issue for most IT organizations. Many IT organizations, however, struggle with this highly complex task. Unfortunately, due to the ongoing deployment of new technologies and services, ensuring acceptable application delivery will become even more difficult over the next few years. As a result, IT organizations need to ensure that the approach that they take to resolving the current application delivery challenges can scale to support the emerging challenges.

Given the complexity associated with ensuring acceptable application delivery, IT organizations will not be successful if they approach the task in a piecemeal fashion. As suggested in this handbook summary, successful application delivery requires four key functions: planning, optimization, management and control. Because they are inter-related, these functions must be looked at holistically. For example, as mentioned, a key planning function is to develop applications that perform well over the WAN. If IT organizations implement this planning function, it lessens their need for optimization. In contrast, if the planning function they implement calls for technologies such as virtualized desktops, or the use of application architectures such as SOA or cloud computing, that would increase their need for optimization.

The full version of the handbook elaborates on the key tasks associated with planning, optimization, management and control and provides decision criteria to help IT organizations choose appropriate solutions. The full version also describes in depth:

- How to apply the application delivery framework to applications such as VoIP.
- The factors that complicate the task of ensuring acceptable application delivery today, as well as some of the factors that are likely to complicate the task in the near term.
- Both the types of, and the value of application delivery managed service providers.
- The changing role of the NOC.
- The role of the CIO in application delivery.
- The specific steps that six IT organizations have taken to improve their ability to ensure acceptable application delivery.

As noted, the 2009 Application Delivery Handbook is available at:
http://webtorials.com/abstracts/2009-Application-Delivery-Handbook.htm

THE HANDBOOK OF APPLICATION DELIVERY

Published by
Webtorials
Editorial/Analyst Division
www.Webtorials.com

Division Cofounders:
Jim Metzler
jim@ashtonmetzler.com
Steven Taylor
taylor@webtorials.com

Design/Layout Artist
Debi Vozikis

For Editorial and Sponsorship Information
contact Jim Metzler or Steven Taylor.
The Webtorials Editorial/Analyst Division is an analyst and consulting joint venture of Steven Taylor and Jim Metzler.