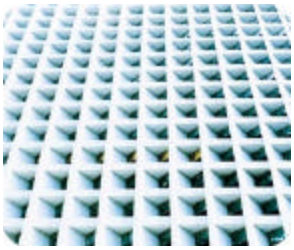

Rich Media Solutions

June 2002

business white paper
industry standard servers

enterprise digital media solution guide

part III: infrastructure

By: Microsoft Corporation - Digital Media Division
HP – Industry Standard Server Group
Approach Inc.

table of contents

abstract

This document outlines considerations and design elements for implementing an enterprise digital media solution. The document particularly focuses on the infrastructure elements required to deploy a digital media solution for e-Learning and corporate communications applications. The paper addresses content distribution and management, media portals and content creation and describes three enterprise scenarios with prescribed high-level solution architectures. The paper assumes the reader is familiar with information technology concepts and has hands-on working experience with designing IT solutions.

introduction

There are many demands for digital media-based solutions from groups within the enterprise. These requests range from encoding ad-hoc video content to deploying a complete Learning management system. For IT stakeholders, these requests can be challenging, because the technology may not be well understood and the requirements not well-defined. Many organizations choose to deploy solutions in a piecemeal fashion, rather than taking a holistic approach to implementing an enterprise solution. The technology is mature and is ready for large-scale deployment. Microsoft® and HP have solutions to meet the demand for enterprise digital media, and together they provide technologies to deliver an end-to-end digital media solution, from creation to consumption.

For many organizations, the best place to start is by addressing the infrastructure necessary to support digital media. However, designing a solution to support enterprise-wide digital media is often the most significant barrier. This white paper will help you consider the design elements for implementing a digital media infrastructure based on Microsoft and HP technologies. This paper will help you to understand what digital media is and how it will affect your existing network and infrastructure. Microsoft and HP have complementary technologies, from Microsoft Windows Media to ProLiant server from the new HP. This paper will detail the infrastructure elements required to deliver an enterprise digital media solution, including:

- Capture options for content creation
- Content distribution and management
- Developing media portals
- Network design considerations
- Security, authentication and authorization

Taking an infrastructure approach

Enterprise digital media solutions, like most enterprise solutions, have many critical components; however, for digital media solutions the key enabler is the infrastructure. The infrastructure includes network devices, servers and software. When designing solutions for digital media technology the biggest challenge is how to manage and distribute this new data type called digital media. Digital media has two unique characteristics which drive the infrastructure changes; files are very large and dedicated bandwidth is required to ensure a quality user experience.

To implement a successful enterprise digital media solution, many things need to be considered, including the business process and the user experience. But the greatest challenge in delivering a reliable, scalable and manageable solution lies in the network and systems infrastructure.

Key Infrastructure Components

Digital media is a data type that IT groups need to understand and determine how to support. Providing support for digital media is not simply about increasing bandwidth, instead it is about optimizing bandwidth and storage on a distributed basis across the network. The network can be designed to centralize content, in essence reducing storage on the edge, or to distribute content to the edge, reducing the impact on the network. There is no one right answer and the correct solution will vary for each company depending on their needs. An enterprise digital media solution will affect the infrastructure by affecting the configuration of routers and switches, increasing demand for network appliances, web servers and storage and requiring additional

software.

routing and switching

Multicasting must be enabled to support live broadcasts or scheduled multicasts. A multicast routing protocol (sparse or dense) will need to be configured on the routers. Switches will also need to be enabled to support multicasting.

networked appliances

The deployment of purpose built appliances for content distribution and management will affect how content is routed and stored on the network. Appliances are likely to become a common device found in all data centers. Appliances support a variety of operating systems and often run in 'headless' operations mode. Appliances are intended to perform one application function and provide remote administration.

server hardware

The server infrastructure will need to support the added storage requirements and web services for the digital media solutions. Digital media is storage intensive, from content creation to distribution, including the local caching required to view the content. This may require additional rack-mountable servers at central data centers as well as on a distributed basis in regional data centers.

server software

Server software such as Windows Media Encoder, Windows Media Services on Microsoft Windows 2000 and other web services will be essential to a digital media solution. Generally, server software can run on a range of hardware platforms, making it easy to deploy on existing equipment as well as new hardware.

trends towards convergence

A digital media infrastructure will serve many business applications and will be highly distributed across the network. Digital media may be perceived today as a special application with limited impact, but in the not-so-distant future, streaming applications will be mission-critical and broadly deployed across the organization.

Support for digital media at an infrastructure level is the first step towards enabling real-time communications on the TCP/IP network. There is increasing demand to converge voice and data communications on the same network, and the deployment of a digital media infrastructure will provide many side benefits as the enterprise moves towards a fully converged TCP/IP network.

keep the enterprise in mind

When getting started with digital media solutions, keeping a broad perspective is useful. The technology itself can often be seen as simple at the unit level. For example, many IT professionals will set up a ProLiant server running Windows Media Services and a Windows client and find it easy to stream some sample content. This type of lab setup is helpful in demonstrating the technology to business users and technology decision makers. However, it is not a good foundation for addressing enterprise design issues and delivering ongoing business value and growth.

If the solution is not designed for the enterprise there are two possible outcomes. The digital media initiative can get stuck in the lab and never get deployed because the enterprise requirements were not considered properly. Or even worse, it does get deployed without being ready for production and is unable to meet the required service levels. Common requirements include security, management of distribution, quality of service, bandwidth management, URL redirection, business processes for content creation and many more.

enterprise considerations

Windows Media Services on ProLiant servers with Windows 2000 provide a stable and robust enterprise platform for deploying digital media solutions. These products and technologies enable content production, distribution, and presentation to be tightly integrated into a highly scalable,

flexible and well-managed enterprise digital media infrastructure.

Here is a listing of some of the key services that are often overlooked but which most enterprises will want to consider when deploying a digital media infrastructure. Adding these services will ensure the solution will scale with future needs.

- Content management and distribution
- Systems management
- Bandwidth management
- Security integration
- Fault tolerance
- Load balancing
- Operations and administration

enterprise scenarios

Three scenarios centered on the application of digital media follow, describing a business objective, an opportunity and a proposed solution, in sample small, medium and large enterprises. These scenarios help illustrate typical applications of the technology along with a high-level view of the primary solution components.

Scenario #1: Small Enterprise

Business challenge: A consulting firm with 500 employees across five locations would like to deploy a streaming solution to manage the distribution of training content. The firm needs a way for subject matter experts to capture and distribute their knowledge on specific business topics. The firm currently uses conventional methods such as written articles and instructor-led training classes as the primary means of knowledge transfer. The firm already has training rooms with audio/visual equipment set up for taping the sessions.

Opportunity: Improve the overall efficiency of training by reducing costs and increasing the rate of employee training. Increase knowledge sharing and have it scale to meet the demands of the business.

Solution: The firm took a holistic approach and created an end-to-end process for capturing and distributing knowledge. By creating new business processes and enhancing the existing training rooms, the firm was able to create 30% more content for knowledge sharing. In addition to capturing more business knowledge, the firm was also able to increase the accessibility and usefulness of that knowledge which is even more important.

The firm implemented a content distribution solution and a media portal to manage the distribution and consumption of training content. The two primary solution components are described.

Content Distribution

To manage the distribution of content, the firm deployed an origin server at the central office and two remote web/media servers to handle requests from the larger sites of 100 and 75 users. The smaller sites will traverse the WAN to get content at the central site. The content management software on the web and media servers controls the synchronization of content on all three servers. As the smaller sites increase in size, it will be easy to add local servers to stage content on.

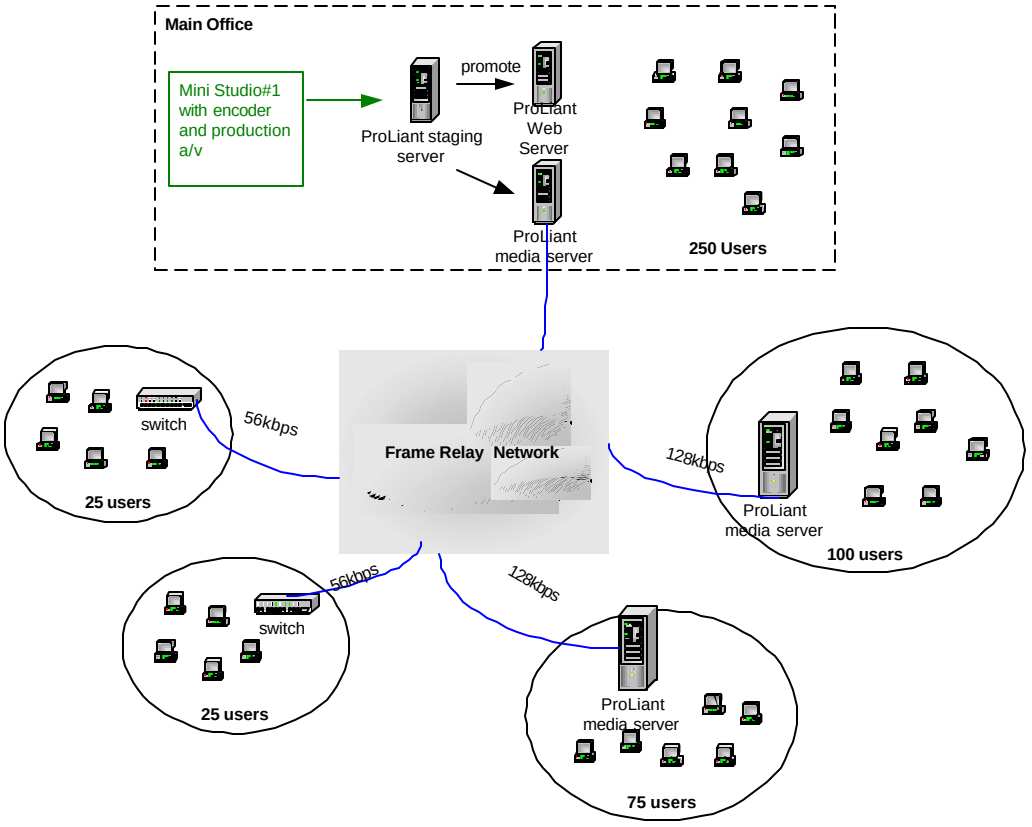

Figure 1: Small enterprise solution

Capture Process

The firm already had several training rooms configured with the required audio/video equipment. The training rooms were modified with additional equipment to handle the encoding and publishing of the content. By taking the microphone and video camera and connecting them to a digital capture card, the video was encoded on the fly as the subject matter expert was presenting. The diagram below shows the configuration of the training room.

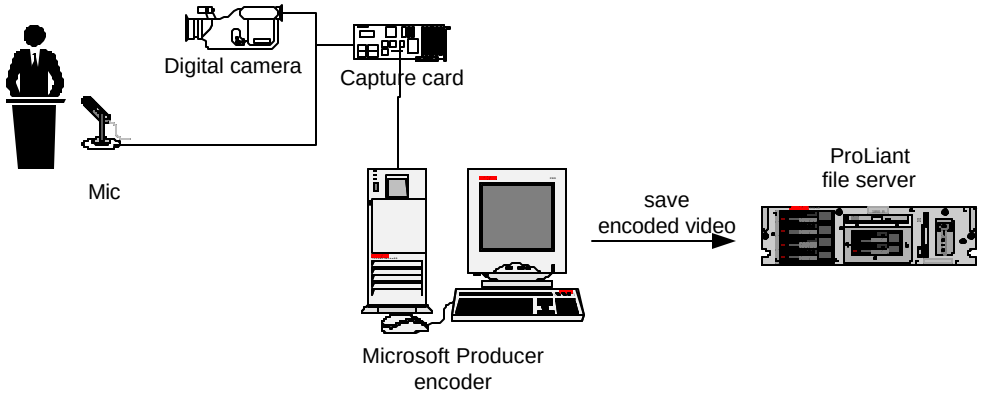

Figure 2: Video capture

Note: See Capture Options on page 28 for more information on capturing content and encoding.

Scenario #2: Medium Enterprise

Business challenge: An insurance company with 10,000 employees across 15 locations specializes in automobile and homeowners insurance. The company's primary focus is handling claims and providing a positive experience for customers while going through the claims process. The insurance company would like to distribute best practices for claims agents who are on the road and the claims counselors located at the regional offices. The best practices are small 2-5 minute video segments used to demonstrate how the agents and counselors should behave when dealing with specific situations.

The company also needs to implement a solution to support live broadcasts for corporate communications. The live broadcasts will include executive communications and previews of company television ads. Allowing employees to preview the company television ads has proven to build company loyalty and align employees with the external client message.

Opportunity: Standardize on best practices by more effectively training employees on complex subjects such as social interaction, customer service and conflict resolution. More effectively communicate and align employees with company goals.

Solution: The insurance company implemented a content distribution network (CDN) to manage the distribution of video segments to all claims agents, along with a custom module developed for Microsoft Exchange to coordinate sending out calendar requests for viewing the content. The Exchange scheduling component helps push reminders out to users, assuring a high degree of online attendance. Users can also browse the company intranet for content.

Content Distribution Network

To facilitate the broad distribution of content, a content distribution network was implemented across all sites. The CDN was designed using edge caching appliances in a forward proxy configuration. A forward proxy involves the cache appliances being used to accelerate client performance by placing the content closer to the client on the high speed LAN. Content can be proactively pushed out to the edge in anticipation of a large number of users requesting the content. Content that will not be requested by a large number of users can be kept on the origin server until it is requested by a remote site. Once the content has been requested, it will populate the cache appliance local to that site and user.

The design of the CDN assumes a maximum user concurrency of 10% for all media services. This means that if a site has 1,000 users, a maximum of 100 users at any given time will request content. This concurrency only applies for on-demand scenarios. Ultimately, the edge cache devices are limited by total bandwidth throughput, not user connections. 50mbps and 100mbps are typical for cache appliances total capacity. In this design, 50mbps appliances were placed on the edge, able to deliver a maximum of 500 100kbps streams or 140 350kbps streams. The largest site of 2,500 users has two cache appliances and all other sites have a single cache appliance.

Live Multicast Support

The network was enabled to support multicast traffic. This makes it possible for a single origin server to propagate a single multicast stream throughout the corporate network at 128kbps or even 300kbps. Since there is only one stream that services all workstations, the bandwidth utilization associated with a broadcast is less than .3% for a 100mbps network, when at 300kbps and less than .1% when at 128kbps. For those sites that do not have multicast support, the CDN provides live stream splitting. The CDN multicasts or unicasts the broadcast stream from the central origin server to the edge cache appliances. Once at the edge, the local users view the stream via unicast. This type of stream splitting pushes the unicast streams to the edge, which in aggregate can be bandwidth intensive.

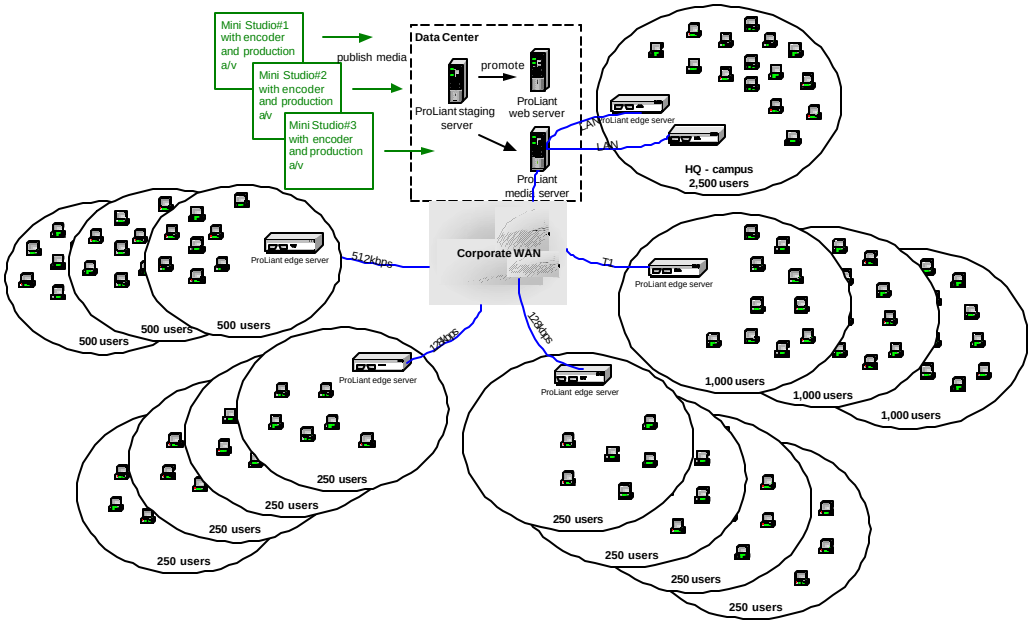

Figure 3: CDN for medium enterprise

Microsoft Exchange Scheduling

Instead of implementing a media portal to access content, it was decided to leverage the Exchange messaging platform to push out and manage requests to users. Exchange was selected because the users of the system were mainly task workers and were not used to browsing an intranet for materials. Users wanted an application they were familiar with. The training group responsible for making sure users view the various courses sends batch meeting requests to users via Exchange a few days in advance of a scheduled course. The meeting request has a brief description of the content and a URL that brings them right to the website hosting the content. Users can accept or decline the requests, depending on their availability. Over the course of a week or so, users are scheduled to the view the content online. When it comes time to view the content, the meeting request will launch a reminder 15 minutes prior to the start time with the URL to get them right into the event. This solution, getting end-users to access content, also works very well when the company needs to get a large number of users to view content in a scheduled broadcast mode.

Capture Process

To capture the content on a regular basis, the company built three mini-studios at their three division offices. The mini-studios cost approximately $40,000 each to build and are end-user friendly so virtually anyone can reserve the studio to create rich media presentations. This enables the business units to create content as they need it to support their business. The studio schematic is shown below.

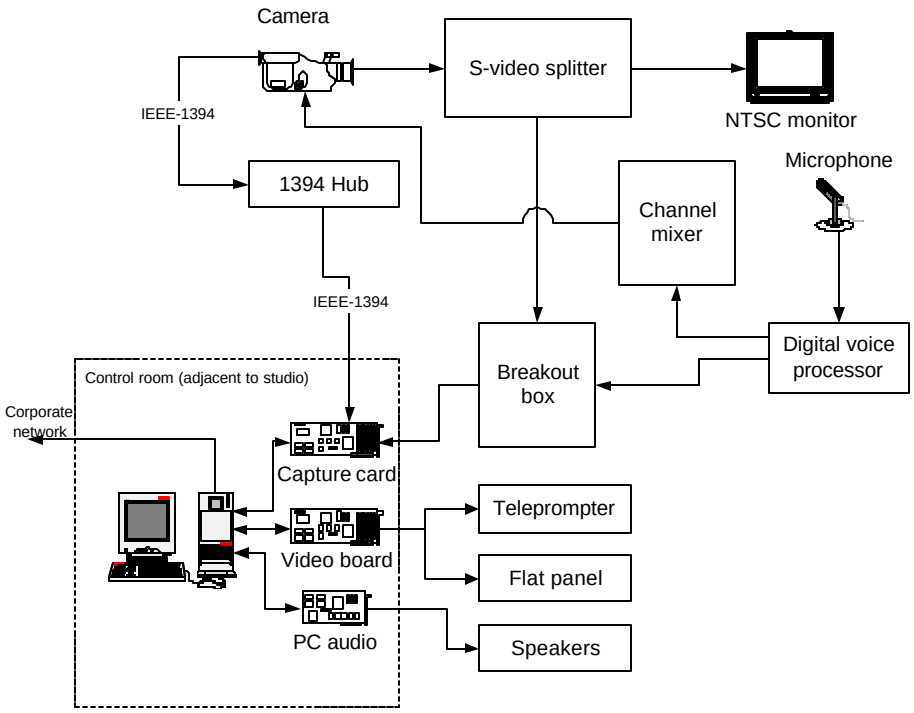

Figure 4: Studio design

Scenario #3: Large Enterprise

Business challenge: A large manufacturing company with 30,000 employees across 100 locations needs to deploy a solution to capture and distribute product information as well as live corporate broadcasts. The solution will be used for product launches, general product information for internal groups as well as for customer information. Reaching customers will require distributing the information external to the organization via the Internet. Customers will be able to go to the public corporate website and view product information. The company also wants to leverage their infrastructure investment for live executive broadcasts to deliver quarterly earnings reports and general HR communications.

Opportunity: Improve general awareness and knowledge of products and align employees with strategic business initiatives by producing and delivering a higher quality experience. Deliver a consistent and compelling message to customers via rich media presentations. Deliver consistent corporate communications more frequently via the corporate intranet.

Solution: The company took a comprehensive approach to solving the problem and initiated several solution teams to deploy a global CDN, develop business processes for creating business content and design a studio to create high quality audio and video content. The three solution teams worked together to integrate with one another and implemented an end to end digital media infrastructure and studio with all of the required business processes and work flow to institute the solutions log term.

Content Distribution Network

The company implemented a multi-tier CDN to manage the global distribution of web and media content. The network was designed to have a core network and an edge network. The core network is responsible for distributing content through the core network from where the outer edge cache appliances can request media and web content. By pushing the content to a mid point in the network the network congestion at the central origin servers is reduced. The network is too

large to accommodate all cache appliance requests directly from the origin servers. This type of tiered distribution also enables different parts of the network to be managed with different policies.

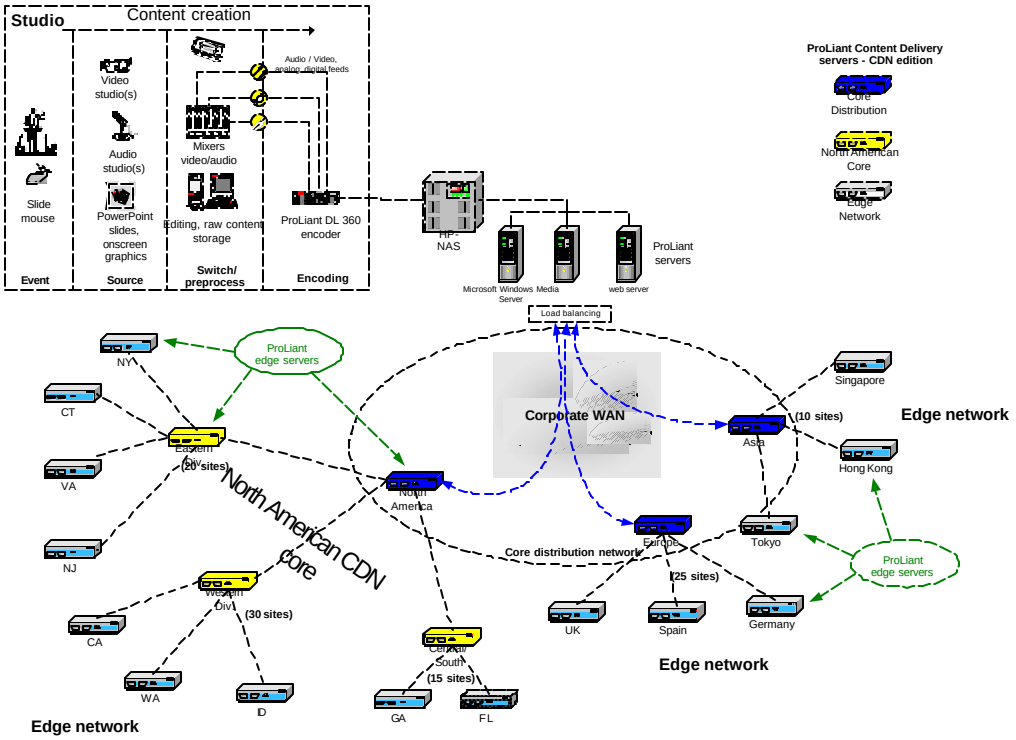

Figure 5: Large organization

media portal

To make the content easy to browse and help employees select the appropriate product information, a media portal was implemented. The media portal resides on the central origin server and its content is indexed and searchable via key words. All the content has a thumbnail of the opening presentation, showing the speaker along with a 50 word abstract on what the presentation covers. All of the content is categorized into line of business tracks.

The marketing managers are able to publish rich media presentations by using an authoring application that synchronizes the video and slides. Once the presentations are complete, the manager publishes the content to a central staging server for testing and approval. The entire process can be executed by the marketing managers, eliminating the need for special IT staff.

The media portal acts as a central catalog for all media presentations, where employees know they can search for archived corporate communications or the latest presentation from various product managers. The media portal is also integrated with the corporate intranet to simplify navigation and search.

enterprise digital media infrastructure

Before starting a detailed design, documenting the solution requirements serves to drive the architecture and design decisions. The infrastructure design can be separated into the following areas:

- Studio integration and content creation
- Content distribution and management
- Media portals and application integration
- Network (LAN/WAN)
- Security
- Systems management

Overlaying all of these design elements will be considerations of scalability and performance, which will also be addressed in this section.

design principles

When designing an infrastructure to support digital media, consider these design principles:

- Demand for digital media will increase significantly over time
- Streaming may not be mission-critical now, but it will become critical to the business as communications converge on IP networks
- This technology is strategic, so it is worth taking a long-term view when making decisions

future growth

The most common demand for digital media solutions comes from corporate communications and e-learning initiatives. Regardless of the specific business requirements, it may be beneficial to take a broader view of these requests with the expectation that once the organization experiences digital media, other business units will look to leverage the infrastructure for other business applications. Try to position the network and infrastructure to easily adapt to future requests and growth.

data point: U.S. spending on e-learning is expected to reach $14.8 billion by 2004, according to International Data Corp. research.

Design Elements for Building a CDN

The distribution and management of digital media content is one of the most important elements of an enterprise digital media solution. This is often considered the greatest barrier to implementing these solutions. Creating an appropriate architecture will require the involvement of multiple IT groups including the networking and the server groups.

Reviewing some basic concepts associated with the distribution of media content will help to make clear what elements are involved in CDN design. For the purposes of this paper, the CDN is internal to the organization, and is deployed as part of the corporate infrastructure.

origin vs. cache

Origin servers are appropriately named; they are the source or origin of the digital media content. Having few origin servers and locating them centrally simplifies the management of content. Windows Media servers on ProLiant platforms will act as an origin server, sourcing content for distributed caches. Media content cannot originate or be sourced directly from a cache; the cache must be directed to get content from a Windows Media server.

cache vs. proxy

A proxy is a server agent that responds to specific client application requests on behalf of the client. The proxy can also be considered an application layer gateway. A cache contains content that has been positioned away from the origin server closer to the client to accelerate the delivery of media content. These terms are widely abused and are often used interchangeably, creating

confusion. Simply put, a proxy is a server process that acts on behalf of the client that may or may not cache content depending on the business rules. The cache stores content on the edge of the network.

ProLiant edge servers can support various network configurations to build-out an enterprise content delivery network (CDN). The server can be deployed as a forward proxy or transparent proxy for client acceleration or as a reverse proxy for server acceleration. (See forward proxy and reverse proxy under Network Topologies below.) A forward proxy requires that the proxy be configured on the client, while a transparent proxy leverages the use of WCCP 2.0 on Cisco routers to redirect origin server requests via local cache servers.

anatomy of a CDN

When designing a CDN, there are three network layers. There is the core where the origin servers and source content reside, there is the core distribution hierarchy where content is moved from the center to the edge, and finally there is the edge where clients access content. Simple CDN solutions will move content from the origin servers directly to the edge. Larger global networks will use a two-tier distribution hierarchy before storing content on the edge. The ProLiant server can support simple content caching as well as multi-tier hierarchies for larger scale deployments.

load balancing/clustering

Load balancing occurs when client requests for content are distributed across several servers. This can be done at a central location in a reverse proxy configuration or on the edge in a forward proxy or transparent cache configuration. An IP load balancing solution can distribute load across several ProLiant servers when extreme scalability is required. IP load balancing will provide a single IP address for requests and will redirect the request to the most available cache device. So if an appliance goes offline, the automatic failover redirects traffic to an available appliance for a seamless transition with no downtime.

UDP vs. TCP

When designing a digital media infrastructure it is important to understand how to optimize performance and impact on the network from a protocol standpoint. Windows Media is streamed over the MMS protocol or HTTP. MMS is a Microsoft proprietary protocol and it provides the signaling from the client to the server. The data portion of the stream is delivered over MMST or MMSU. MMST is TCP-based and MMSU is UDP-based. It is important to understand the distinction between TCP and UDP. TCP is a reliable transport and provides a dedicated connection between the client and server. The delivery of each packet is acknowledged and checked for validity. Although TCP is trusted, it has more overhead which impacts the network.

UDP is considered an unreliable transport. UDP packets are sent on a best effort basis and lost packets will not be retransmitted. This may appear to be less desirable; however UDP is the preferred protocol for streaming media. Since streaming content is time-sensitive, it is preferable to play the content immediately even if a few frames are missing. To wait for TCP to retransmit and sequence the packets would create undesirable delays. Since UDP does not have the error correction or session management, there is less overhead and less impact on the network. For intranet-based solutions, UDP should be the preferred choice for all streaming media. If the traffic must traverse firewalls, TCP or HTTP might be a requirement.

bandwidth management

Windows Media and Windows 2000 bandwidth management can be leveraged in conjunction with other bandwidth management policies and QoS systems, such as network QoS, CiscoWorks2000 QoS Policy Manager, and Content Flow Monitor.

Bandwidth management can be achieved by implementing policies that leverage non-business hour network availability or only trickle feed distribution limiting the network usage during business hours. Additionally, multicast and cascaded caching devices can provide cost effective use of bandwidth.

Network transport support will leverage predominantly IP-based protocols, including, in order of

preference: multicast, unicast UDP, TCP, and HTTP. When possible, multicasting will provide optimum bandwidth utilization for content distribution.

multicast

Multicast traffic allows multiple clients to view the same stream and can scale independent of the number of clients. Multicasting, when implemented on the network is a very powerful tool for live broadcasts and distributing content. Multicasting must be enabled on the routers and in most cases switches are enabled in their default mode. Multicast has the benefit of allowing a large number of users to tap into one stream, conserving network bandwidth and reducing server load.

Routing: To turn multicasting on, a 'multicast routing' protocol must be selected. This can be PIM (Protocol Independent Multicast), sparse and/or dense. Similar to unicast routing protocols OSPF and RIP, the multicast routing protocol is responsible for helping to propagate the active multicast groups and moving the multicast stream to network segments that have requested the stream.

Multicast groups: Multicast addresses start at 224.0.0.1 and goes through 239.255.255.255. The addresses starting at 224.0.0.1 through 224.0.1.255 are considered 'well-known' or reserved addresses and should be avoided when issuing multicast addresses. It is recommended to select a narrow range of addresses and leverage MADCAP to automate the management of these addresses. MADCAP is similar to DHCP, but for multicast addresses. Windows Media Services 4.1 does not support MADCAP, so the multicast addresses for Windows Media server must be manually configured.

Note that if client computers are on shared network segments, (hubs), those multicast streams will be seen by all clients, effectively reducing the total available bandwidth for everyone. If the clients are on layer 2 switches the impact can vary. Layer 2 switches normally propagate by default, even those that have not joined the multicast group. Several switch options are available to manage this depending upon the switch model. IGMP snooping, CGMP and 802.1p configure the switch to only send the multicast traffic to ports which have requested it. Reviewing the switch documentation is helpful before enabling multicasting. IP multicasts do not work well on token ring networks and it is recommended to avoid enabling multicast on token ring.

WAN considerations: When enabling a multicast across a WAN, it is important to understand the impact on the logical network as opposed to the physical network. A frame relay network can be viewed logically, the frame relay cloud and the individual permanent virtual circuits (PVCs). In the model where multiple PVCs come together into a central data center, those PVCs will terminate on a single physical port. It is common to have the aggregate bandwidth of the PVC exceed the total port speed. This can be done because statistically the aggregate bandwidth of the PVC will rarely exceed the port speed. However, in a multicast scenario the PVCs will be treated as individual WAN circuits requesting the multicast stream, possibly exceeding the total port speed. For WANs with dedicated circuits, this is not an issue.

other tips for multicasting

- Watch for network loops, these can be reduced by proper value of TTL (hops) to max within a given enterprise.
- Multicast is UDP-based and some packet loss is normal.
- A low amount of packet duplication is normal for UDP networks. Multicast is an intentional packet duplication, but duplication over the network segment can indicate a problem.
- Administrative scope range is 239.0.0.0 to 239.255.255.255 used for private networks.
- MCAST.EXE is a utility provided by Microsoft to test multicasting. It simulates a multicast server and destination, exercising the routing protocols to propagate the multicast stream across the network.

network topologies

A ProLiant server can be implemented as a forward or reverse proxy. This simply defines where the proxy cache is located. In forward proxy, the cache is placed closer to the client, in an attempt to accelerate client performance. In a reverse proxy, the cache is located closer to the server, basically front-ending the origin servers. This design is intended to accelerate performance for the servers.

Forward proxy: When designing an enterprise CDN, a configuration intended to distribute content to the edge in order to improve the client experience and reduce network bandwidth requirements is considered a forward proxy. A forward proxy configuration places the content closer to the end user, better leveraging the LAN speeds of the network for higher quality content. Internet service providers like Speedera Networks and Akamai have essentially built out global networks for content delivery using a forward proxy configuration by placing edge caches near content consumers. This model is also best when content is being pushed out to specific groups of users who are specifically intended to receive content, like training and corporate communications.

Reverse proxy: A reverse proxy is intended to accelerate the server performance without specifically knowing which clients will be requesting content. Reverse proxies are often used when hosting an Internet site when the only part of the network under your control is the immediate server environment.

A reverse proxy topology can make sense in a large enterprise when content is being hosted by a specific division or group, which does not have the ability to place edge devices near the client. For the largest of organizations with hundreds of thousands of employees, a reverse proxy design is often the only economical decision as the networks are built out to support a forward proxy approach.

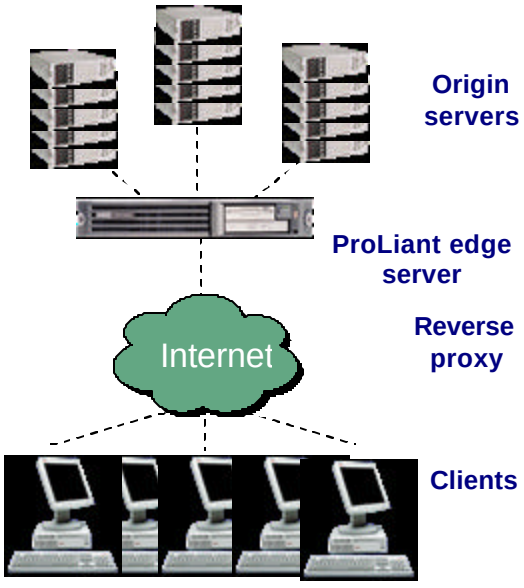

Figure 6: Reverse proxy

Transparent caching: Transparent caching is when no client configuration is required to redirect client requests to the local cache. This can be achieved in several ways, one of them being the WCCP (web caching communication protocol) from Cisco. WCCP 2.0 is a protocol that runs on Cisco routers and filters on the IP address and port. WCCP keeps track of the cache

servers connected to local network segments and the origin servers they proxy. When clients request content from those origin servers, the router will redirect the client request to the local proxy. The proxy will then serve the content to the client if it is in the cache, or request the content from the origin server if it has a cache miss. WCCP makes it very easy to implement edge caching for large networks, without having to make client changes. Without WCCP, each client will have to be configured with the IP address of the local proxy.

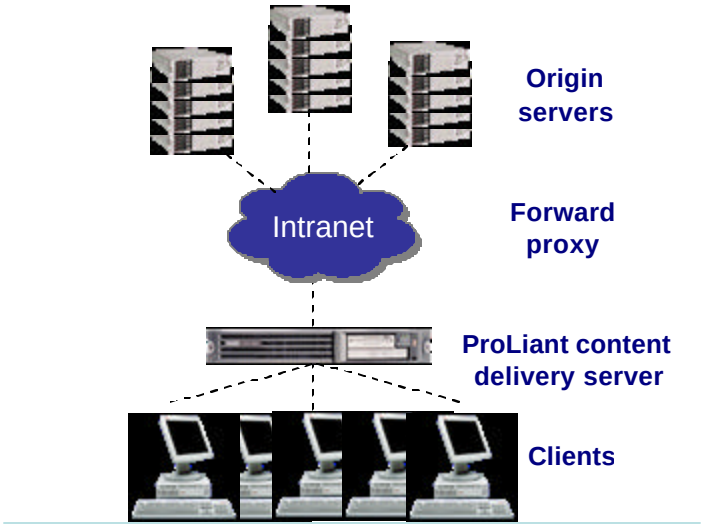

Figure 7: Forward proxy

Web and media content flow: When users view rich media content there are essentially two traffic flows of content from the server to the client computer; streaming media and web content. For example, rich media presentations created with Producer are made up of two types of content, Windows Media video and other web objects like jpegs, HTML and asp files. The Windows Media video and the web objects are all integrated at the client's browser for a synchronized rich-media experience. The web content and the streamed content have different distribution requirements and are often managed on separate origin web servers and origin media servers. The distribution will also vary depending if it is an on-demand or a live broadcast.

For live and on-demand streaming content, the general application process is:

- Promotion - from an initial promotional mechanism (email, Outlook calendar, intranet), the user is directed to open the URL for the Media Guide where all the content and events are catalogued and listed.
- Caching - all static content requests will be serviced by cache hosts. Static content is usually considered content that is reachable via static URLs. Once the system has determined a user's requested content fidelity, all subsequent requests for content are serviced by cache hosts. Multicast media is not cached. Media in on-demand content is considered cacheable.
- Content distribution - all cacheable content is published to origin web and media servers and is dynamically distributed via the first browser request to downstream cache positions in data centers. For the first user content request received by the system, the cache host will fetch the content from the origin servers for that page or page element (i.e. PowerPoint slide flip). The user's browser never directly requests the content from the origin web server. The first request indirectly goes to the origin web server via a request from the cache host. All subsequent requests for the same content are serviced by the cache.
- Initial user requests - the user selects a live event or on-demand content, which loads on their browser the web content of the event, containing all the web elements of the content's base page. Live event users' arrivals at the starting page are predicted to occur over an average of 5 minutes before the event and five to ten minutes after the event has started, distributing

the infrastructure loading over time. This content is primarily HTML, pictures and graphics. Once the base web elements have loaded on the browser, it is triggered to request media which, for a live event, begins to play from the multicast stream, or otherwise streams directly from the cache host.

- Encoding - to facilitate live event multicast convergence and prepare the encoding, streaming of empty content begins to allow the network and encoders to converge on the upcoming event start. For on-demand content, encoded content and web pages are edited for quality and published to origin servers for distribution.
- Waiting for event (live) - users' embedded player, joins a multicast group and awaits the event's start, which is triggered by the first synchronous script update in the stream (i.e. the first PowerPoint flip).
- Event begins (live) - usually triggered by the studio, streaming begins. Frequently, the encoding of four streams occurs simultaneously, as determined by fidelity requirements and these feed the origin media and web servers.
- Synchronous requests network and server loading (live) - an important point for performance engineering is to be aware of the dynamics of the highly synchronous system and network events around streaming live content. A popular live event may trigger thousands of users to request the same content at exactly the same time, causing high peaks in infrastructure loading. The digital media infrastructure effectively manages these peaks.

To understand the general sequence of events during live and on-demand content viewing better, the network flows are diagrammed and described on the next two pages.

Live Event Network Flows

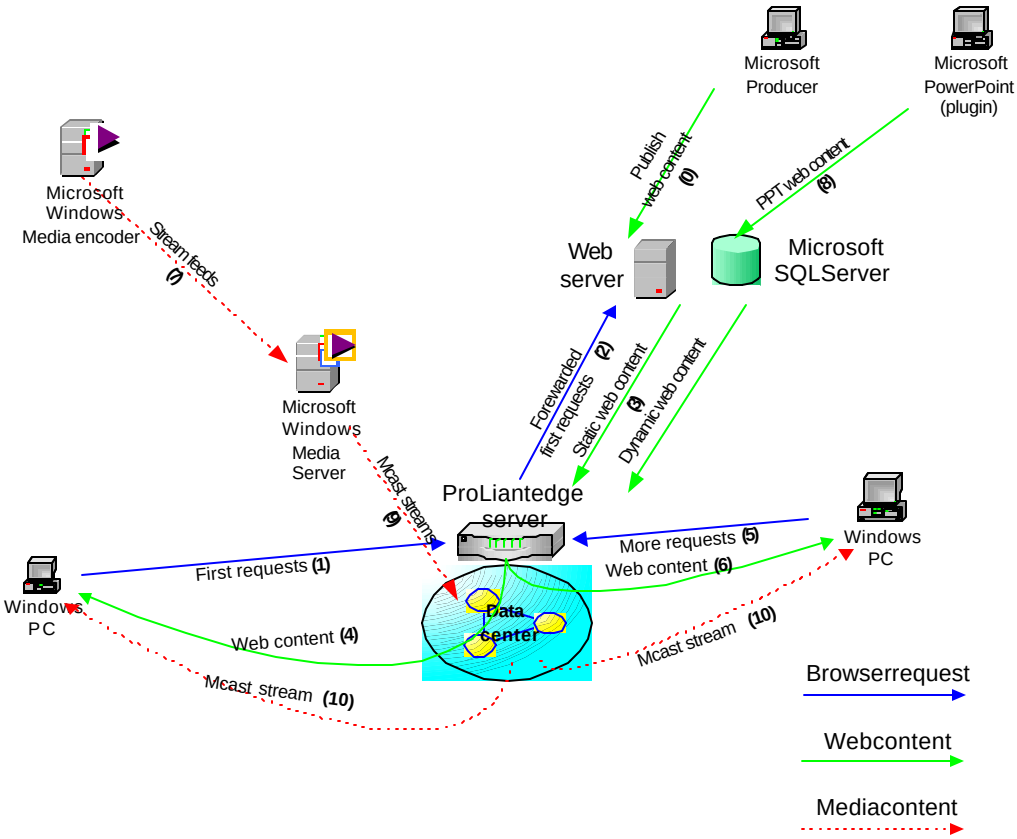

Figure 8: Live network flows

On-demand Network Flows

1. Web static content is pre-published to origin web server.
2. At the event pre-start, user's browser requests static page elements of event's content.
3. For each site, nearest cache receives request and forwards it to origin web server (or parent cache host).
4. Origin web server responds to request of first user, content is cached at cache host for future requests (for same site/fidelity).
5. Cache host forwards response back to first requester. Embedded player joins multicast.
6. Subsequent requests for more of the same static content from other same site browsers or downstream cache hosts for content of first requester (1). Embedded players join multicasts.
7. Cache host services request and does not forward request to origin. Although not shown, technically nearest cache host with content will service requests.
8. A few minutes prior to event start, encoders are primed (triggered by studio), to begin feeding media content to origin media servers. Priming allows systems and network multicasting to commence.
9. Event begins with the first PowerPoint slide being displayed. Synchronization of web and media content is established. Studio initiates camera/microphone feeds to encoders at airtime.
10. Origin server begins to multicast event.
11. Browsers receive multicast stream (via pending mcast prior request), which triggers subsequent synchronous web requests at each PowerPoint flip.

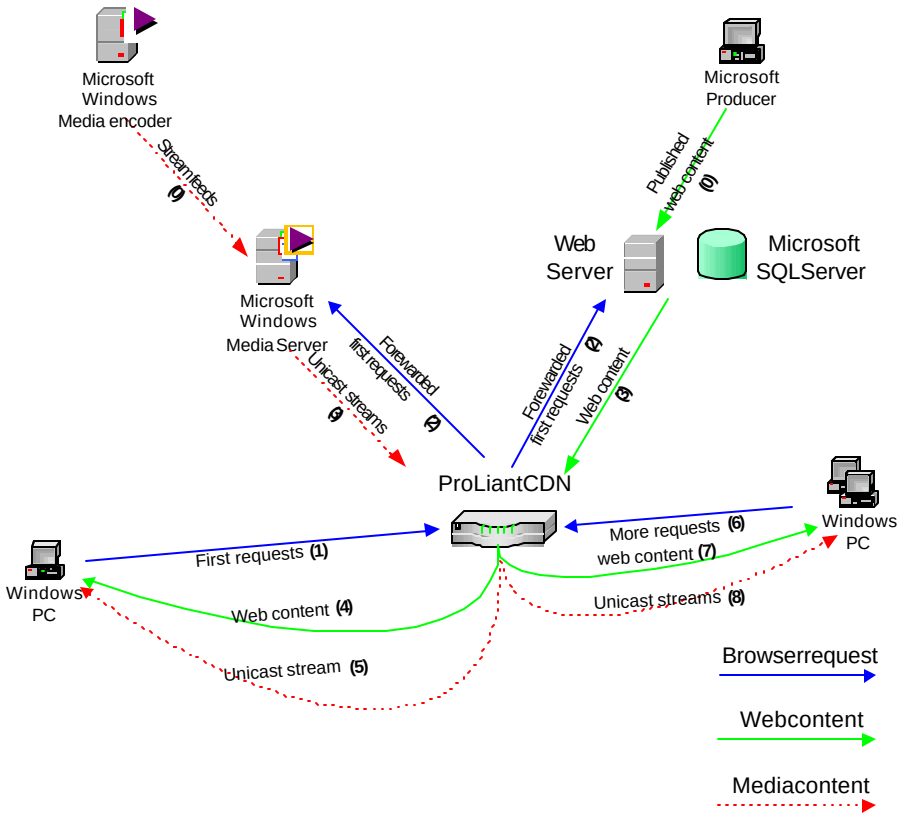

Figure 9: On-Demand network flows

1. All web and media content (usually edited) is pre-published to origin web and media servers.
2. User's browser requests web elements of content.
3. For each site, nearest cache receives request and forwards it to origin web and media servers (or to parent cache host).
4. Origin web and media servers respond to request of first user, content is cached at cache host for future requests (of same site/fidelity).
5. Cache host forwards response back to first requester. Embedded player is instructed to play media content.
6. Media streams to first requester from cache hosts.
7. Subsequent requests are made for same content (web and media) from other browsers of same site/fidelity or downstream cache hosts for content of first requester (1).
8. Cache host services requests and does not forward requests to origin servers if content already found in cache. Although not shown, technically the nearest cache host with content services requests.
9. Media streams to subsequent requesters (of same site/fidelity) from cache hosts. Unlike synchronous live events, IP unicast is used for media, since each user can view content anywhere, anytime.

where to position the ProLiant server used as edge servers

For the delivery of intranet-based content, the edge server should always be placed between the client and the origin server. For best results the server should be placed on the local network segment of the client. However this may dictate too many cache devices, so the design of the CDN needs to balance putting the cache close enough to the client to deliver a quality experience with the total number of edge devices that is desirable. Placing the media content on a local cache not only saves network bandwidth across the WAN, but it also eliminates latency and packet loss from pushing content across congested network segments, improving the user experience.

cache control

An edge server allows for very granular cache control, where administrators can specify rules for what content is cacheable. Administrators can also control 'freshness' with features like 'never-cache', 'pin in cache' and 'revalidate after'. To help eliminate the caching of duplicate content, the appliance server has content fingerprinting, which can intelligently detect duplicate content independent of URL. The edge server also obeys cache directives in HTTP headers.

content management

Content management is another very important dimension to an enterprise digital media infrastructure. Content management is a somewhat nebulous term, but for the purposes of this paper, it relates to the higher level tasks required on top of the basic distribution of content. This means tasks such as:

- Reporting
- Policy based management
- Indexing

Reporting: As an organization produces significant amounts of content and begins to distribute content around the enterprise, it is essential to keep track of this activity. Reporting relies on log files stored on Windows Media origin servers and ProLiant servers. Edge servers can employ extensive Java-based charting modules including MRTG graphing for quick and easy analysis of activity. This is very helpful when identifying the most popular or least popular content and for

internal billing and charge-backs.

Policy-based management: Windows Media and HP both support policy-based management, which enables policies to be defined in Active Directory that control which users and groups have rights to view content. The ProLiant server will pass through authentication challenges from the origin server to the client. Policy-based management is a powerful way to assign content to user groups and have the ACL follow the content through the CDN. This makes distributing and securing sensitive content much easier. For large organizations, with lots of content, policy-based management is helpful for streamlining administration and improving security.

Media portal considerations

Once the digital media infrastructure is in place and operational, it is time to look at how content is accessed and managed. Content will come from many sources and will come in all shapes and sizes, video, audio, HTML, PowerPoint presentations, and static images. Seamlessly connecting users with the content is a challenging aspect of enabling the digital media infrastructure. A central Web location is often a requirement where users can locate and search for the content they want to see.

One of the most effective means of displaying content for users to select is through a media portal. This provides a centrally controlled destination for publishers and users of digital media content. It defines a standard structure for publishing content, a central location for storing content and a single point for end-users to find and access content. Microsoft and HP provide a suite of products to build a robust media portal solution that fits seamlessly into any digital media infrastructure.

A media portal solution can be developed using standard platforms such as Microsoft IIS and SQL Server 2000 using a multi-application layer model, web layer, middle layer and database layer. Solutions can also be developed using portal products such as Microsoft SharePoint™ Portal Server. Platforms like SharePoint provide much of the web interface and content management required making it easier to develop a media portal. Ultimately, it comes down to the application requirements, but SharePoint is a good place to start.

Building a SharePoint media portal server

SharePoint Portal Server, in addition to its enterprise-level search and mainstream document management capabilities, provides a rich out-of-the-box web portal.

SharePoint Server creates a web portal, called the dashboard site. This site offers centralized access for finding and managing information in the SharePoint Portal Server. The dashboard site can provide web-based access to information stored both inside and outside the enterprise, and allows users to find and share documents, regardless of their location or format. The dashboard site offers a single, customizable source for accessing information drawn from a wide variety of content sources, including Microsoft Office documents and website content, and it preserves document and content security. The SharePoint dashboard site is easily integrated into the CDN, providing end-user portal and content origination functionality.

An enterprise might use the dashboard to incorporate various tools such as Microsoft NetMeeting, to allow people to share applications and collaborate face-to-face over the internet, or via an enterprise intranet. It might choose to integrate Windows Media Player, or to deliver dynamic digital media content through Producer. Corporate communications, online training materials, and business broadcasts from the Internet or from a corporate intranet site can also be hosted here.

The functionality of the SharePoint dashboard portal is provided by web parts components. These enable the creation of reusable components of web-based content, with standard sets of properties that control how the web part is rendered in the digital dashboard. Web parts components allow developers and users to define, manage, and control content within the enterprise portal. Web parts components include:

- Basic Metadata - information such as title, description, and the date the web part was last modified.

- Content - information about the type of content the web part contains, and the source from which the web part gets its content. For example, the content can be embedded in the web part itself, or it can come from a URL on the internet.
- Appearance - specifies the height and width of the web part.
- Execution - specifies how often the content is refreshed, and whether the web part is isolated from other web parts in the system.

Web parts can deliver any web-based content such as XML, HTML or script. Developers using web parts can build dashboards that allow extensive user customization, including the ability to change dashboard layout, add or remove content, or select a personalized style.

Since the dashboard site is made up entirely of web parts, it can be easily customized to suit business needs. Specialized web parts can be obtained from third parties, or may be built using familiar tools like Microsoft Office XP Developer. Because web parts share a common schema they can be easily reused across dashboards, and catalogs of web parts can be created that allow system administrators to manage and distribute digital dashboards throughout the enterprise.

Security

As with any enterprise solution, security is one of the most important aspects to be addressed. Security considerations guide how the solution is constructed, as well as how content is controlled and who has access to it. Incorporating Windows Media Services into the solution provides built-in security capabilities and the services seamlessly integrate with Window NT Server 4.0 and Windows 2000 Server for controlling access to digital media content.

Windows Media and Windows 2000 support other security features and extensions such as auditing, Active Directory policies, security management, event notification, intrusion detection and custom security extensions.

Three key aspects of security are authentication, authorization and rights management.

- A Windows Media-based digital media infrastructure easily leverages authentication services found within Windows 2000, which supports the most common types of authentication.
- Windows Media provides very granular and efficient management of content authorization through tight integration with Windows 2000 object access control list as well as other customized access control capabilities.
- Digital rights management is easily achieved through the addition of Windows Media Rights Manager software.

authentication

Authentication is the means of validating a user's identity, typically by using a logon ID and password. As a user attempts to log in, their user ID and password are used to look up their specific security privileges and identify who they are. Authentication is not enabled by default when Windows Media Services are installed. There are typically three authentication schemes available in the Windows Media Administrator.

- Windows NT LAN Manager (NTLM) authentication and NTLM account database authentication.
- HTTP-BASIC authentication and NTLM.
- HTTP-BASIC authentication and membership services.

authorization

Authorization is a security feature for granting or denying access to protected resources based on the privileges of an authenticated user. For Windows Media Services, protected resources include content or media for which you want to control access, such as real-time content. Authorization works hand-in-hand with authentication, which confirms the user identity. In general, a user who

fails authentication is not allowed to access the requested resource.

Security authorization is typically set at the Windows Media origin server by established content management and/or security policies. These authorizations are then propagated by the content delivery infrastructure. Security authorization (access control) is usually done at the content object level. This enables individual authorization structures associated with content to be replicated across to any number of downstream media servers, caches, and web servers.

By enabling access control list (ACL) checking, a Windows Media server can check the access privileges of an instance of Windows Media Player against the access rights for any directory, .asf, .wma, or .wmv file or stream. In addition, you can restrict access to multiple files simultaneously by assigning an ACL to the directory where the files are stored. You can set an ACL for stored content on an NTFS file system partition by assigning an ACL for the file or the physical directory where the file is stored. Furthermore, you can set an ACL for on-demand content stored on a FAT partition by setting an ACL for the registry key associated with the on-demand unicast publishing point. You cannot use ACL checking without enabling an authentication mechanism because unknown users cannot be authorized. A Windows Media server can also grant or deny access to streams based on the IP address of the client.

digital rights management (DRM)

Another aspect of managing digital media content is handling the usage rights associated with a specific piece of content. With the increase in peer-to-peer file sharing and content piracy, the issue of copyright protection is a high priority for content creators, owners and distributors. Microsoft provides a robust means of DRM via Windows Media Rights Manager software.

DRM encrypts content and creates a license key which is required to access the content. Business rules are created in association with the license keys which govern the rights users have to access the content. The content can be distributed freely. When users access content there is an acquisition URL which points to the license key, at that point, the user is authenticated, and their rights are determined. The Windows Media Rights SDK provides the leading DRM platform for secure distribution of content.

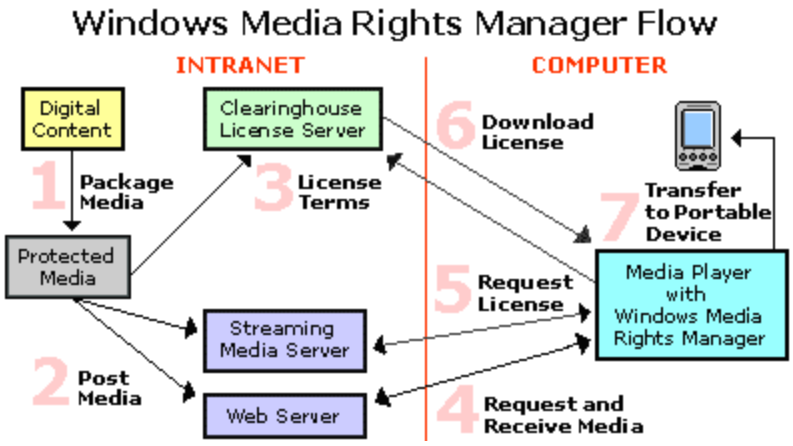

Figure 10: Windows Media DRM

firewall traversal

Firewall traversal for Windows Media can often leverage existing infrastructures and methods previously put in place for other internet protocols such as HTTP and FTP. Windows Media players may access Windows Media servers via HTTP as a default, often requiring little or no configuration changes to existing firewall policies. Other firewall traversal methods are also available over UDP and TCP, which offer better performance, but usually require more complicated firewall traversal policy considerations and changes.

Providing firewall traversal support for Windows Media is well supported by most enterprise firewall products and is secure and reasonable for administrators to manage. Integrating firewall traversal into internet proxy and CDN strategies can offer substantial performance improvements. Additionally, it will help reduce the overall load on the firewall and Internet gateway services.

Generally, two scenarios cover Windows Media firewall traversal needs:

- Providing secure access to Windows Media content found outside the enterprise network
- Allowing users outside the enterprise network access to Windows Media content found within the network

firewalls and Windows Media MMS

Windows Media client players normally access the Windows Media server through Microsoft MMS protocol. MMS attempts to connect to the media server in the following order: IP multicast, UDP, TCP and finally HTTP. This is the prioritized media transport method, with multicast being more network-efficient than UDP, that is more efficient than TCP, and so on. Real-time applications such as digital media normally don't use the additional error recovery and management of TCP and HTTP protocols. Firewalls supporting UDP for Windows Media can provide performance improvements over HTTP proxies, and improve user experience.

The protocols and ports used in MMS are:

- UDP: 1755 server destination port
- TCP: 1755 server destination port
- HTTP: 80 server destination port
- Player source port client port: 1024-5000 (dynamic)

Scalability and Performance

There are many constraints in the infrastructure which can affect overall performance. Key areas to be aware of are:

- Hardware (CPU, memory, disk I/O, net I/O)
- Bandwidth and concurrent usage
- Cache hit ratio
- Fault tolerance

Hardware Requirements for Digital Media Infrastructure

Component	Bottleneck	Comment
Encoding Servers	CPU and memory bound	Requires one CPU per capture card, scales linearly.
Origin servers	CPU and network I/O bound	
Edge server	Disk and net I/O bound	
Content server	Disk and net I/O and storage	Requires fast disks and good Net I/O throughput.

estimating maximum concurrent usage

It is important to determine the maximum number of users that will view on demand content concurrently. It is unrealistic to design a solution that can handle 100% concurrency for on-demand content, for two reasons. First, if all users are viewing content at the same time, multicast is a better option. Second, the number of times a solution will be required to handle very high levels of concurrency is too few to justify the investment in the infrastructure required to support this. It is reasonable to expect business users to spread their access to content over the entire

business day. Maximum concurrencies can range from 8% to 15% of the total user population for large organizations, depending on the company and business requirements.

At a site that has 1000 users with a maximum concurrency of 12%, the infrastructure would need to support 120 streams at any given time. If the standard bit rate of streamed content was 128kbps, then the total bandwidth utilization on the edge appliance would be 15,360 kbps (15.3MB). These numbers are important to understand when specifying how many cache appliances or media servers to deploy and where they should be located.

cache hit ratio

Cache hit ratio is an indication of how much content is retrieved from the appliance servers. This is an indication of how effective the cache is being in reducing bandwidth requirements. Cache hit ratios should be in the 90% range, depending on the nature of the content. Some dynamic web content will not be cacheable, which will drive the hit ratio down.

fault tolerance

In any production environment, some form of fault tolerance is necessary to prevent prolonged service outages. For digital media, fault tolerance can be achieved at the hardware level and in the design level.

At the hardware level, redundant network devices and RAID disk configurations contribute to a fault-tolerant design, but are common to most applications, not just digital media. Generally, the same fault-tolerant design techniques apply to digital media as to other applications.

At the design level, there are ways to build an infrastructure that can handle unplanned outages and still support your users. One technique is to use the Windows NT Load Balancing software (WLBS) with Windows Media Services in conjunction with ProLiant VIP failover capabilities. If one server goes down from a planned or unplanned outage, WLBS automatically removes the server from the WLBS cluster and redirects users to other servers in the cluster.

ASX files can also be used to provide fault-tolerance for users. If a user tries to connect to a server and the server is unavailable, they will be presented with an error. However, an ASX file can contain multiple REF entries in its URL to specify multiple locations where users can connect to a stream. If the connection to the server's first REF tag fails to connect, the Windows Media Player will automatically and transparently to the user, roll-over to the next server in the list and try to connect. If all of the servers fail to connect, then the Windows Media Player will return an error to the user.

Capture Options for Content Creation

From an infrastructure perspective, studios are a challenge to integrate. The challenges include the actual configuration of the studios themselves and the conversion of audio and video from analog to digital format. Many times large organizations already have professional studios for producing video content. In other cases, companies outsource their video production, or perhaps only have a minor investment in a camera and some lighting equipment. Regardless of the scale and level of sophistication, studio integration has to be addressed in order to create video content.

studio components

When designing a studio, there are three major design elements:

- Audio/acoustics
- Video
- Lighting

Audio is often assumed to be the easiest media component to capture or produce, but it is often

the most difficult. Ambient noise is difficult to eliminate, and steps should be taken to reduce it to improve audio quality:

- The studio should be insulated with a material like 'Sonex', to reduce noise reflection
- High-quality directional microphones will reduce ambient noise and are designed to cancel out background noise
- Use a mixer to compress the audio, reducing dynamic levels to an acceptable range and filter out noise

Producing quality video is generally straightforward and achievable. There are many choices of camera to use, however an entry-level digital video camera from SONY or Canon is a good choice. Quality video production is primarily a factor of lighting. Video quality can be improved by taking the following steps:

- Use high-wattage adjustable incandescent professional lighting
- Use dark colored walls to control light reflection and balance the overall lighting
- Use a tripod to hold the camera steady and allow for height adjustment
- Position a monitor directly above or below the camera, so the subject can read a transcript or PowerPoint slide as they are on camera. If the subject is six feet or more from the camera, the angle between the camera and monitor is small enough not to create a perceptible discrepancy in focus or eye contact
- Use an NTSC monitor for preview

For more information on implementing a mini-studio read, "Building a Mini-Studio for Creating Online Presentations with Microsoft Producer for PowerPoint 2002" from Microsoft.

modes of content creation

Professional - A professional studio is a broadcast-quality studio used to create and acquire audio and video. Professional studios are designed so that the authoring environment is completely controlled. Special lights, microphones, sound absorbing coverings and low background noise increase the quality of the studio output. Professional studios can be costly, and are usually built only when an enterprise regularly generates a considerable volume of video and rich media.

Enterprise - Enterprise studios are more common in companies. These studios typically serve multiple purposes, rather than providing a dedicated studio space. An enterprise studio might double as a shared conference room, or a small office that contains equipment specifically for creating rich media. An enterprise studio might also be equipped with A/V equipment, in a location to which users can easily gain access. These studios usually have a mid-range camera and microphone configuration that is easy to use and reposition, and they are configured so that users can easily find and use the equipment to generate audio and video.

Desktop – Desktop studios are by far the most common configuration found in enterprises today. Desktop configurations need little hardware, and can generate reasonable quality output. However desktop configurations are prone to many environmental problems, such as poor lighting, and unacceptable levels of background noise.

encoding servers

The Windows Media Encoder is available for download from www.microsoft.com and it is the primary server software used for encoding content into Windows Media format. Encoding content should be done by a high-end machine with enough available CPU and memory. A dual processor ProLiant DL360 G2 server with 1 gigabyte of RAM is ideal. Using one computer to do encoding will not be enough especially if a high volume of content is being created. Encoding servers fall into two groups, those used for on-demand content encoding and those used for live

encoding. For enterprise scenarios it is best to avoid using a workstation as the encoder, as the encoder should be freely available and able to be remotely managed.

Windows Media Encoder for on-demand content will be present in the studio and will be accessed regularly to configure and adjust source input. Encoding for on-demand content can be done as part of the post-production process, so the encoder can be located anywhere. The encoding process can also be done as the action is being captured. In this case, the encoders will have to be close to the audio and video equipment, so make sure that the studio environment can accommodate computer equipment.

Windows Media Encoders for live encoding have to be close to the A/V equipment but they have to be dedicated for encoding content for immediate live multicasting. When implementing a live multicast, reducing delay is essential, so the most powerful computer available should be used to keep the encoding as real-time as possible.

The encoder is a single point of failure, so care must be taken to ensure the encoding process is not interrupted, otherwise the live stream will also be interrupted. It is preferable to run dual encoders on a ProLiant server platform to eliminate this problem when the integrity of the stream is critical. The A/V source must be split and input into two encoders running on two different ProLiant servers. The two encoders then feed either a single origin server or dual origin servers on the ProLiant platform.

capture cards

A video capture card is needed to receive inputs from analog and digital sources. For example, the Viewcast Osprey line of capture cards supports input for composite video, SDI video and IEEE 1394. The Windows Media Encoder software will then take the digitized content from the capture card and encode it in Windows Media format. It is preferable to have one CPU per capture card, so a dual processor ProLiant DL360 G2 or DL380 G2 server can support two capture cards running concurrently.

fidelity

Content can be encoded at different fidelities depending on the business and network requirements. The fidelity controls the quality of the stream, for both audio and video. It can be controlled by changing the size of the video frame or window and the bandwidth of the audio. It is common to encode content at several fidelity levels to support a broad set of user requirements. Users over a dial-up VPN may only be able to support a 16kbps stream, while users on the corporate network can support 300kbps.

Consider the fidelity of the content as a way to control the quality of the stream delivered. Attempting to deliver a 300kbps stream over a 128kbps link would create a bad user experience. A 90kbps version of the same content would be much more satisfactory. The Windows Media Encoder supports the full range of fidelities from low-end audio only to high-end DVD quality video. Creating appropriate fidelities is a matter of knowing your network and your business setting.

preparing for a Microsoft Producer session

Producer allows users to combine audio, video, PowerPoint slides, HTML, and still images into a rich media presentation. Producer has several built-in Wizards that allow users to import content from a variety of locations. To create a Producer presentation, the user must first know what content and information they want to include in the presentation.

It is helpful to list the content required before launching the Producer session:

- PowerPoint presentation(s)
- Video
- Audio

- HTML
- Digital images

Archived audio and video content can be imported into Producer, and new audio and video content can be captured directly using the Producer Capture Wizard or the New Presentation Wizard. Content authors can synchronize audio and video with slides, HTML or images by using the Capture Wizard or the Synchronize Wizard, and can directly manipulate media elements on the timeline. These projects can be saved in pre-published form, and content can be added or edited later.

PowerPoint 'flips' in live broadcast

When producing a live broadcast with a presenter on stage or in a studio, synchronizing PowerPoint slides is often a requirement. Since the broadcast is live, interlacing the PowerPoint flips into the encoded stream is a bit more involved than when creating on-demand content. For on-demand content, this is done with an authoring tool, which embeds the HTML commands in the stream in a post-production process. For a live broadcast, the HTML commands to flip the PowerPoint slide must be initiated from the podium where the presenter is standing, and then embedded into the encoded stream in real time. There are two ways to achieve this:

- A person in the production booth of the live event can use the Windows Media Encoder to embed the HTML commands manually as they see the presenter change the slide.

Online Broadcaster for PowerPoint is an add-on tool that allows the presenter to click through a PowerPoint deck that will send DCOM calls to the encoder to embed the flips for the web viewing audience.

enterprise deployment

Deploying digital media capabilities presents many of the same infrastructure challenges networking groups have dealt with when implementing more traditional web services. However, these challenges are significantly increased because of the complexities introduced by the need for greater and more reliable bandwidth. The effects of network congestion and general latency on standard HTML content are not acceptable when streaming digital media. A careful and thorough deployment strategy will help to ensure a quality end-user experience. Here are some guidelines and principles to consider when embarking on a digital media deployment.

Most organization have their own established solution development processes so what is presented below is a general framework to ensure the business and technology objectives are clearly aligned, solution designs are fully considered and the production deployment of the solution is properly planned.

The project is often best defined and structured into two distinct phases:

- Phase A – Requirements, architecture and preliminary design
- Phase B – Final design, build, implementation and pilot

The project team, analysts and network designers, will work to define the system requirements and design the digital media infrastructure. They will work collaboratively with all parts of IT and the necessary business units to ensure that the solution adheres to current IT standards while achieving business goals.

Phase A

Phase A can be executed in three parts; systems requirements, conceptual architecture and preliminary design. The objective of Phase A is to define and document a preliminary design, which can be used to plan for the final design and deployment which is carried out in Phase B. It is usually best to conclude Phase A with a preliminary design so as to provide enough clarity on the final solution for planning, budgeting and resource purposes.

systems requirements

The requirements team will perform an assessment of the current network and determine the modifications required to create the digital media infrastructure. They will work closely with network administrators, infrastructure engineers, directory services, security and application development to capture the specific needs and constraints of each area. This is vital information when considering the design elements of the solution. Although the business is involved, depending on the nature of the solution and the type of content to be streamed, the majority of requirements gathering will be confined to the network and related information systems.

conceptual architecture

The system requirements are then used to produce a conceptual architecture for the digital media infrastructure. This defines the design principles and guidelines for streaming on the network and building out the infrastructure. The architecture will be an important resource for future design changes and enhancements.

The architecture is a necessary precursor for producing the preliminary design. It also ensures that the detailed designs adhere to the original principles and standards of the project.

preliminary design

The preliminary design is intended to provide a vision of how the final solution will be implemented and a plan for solution deployment. It is also needed to determine the hardware, software and services required to complete the deployment and estimating the total cost to complete the project. Based on budget approvals, technical direction, solution capabilities and overall approval to proceed, the detailed design and solution deployment is then performed in Phase B.

Phase A scope

Phase A operations typically consider the following areas:

- Network services
- Streaming services
- Content management and distribution
- Security
- Operations and administration

network services

Assess and design the necessary changes to the network required to support digital media. Address how the network will support live multicasts, distribution and the implementation of Windows Media and ProLiant servers for stream distribution and stream splitting. This will cover modifications to the existing network and any new infrastructure required. It will also include:

- CDN hardware-specific designs
- Protocol standards (multicast/unicasts, UDP, TCP, HTTP)
- Bandwidth management
- Capacity planning

streaming services

Design the components to deliver the core streaming services for the digital media infrastructure. Design and establish the configuration of Windows Media and ProLiant servers. The team will also assess and design the systems management components of the solution. The design will include the placement, configuration and systems integration of:

- Encoders (codecs)
- Origin server design
- Players and client configuration

content management and distribution

Design the systems required to manage and distribute digital media content across the organization. This will allow content to be managed centrally across the digital media infrastructure. This typically applies mostly to video on demand content and programmed broadcast content. These designs will be very dependent on the Windows Media origin servers, ProLiant server infrastructure and the network. Reporting should be addressed in this area to provide statistics on system usage ('Nielsen ratings'). Design elements will include:

- Windows Media server designs
- ProLiant server configurations
- System reporting
- Storage
- Fault tolerance
- Load balancing

security

Assess and design security requirements to provide the proper authentication and authorization throughout the system. Security may be a pervasive theme throughout the project, specifically looking to design a solution that considers:

- Directory services integration
- User authentication

- Authorization

operations and administration

Consider the IT operations and administration tasks required to maintain the digital media infrastructure. This should also be a pervasive theme in the project. The operations and administrations requirements of the solution must be documented and agreed to at a high level. This is important to consider so the recurring costs of the solution are understood.

- Reporting
- Training

Phase A deliverables

Typical deliverables at the conclusion of Phase A are:

- System requirements
- Conceptual architecture
- Preliminary design
- Phase B project plan with cost estimates

Review all project deliverables prior to completion. The architecture and design documents will contain logical and physical diagrams as necessary to describe the solution.

Phase B

Phase B will complete the project, implementing the production digital media infrastructure. The scope of Phase B is typically determined in more detail at the completion of Phase A, and is based on the specifics derived from Phase A. However, at a high level, Phase B will deliver:

- Final digital media infrastructure designs - these are the final designs that will be used to make modifications and changes to the network, establish new network standards, purchase new equipment, plan tests, and set up lab and pilot configurations.
- Lab build out and system testing - to ensure that the production deployment goes predictably, the solution is mocked up and tested in a controlled lab setting. Although the entire design is not physically built in the lab, a broad cross-section of components are configured and tested to ensure proper functionality and performance.
- Production deployment - this is the stage where all physical changes to the network are performed. All hardware is configured and located and all software is installed according to the design specifications.
- Production system testing - prior to exposing the final solution to the entire enterprise, a representative sample of the deployed solution is tested under very controlled but real work conditions. Real production content is used under real production conditions. The results are carefully monitored to validate that the solution is performing properly.

Production pilot - the final stage before making the solution available everyone in the enterprise is to conduct a pilot. This involves a small, predetermined group of regular users; use the solution in an everyday fashion. The results are measured and monitored carefully and any last-minute modifications are made before making the solution generally available.

summary

In this paper, we have reviewed three business scenarios in an attempt to demonstrate how Microsoft and HP products and technology can be applied to meet the demands for digital media in the enterprise. We have also reviewed the technology components required to deliver an enterprise solution, including the implementation of a CDN, media portal and options for content capture. The technology is mature and there are clear benefits to the business.

This document is part of a three-part Enterprise Digital Media Solution Guide series. Please also see the other papers in the series that discuss the business benefits and scenarios for using digital media in corporate communications and e-learning.

- Enterprise Digital Media Solution Guide PART I: E-Learning
- Enterprise Digital Media Solution Guide PART II: Corporate Communications

for more information

The best way to move forward will be different for each organization. Take the time to understand your business requirements and develop a plan to assess the infrastructure. For more information see the list below for Microsoft and HP resources to help you move forward.

HP Online Resources

ProLiant solutions for Streaming Media-
http://www.compaq.com/solutions/showroom/streamingmediasolutions.html
ProLiant servers
http://www.compaq.com/products/servers/platforms/index.html
Storage solutions
http://thenew.hp.com/country/us/eng/prodserv/storage.html
Client and hand-held solutions
http://thenew.hp.com/country/us/eng/prodserv/notebooks_handhelds.html

HP Services

HP Services has the expertise to help your business deploy enterprise solutions. For more information visit:

http://www.compaq.com/services/

Mobile and Media Systems Lab
http://www.hpl.hp.com/research/cp/cmsl/publications/streamingmedia_publications.htm

Microsoft Online Resources

The Windows Media website has a large selection of resources to help you get started, with everything from "How tos" and business justifications to case studies.

Windows Media
http://www.microsoft.com/windowsmedia
Windows Media enterprise
http://www.microsoft.com/windows/windowsmedia/enterprise.asp
Windows Media download center
http://www.microsoft.com/windows/windowsmedia/download/default.asp
Microsoft Producer
http://www.microsoft.com/windows/windowsmedia/technologies/producer.asp
Windows Media Rapid Economic Justification (REJ) white paper
http://www.microsoft.com/windows/windowsmedia/enterprise/value.asp#rej
Executive Broadcast white paper
http://www.microsoft.com/windows/windowsmedia/enterprise/deploy.asp
Windows Media deployment resources
http://www.microsoft.com/windows/windowsmedia/enterprise/deploy.asp

Microsoft Consulting Services

Microsoft Consulting Services (MCS) has a national presence and the expertise to help you implement an enterprise digital media solution. The MCS home page is:

http://www.microsoft.com/business/services/mcs.asp

Microsoft Partners

Microsoft has over 32,000 Certified Partners worldwide that can also help you get started with the expertise on Windows Media and enterprise solutions.

Microsoft partners

http://www.microsoft.com/business/partners/

Windows Media service providers
http://www.microsoft.com/windows/windowsmedia/service_provider/programs/wmsp.asp

About Approach Inc.

Approach was a contributing author to this paper in conjunction with Microsoft and HP. Approach is a consulting company advising on and delivering digital media solutions for the enterprise. Approach has significant expertise in designing and implementing Windows Media based solutions for the enterprise.

For more information, email digitalmedia@approach.com or visit www.approach.com

legal notice

appendix a - primer on digital media

What is Streaming?

Streaming media is a method of delivering digital media content across a network. Streaming occurs when a client begins to download content from a server and is able to play the content before the entire file has been downloaded. The client plays the content as the bits get transferred across the network in real time. Digital content can be streamed via *unicast* or *multicast*. A unicast stream serves a single client with a single stream from a central server, while a multicast can serve multiple clients with a single stream, similar to a radio or television broadcast.

Codecs

Streamed content is compressed by applying compression algorithms to the raw data. In order for the stream to be "played", it must be decompressed back to its original form using decompression algorithms. These compression and decompression algorithms are known as "codecs". Windows Media provides 'codecs' as part of Windows Media Encoder to encode content into the Windows Media format. Audio and video files are compressed because in their raw state they are very large, and so are not easily stored, manipulated and distributed. The encoded content is more manageable and can be more easily delivered over IP networks using minimal bandwidth.

Bit Rates

The bit rate determines the bandwidth requirements to stream a file. The bit rate is determined by the amount of compression applied by the codec during the encoding process. Windows Media Technologies is a leader in developing codec technology. Since streaming digital media is sensitive to network bandwidth conditions, Microsoft has developed codecs that enable a single encoding process to create a single Windows Media format stream containing multiple bit rates.

Media Players

The Windows Media Player is used to play encoded digital media content. Digital media is delivered by a Windows Media server or a cache appliance. A Windows Media server is specifically designed to communicate with a connected Windows Media Player. The media server is able to meter the delivery of data packets based on the feedback it receives from the player. This dynamic control of the stream allows a smooth, uninterrupted user experience.

Digital Media

Digital media is a type of media or a content format that has been digitized. Digital media is arranged into frames and can be streamed from a server to workstations because the content can be distributed via a serial stream of packets over a network.

Content Delivery Network

A content delivery network (CDN) is a series of network devices and appliances used to manage the distribution of digital media content from a central location. The ProLiant server is an example of a device that can be used to build a CDN. All of the distributed edge servers work in concert to move digital files through the network.

appendix b - digital media framework

The framework for an enterprise digital media solution has three main areas:

- Content creation
- Content distribution and management
- Content consumption

Considering the specific requirements of each area will enhance the quality, scalability, manageability, and reliability of the resulting solution.

The three solution areas are interdependent; a weakness in any one can undermine the effectiveness of the overall solution. For an end-to-end solution to be viable and add value, content, created cost-effectively on a regular basis is needed. This can then be distributed and viewed by the target audience.

Content Creation

Content creation can be seen as a set of guidelines, processes and tools which enable personnel to efficiently create and prepare digital media content. There are four primary infrastructure requirements for creating content:

- Encoding
- Storage
- Authoring tools
- Development tools

Companies will want to address video production needs if they are creating original content in a professional studio setting. This is not addressed in this paper, but is worth considering when implementing a digital media solution.

Encoding

Encoding is the process of digitizing analog audio and video input into a desired format for distribution via the enterprise network, CD-ROM or download. Encoders are hardware- and software-based services that take input from stored files in CDA, .mp3 .avi or .wav format. Encoding is a two-phase process, where analog content is first digitized via hardware-based encoder cards that produce "raw" .wav or .avi output. This output is then encoded by software-based encoders into the specified format with the desired bandwidth and quality settings.

Storage for Original Content

When producing original audio or video content, encoding existing VHS or Beta tapes or encoding final produced content, storage requirements are significant. It is best if all original content can be saved in its raw, highest-quality format so it can be re-purposed in the future if required.

Content production groups will generally want to maintain their own libraries of content, in its various phases of production. Often, multiple versions of a content element will exist to facilitate review and simultaneous work by several groups, in a similar way to when a large document is written. Scalable and reliable storage is essential.

Authoring / Development Tools

The right authoring tools and development tools are key to a good final product. The value of digital media lies in the integration and synchronization of content to make compelling presentations. There are two ways to achieve this, either use authoring tools or more sophisticated development tools.

- Authoring tools simplify the process by allowing drag and drop functionality in a graphical user interface. Two examples are Producer and Movie Maker, which are both intended to be end user applications. Distributing the content creation process out to end users will help

increase the rate at which content is published and also reduce the burden on central IT groups.

- Development tools, typically in the form of SDKs are intended for web developers to program custom solutions. There are situations where it will be desired to create more sophisticated content for special events, such as integration of Flash or DHTML into a presentation.
- The use of authoring applications is preferred if the creation process is going to be part of the business process on a regular basis. Leveraging development resources to create digital media presentations can be costly and generally more time consuming.

Content Distribution and Management

Once a piece of digital media content has been produced, it needs to be distributed to the intended audience in a timely and cost effective manner. Enterprises will want to manage this distribution process to ensure appropriate availability and security. The majority of the infrastructure work for distribution and management falls into three areas:

- Storage
- Content delivery network
- Media services

Storage

As content is produced and made available to end users, large amounts of network and server storage will be required. Typically, content will be centrally managed but highly distributed. This means that the source content is found in one central location, but copies of the content are found near the edge for user consumption. Storage solutions can be implemented in a variety of ways, but one large central repository with many smaller repositories on the edge is a common design principle.

When designing a storage solution, consider the access requirements, because this will often dictate the solution with the appropriate bandwidth for getting storage off the disks. Typical options include direct attached storage, network attached storage or a storage area network.

Content Delivery Network (CDN)

The CDN handles the distribution and management of the digital media content. This is the heart of any digital media solution.

The CDN will typically need to handle two forms of streaming content: live and on-demand. Delivering "live" digital media is similar to a live television broadcast in that the streamed content is delivered in real time, so that all participants view it at the same time while the event itself is taking place. On-demand delivery of digital media is similar to a VCR player, allowing viewers to select content and then to control its playback by rewinding, pausing and fast forwarding.

Digital media will impose demands on the network. Introducing it will most likely require a cache-based CDN solution. Caching technology enhances network capabilities without needing to increase bandwidth or upgrade network components. Adding content distribution capabilities gives the ability to deliver live and on-demand content throughout the enterprise without impacting the network's performance or bandwidth needs. Content can be proactively scheduled to be distributed to specific locations within the enterprise ensuring availability and improving viewing quality while minimizing WAN expense.

CDN Management

Adding web-based management tools to the CDN allows administrators to monitor stream quality, content bandwidth utilization and cache location performance. Content can be prioritized to ensure that business-critical content is delivered with the highest quality.

Media Services

Media services are a set of digital media components that determine the encoding standards, media servers, use of unicast or multicast distribution and media player functionality for playing media files.

Content Consumption

Content is accessed from a PC or PDA and can be viewed with a stand-alone media player or with an embedded player found inside an intranet portal or web page. This part of the solution framework has three infrastructure requirements:

- Device support (PC, PDA etc)
- Media player support
- Media portal

A user may be directed to a media portal to locate materials for viewing. Once the content is located, the user expects a quick and easy way of viewing it.