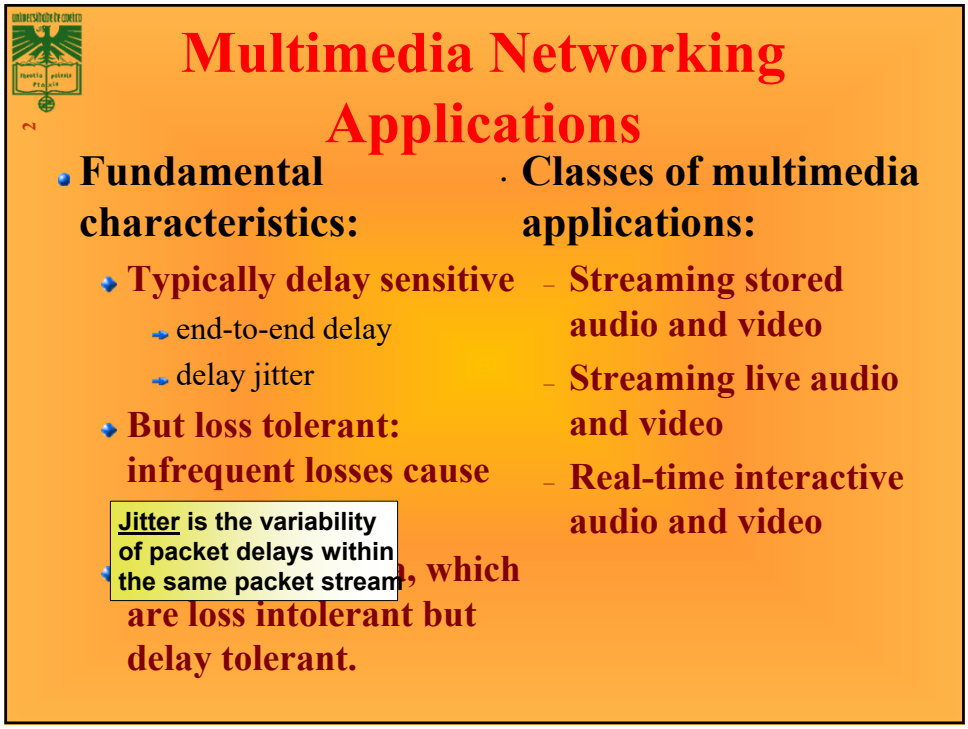
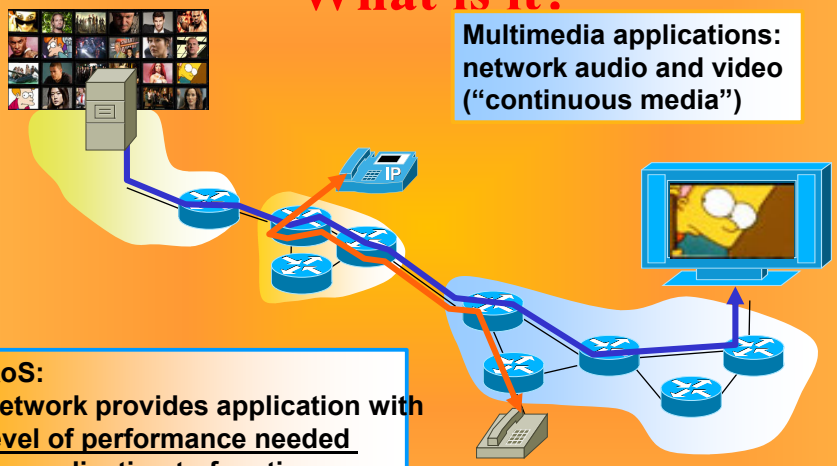

Bringing it together
Multimedia in IP
(Web view)

1


Multimedia Networking Applications
Fundamental characteristics:
Typically delay sensitive
end-to-end delay
delay jitter
But loss tolerant: infrequent losses cause
Jitter is the variability of packet delays within the same packet stream
, which are loss intolerant but delay tolerant.
Classes of multimedia applications:
Streaming stored audio and video
Streaming live audio and video
Real-time interactive audio and video
2

2

Multimedia, Quality of Service: What is it?



Multimedia applications: network audio and video ("continuous media")

QoS:
Network provides application with level of performance needed for application to function.

3

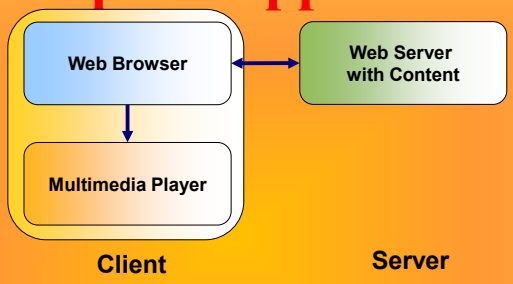
Internet Multimedia Support

- Integrated services philosophy.
  - Requires dedicated links/channels with QoS requirements.
- Differentiated services philosophy.
  - Fewer changes to Internet infrastructure.
- Best effort.
  - No major changes.
  - More bandwidth when needed.
  - Application-level control and distribution.

Would require QoS
Only possible in private networks or operator infrastructure

4

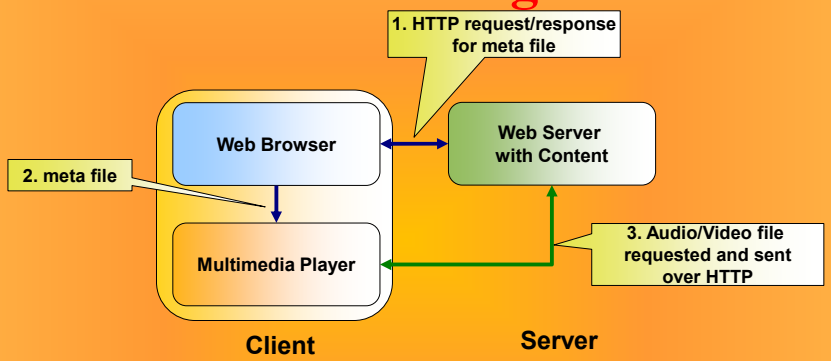
Internet Multimedia: Simplest Approach



- Audio or video stored in file.
- Files transferred as HTTP object (or using P2P).
  - Received in entirety at client as a file.
  - Then passed to default player in client.
- Audio&video is not streamed!
- No "pipelining", long delays until playout!

5

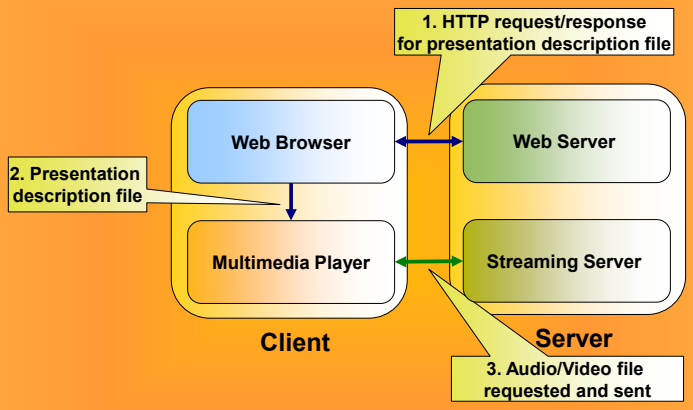
Internet Multimedia: Web Streaming



- Browser GETs metafile.
  - Content negotiation may happen.
- Browser launches player, passing metafile
- Player contacts server.
- Server streams audio/video to player.

6

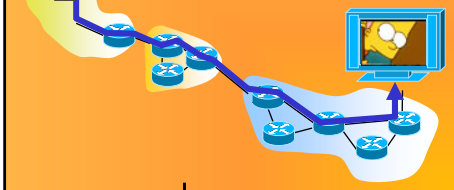
Streaming from a streaming server



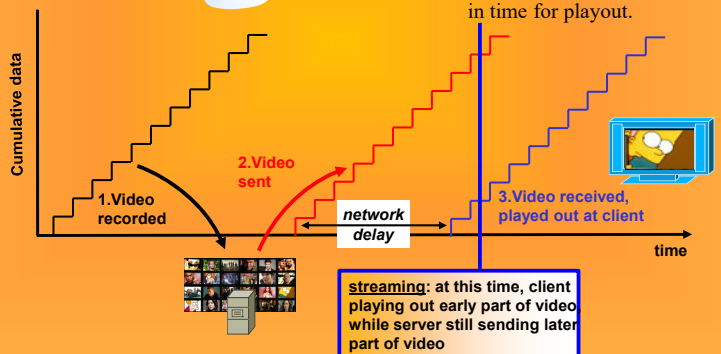
- This architecture allows for non-HTTP protocol between server and media player.
- Can use UDP or TCP transport.

7

Web Streaming: Stored Multimedia

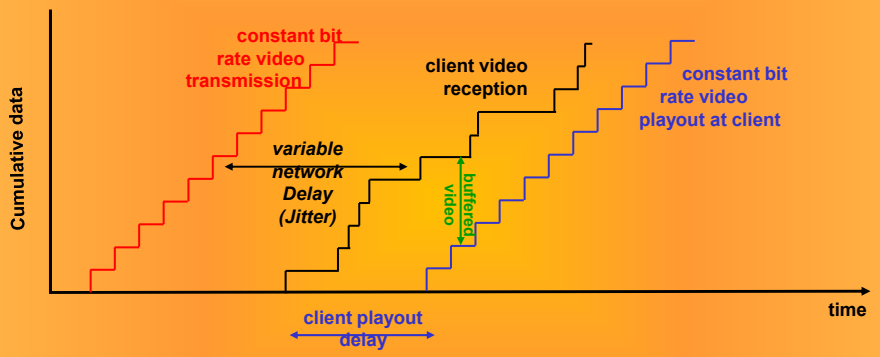


- Streaming:
  - Media stored at source is transmitted to client.
  - Client visualization begins before all data has arrived.
    - Timing constraint for transmitted data: in time for playout.



8

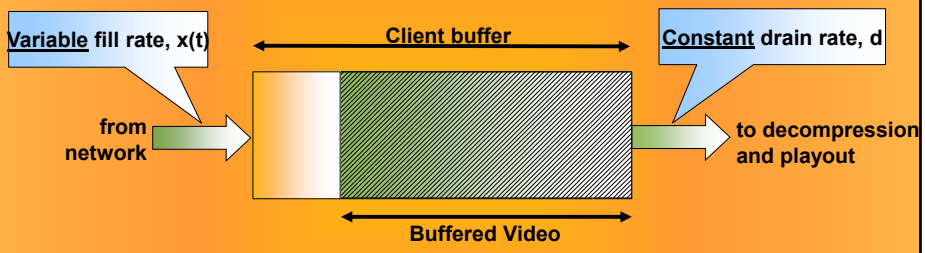
Streaming Multimedia: Client Buffering



- Client-side buffering, playout delay compensate for network-added delay, delay jitter.
  - Size of buffer/playout delay are parameters that can be adjusted dynamically
  - For VoIP, delay between 2 packets is about 20ms (plus *jitter*).

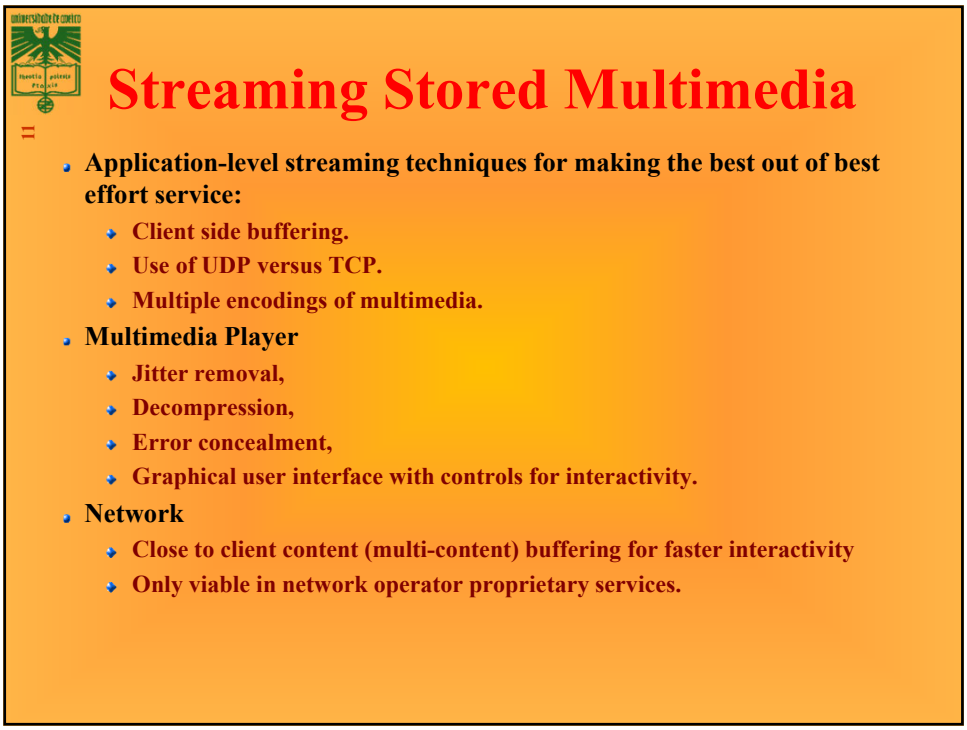
9

Streaming Multimedia: Client Buffering

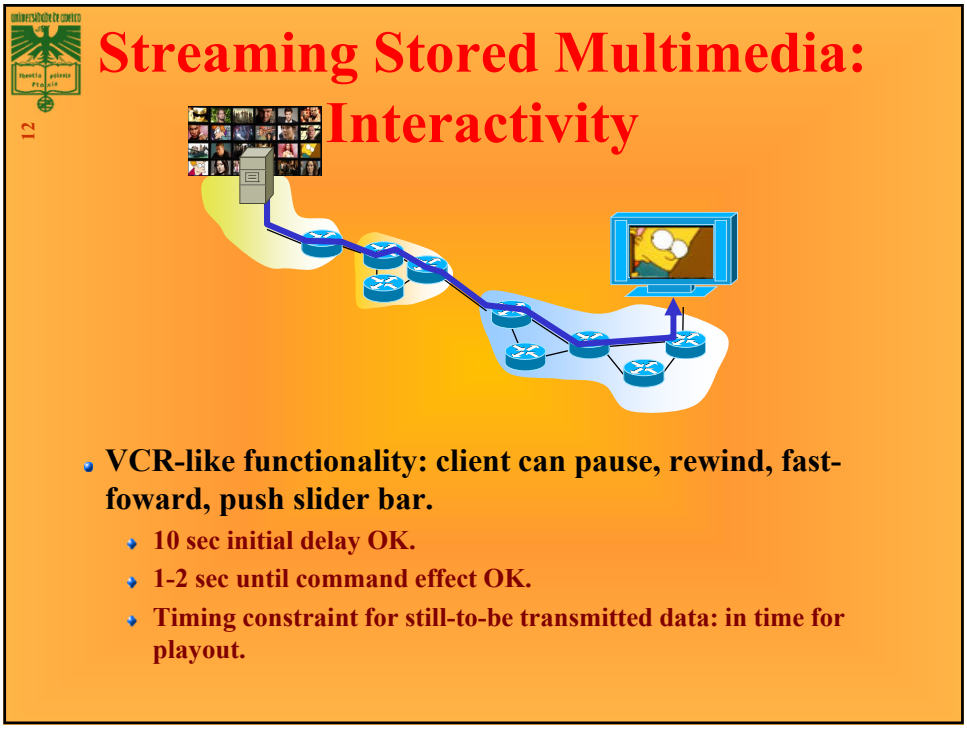


- Client-side buffering, playout delay compensate for network-added delay, delay jitter.
  - Size of buffer/playout delay are parameters that can be adjusted dynamically

10


11
Streaming Stored Multimedia
Application-level streaming techniques for making the best out of best effort service:
Client side buffering.
Use of UDP versus TCP.
Multiple encodings of multimedia.
Multimedia Player
Jitter removal,
Decompression,
Error concealment,
Graphical user interface with controls for interactivity.
Network
Close to client content (multi-content) buffering for faster interactivity
Only viable in network operator proprietary services.

11


12
Streaming Stored Multimedia: Interactivity
VCR-like functionality: client can pause, rewind, fast-foward, push slider bar.
10 sec initial delay OK.
1-2 sec until command effect OK.
Timing constraint for still-to-be transmitted data: in time for playout.

12

13

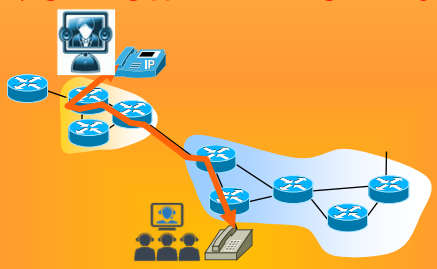
Streaming Live Multimedia

- Examples:
  - Internet TV/radio show.
  - Live sporting event.
- Streaming
  - Playback buffer.
  - Playback can lag tens of seconds after transmission.
  - Still have timing constraint.
- Interactivity
  - Fast forward impossible.
  - Rewind, pause possible!

13

14

Interactive Real-Time Multimedia



- Applications:
  - IP telephony, video conference, online-game multimedia actions, distributed interactive worlds.
- End-end delay requirements:
  - Audio: < 150 msec good, < 400 msec OK
    - Includes application-level (packetization) and network delays.
    - Higher delays noticeable, impair interactivity.
- Requires session initialization
  - Advertise its IP address, port number, encoding algorithms, required contents, available contents

14

15

UDP Streaming vs. TCP Streaming

- **UDP**
  - **Server sends at rate appropriate for client .**
    - Often send rate = encoding rate = constant rate.
    - Then, fill rate = constant rate - packet loss.
  - **Short playout delay (2-5 seconds) to compensate for network delay jitter.**
  - **Error recover: time permitting.**
- **TCP**
  - **Send at maximum possible rate under TCP.**
  - **Fill rate fluctuates due to TCP congestion control.**
  - **Larger playout delay: smooth TCP delivery rate.**
  - **HTTP/TCP passes more easily through firewalls.**

15

16

HTTP/TCP Streaming

- **Multiple versions with distinct/complementary characteristics are generated for the same content**
  - **With different bitrates, resolutions, frame rates.**
- **Each version is divided into time segments.**
  - **e.g., two seconds.**
- **Each segment is provided on a web server and can be retrieved through standard HTTP GET requests.**
- **Examples of protocols:**
  - **MPEG's Dynamic Adaptive Streaming over HTTP (DASH).**
    - Standard ISO/IEC 23009-1. YouTube's default.
  - **Adobe HTTP Dynamic Streaming (HDS).**
  - **Apple HTTP Live Streaming (HLS).**
  - **Microsoft Smooth Streaming (MSS).**

16

17

User Control of Streaming Media: RTSP

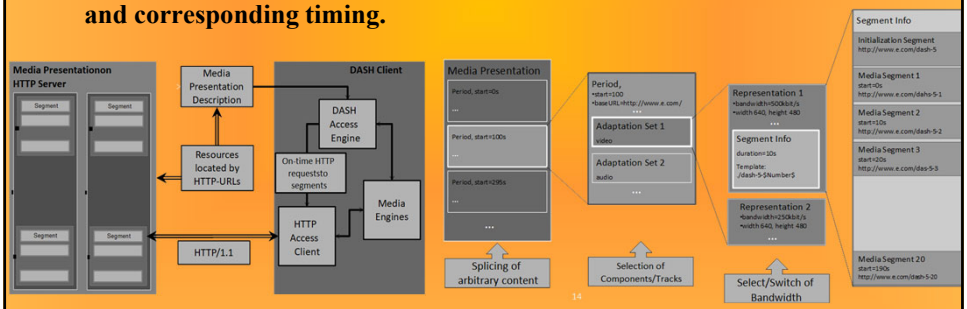
- RTSP (Real Time Streaming Protocol): RFC 2326
  - Client-server application layer protocol.
  - For user to control display: rewind, fast forward, pause, resume, repositioning, etc…
- Does not define how audio/video is encapsulated for streaming over network.
- Does not restrict how streamed media is transported.
  - Can be transported over UDP or TCP.
- Does not specify how the media player buffers audio/video.
- RTSP messages are also sent out-of-band:
  - RTSP control messages use different port numbers than the media stream: out-of-band
    - Port 554
  - The media stream is considered "in-band"

17

22

Dynamic Adaptive Streaming over HTTP (DASH)

- Developed to be an Open Standard Delivery Format.
  - MPEG DASH ISO/IEC 23009-1.
- Video streaming solution where pieces of video streams/files are requested with HTTP and spliced together by the client.
  - Client entirely controls delivery.
- Media Presentation Description (MPD) describes accessible Segments and corresponding timing.



22

23

WebRTC

- Peer-to-peer connections.
  - An instance allows an application to establish peer-to-peer communications with another instance in another browser, or to another endpoint implementing the required protocols.
- RTP Media.
  - Allow a web application to send and receive media stream over a peer-to-peer connection (discussed in a minute)
- Peer-to-peer Data
  - Allows a web application to send and receive generic application data over a peer-to-peer connection.
- Peer-to-peer DTMF.

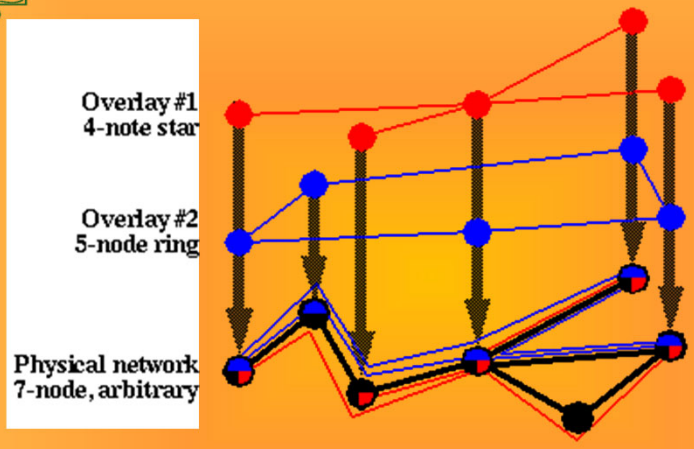
23

CDNs

Everyone in the same network ?

24

Recall: what is an Overlay ?



What is the topology of this network?

WHICH network??

www.isi.edu/xbone

26

Overlay Networks: Overview

- **Networks built using an existing network as substrate (*Virtual Networks)***

Internet

  - **Initially an overlay on the POTS (Plain Old Telephone System) network**
- **Overlays are a (quasi) structured virtual topology above the basic transport protocol level that facilitates deterministic search and guarantees convergence**
  - **Overlays could consist of routing software installed at selected sites, connected by encapsulation tunnels or direct links**
- **Examples of overlays:**
  - **MBone, 6Bone**
  - **P2P (Napster, FreeNet, Gnutella, Bittorrent)**
  - **Cooperating Caches**
  - **Server Farms**
  - **Content Distribution Networks (CDNs)**

27

Content Distribution Networks

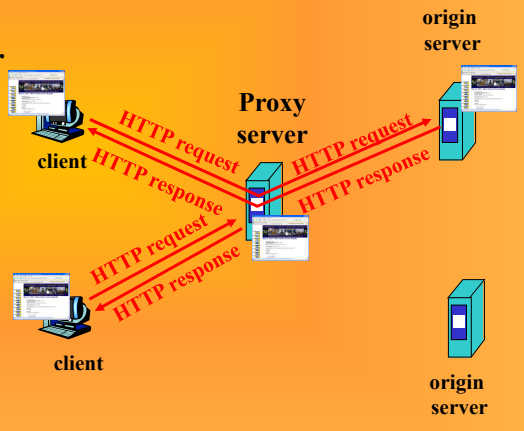
Client-Server and distribution models
Caching and load balancing

28

(recall FR): Web caches (proxy server)

Goal: satisfy client request without involving origin server

- user sets browser: Web accesses via proxy server
- browser sends all HTTP requests to proxy
  - object in cache: cache returns object
  - else proxy requests object from origin server, then returns object to client



30

More about Web caching

- Proxy server acts as both client and server
- typically proxy server is installed by ISP (university, company, residential ISP)

Why Web caching?

- reduce response time for client request
- reduce traffic on an institution's access link.

31

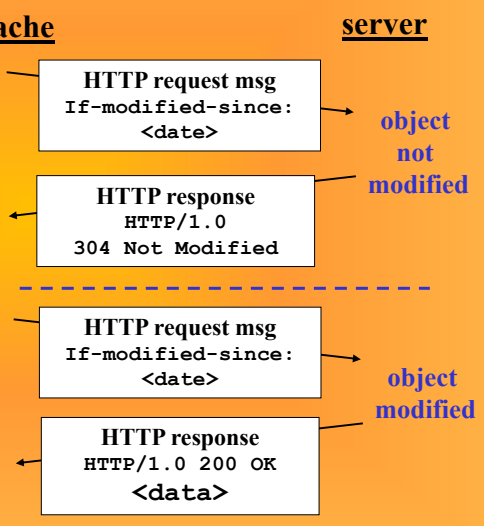
Conditional GET

- **Goal:** don't send object if cache has up-to-date cached version
- cache: specify date of cached copy in HTTP request

```
If-modified-since:
  <date>
```

- server: response contains no object if cached copy is up-to-date:

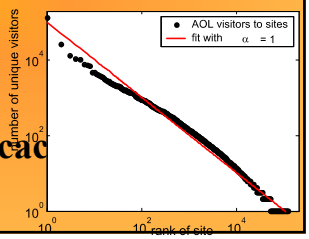
```
HTTP/1.0 304 Not
  Modified
```



36

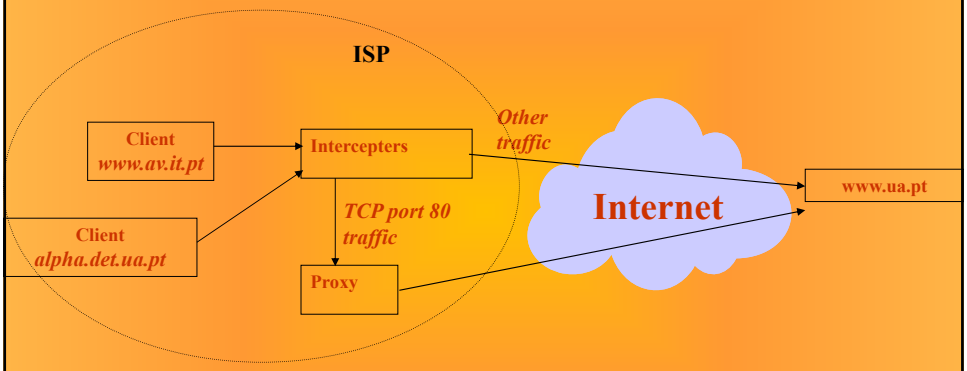
Optimizing performance

- **Where to cache content?**
  - **Popularity of Web objects is Zipf-like**
    - a few elements that score *very* high (the left tail in the diagrams)
    - a medium number of elements with middle-of-the-road scores (the middle part of the diagram)
    - a huge number of elements that score very low (the right tail in the diagram)
  - **Small number of sites cover large fraction of requests**
- **Given this observation, how should cac replacement work?**



37

Potential content network structure: Caching Proxies

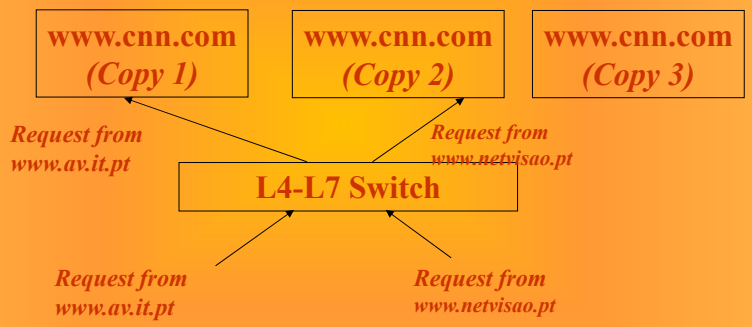


- **Mostly motivated by ISP business interests – reduction in bandwidth consumption of ISP from the Internet**
  - **Reduced network traffic**
  - **Reduced user perceived latency**

41

Potential content network structure : Server Farms

Simple solution to the content distribution problem: deploy a large group of servers on site

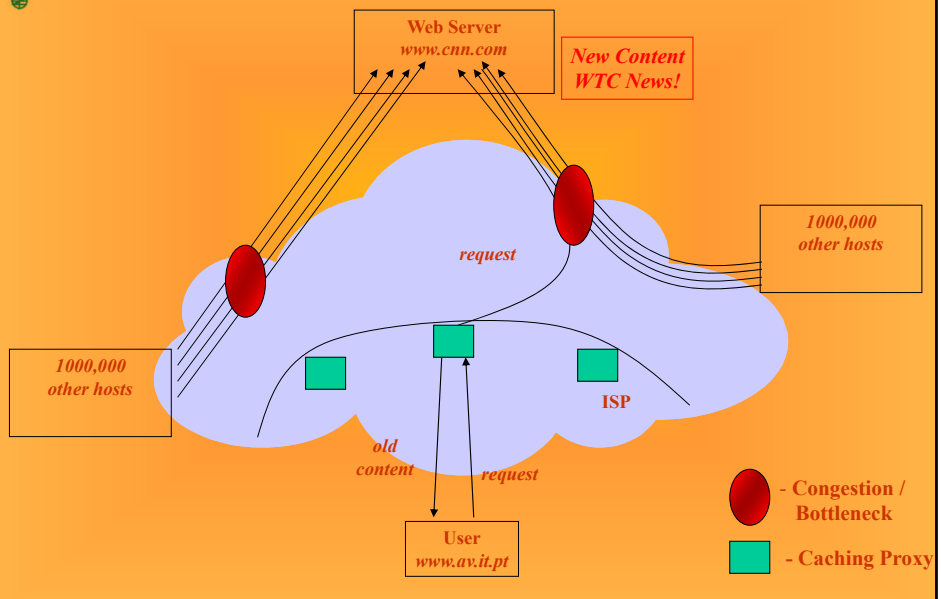


- **Arbitrate client requests to servers using an "intelligent" L4-L7 switch**
  - **Quite used today**

42

Flash Crowds:

Consider, On September 11, 2001



43

44

Why Not Web-only approaches for content networks?

- Integrating file caching in proxies
  - Optimized for 10KB objects
  - 10GB = 1.000.000 x 10KB
- Memory pressure
  - Disk access is 1000 times slower
  - Working sets do not fit in memory
- Waste of resources
  - More servers needed
  - Provisioning is a must

44

Problems with *Server farms and Caching proxies*

- Server farms do nothing about problems due to network congestion, or to improve latency issues due to the network
- Caching proxies serve only their clients, not all users on the Internet
- Content providers *(say, Web servers)* cannot rely on existence and *correct* implementation of caching proxies
- Accounting issues with caching proxies.
  For instance, *www.cnn.com* needs to know the number of hits to the webpage for advertisements displayed on the webpage

45

CDNs

47

The Internet in 2023, every minute

49

50

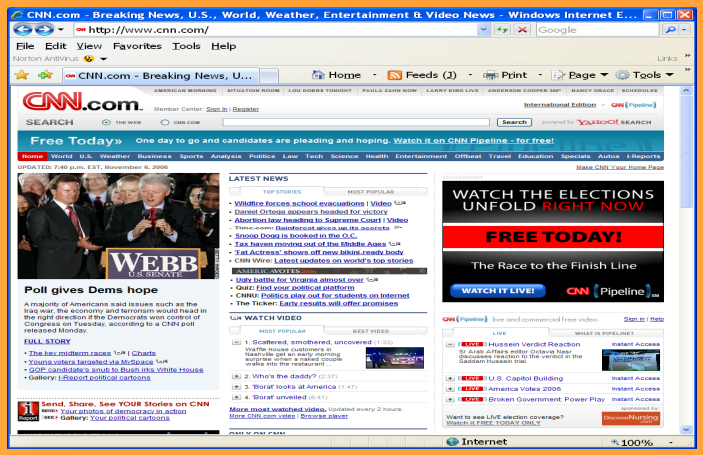
Lots of multimedia

	TOTAL MOBILE VOLUME		DOWNSTREAM VOLUME		UPSTREAM VOLUME	
	Category	Total Volume	Category	Total Volume	Category	Total Volume
1	Video	67.60%	Video	70.35%	Video	37.11%
2	Social Networking	12.16%	Social Networking	12.27%	Messaging	18.23%
3	Messaging	5.89%	Messaging	4.78%	Web Browsing	11.95%
4	Web Browsing	4.51%	Web Browsing	3.83%	Social Networking	10.96%
5	Marketplace	2.77%	Marketplace	2.86%	Cloud	9.81%
6	Gaming	2.41%	Gaming	2.43%	File Sharing	4.27%
7	File Sharing	1.97%	File Sharing	1.77%	VPN	3.65%
8	Cloud	1.79%	Cloud	1.06%	Gaming	2.11%
9	VPN	0.79%	VPN	0.53%	Marketplace	1.82%
10	Audio	0.11%	Audio	0.12%	Audio	0.10%

50

51

CDNs Target environment?



Most Web files are small (1KB ~ 100KB)

(initially....)

51

52

Motivation

- IP based networks
- Web based applications have become the norm for corporate internal networks and many business-to-business interactions
- Large acceptance and explosive growth
  - Serious performance problems
  - Degraded user experience

  For a large set of applications, including VIDEO access
- Improving the performance of networked applications
  - Use many sites at different points within the network
    - Stand alone servers
    - Routers

52

CDNs basics

- What is a CDN?
  - A network of servers delivering content on behalf of an origin site
    - A number of CDN companies well established now
      - E.g. Akamai, Digital Island, Speedera, CDN77, Cloudfare, Stackpat
    - Many companies are exploring CDNs
      - Avoid congested portions of the Internet
- Consist of
  - Edge servers deployed at several ISP (Internet Service Provider) access locations and network exchange points
- Large-file service with no custom client, no custom server, no prepositioning
- Improve the response time of an Internet site
  - Offloading the delivery of bandwidth-intensive objects, such as images and video clips
- Intelligent Internet infrastructure that improves the performance and scalability of distributed applications by moving the bulk of their *computation* to servers located at the edge of the network
  - Applications are logically split into two components
    - Executed at an edge server close to the user
    - Executed on a traditional application server

53

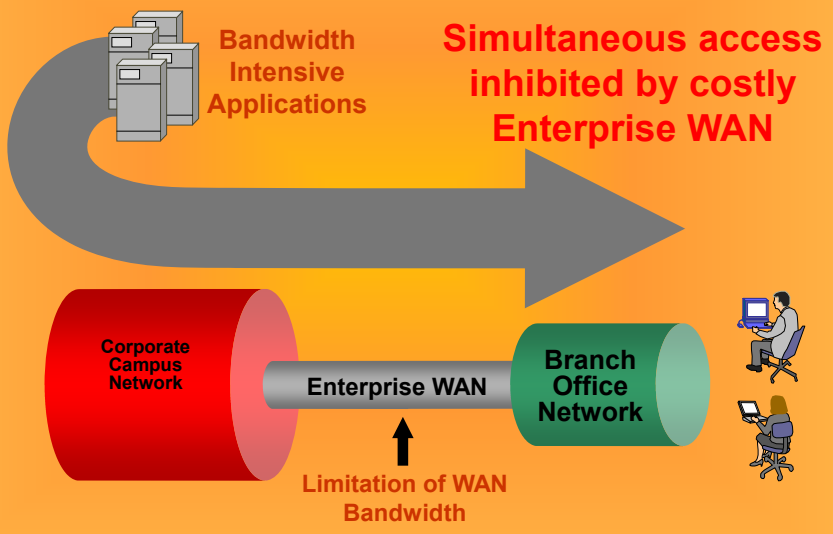
54

CDN Generations

- **First generation (early 90ies)**
  - **Accelerate the performance of web sites**
  - **Support increasing volumes of traffic**
    - Key disruption event: 9/11
    - Akamai technologies created
- **Second generation (early 2000ies)**
  - **Support high volumes of multimedia traffic**
  - **Audio/video intensive networks**
    - All ISPs developed/used CDNs
- **Third generation (2010+)**
  - **Cloud computing**
    - Amazon cloud (2008)
  - **UGC (user generated content)**
  - **P2P and interactivity**
    - AT&T distributed data centers (2011)
  - **Mobile support, and device adapted content**

54

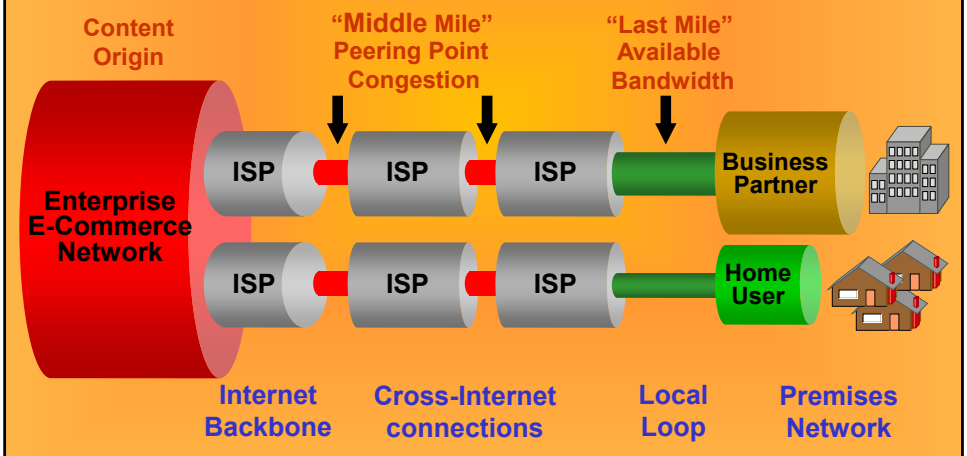
Internal Enterprise Applications



55

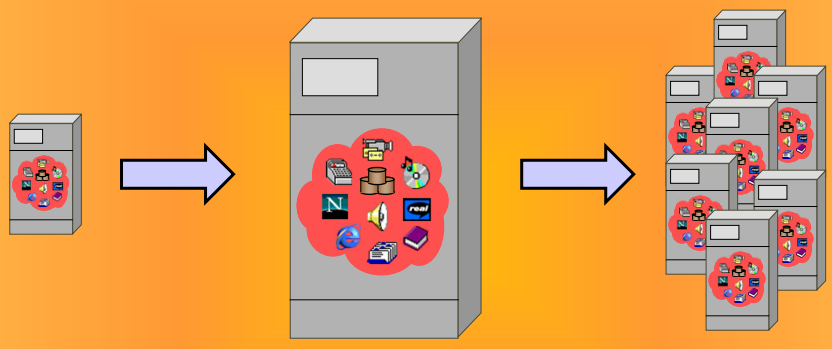
External Enterprise Applications

E-Commerce applications limited by congestion points within the Internet



56

Content Scaling



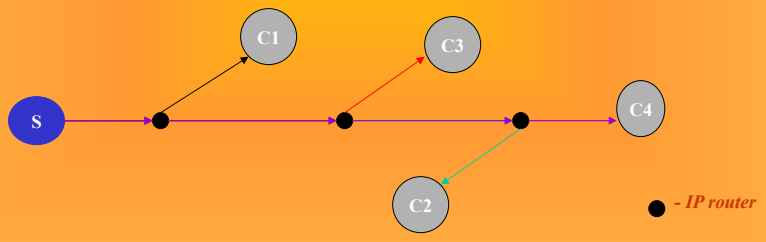
- Need to scale content to handle numerous clients
  - One can only scale 'vertically' to a point

Multiple servers/locations introduces new issues

57

Early Motivations for Content Networks (1st generation)

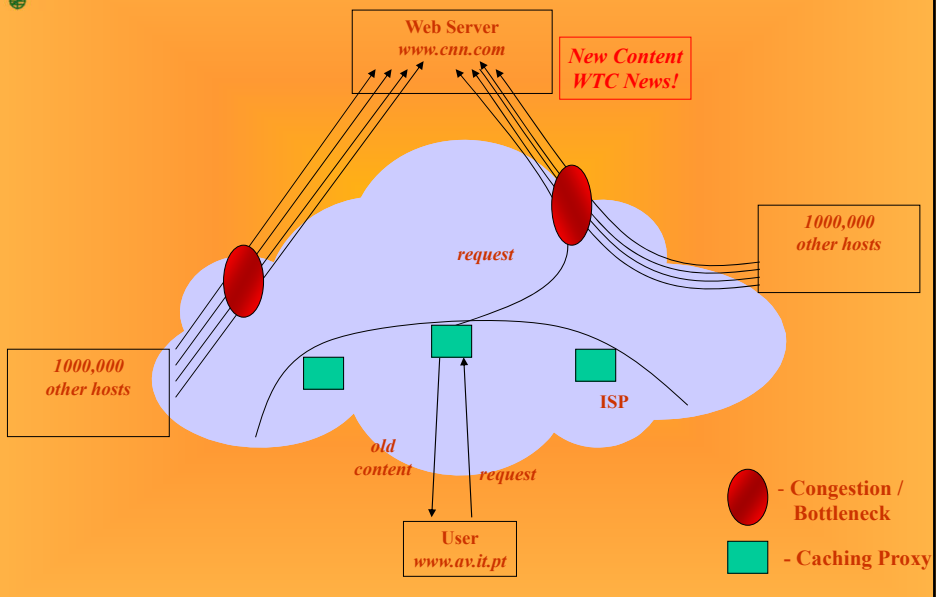
- More hops between client and Web server => more congestion!
- Same data flowing repeatedly over links between clients and Web server
- Origin server is bottleneck as number of users grows
- Flash Crowds *(for instance, Sept. 11)*
  - *The Content Distribution Problem:* Arrange a rendezvous between a content source at the origin server *(www…com)* and a content sink *(users)*



58

Flash Crowds:

Consider, On September 11, 2001



60

61

Flash crowd solution: CDNs..

What is a CDN?

A network of servers delivering content on behalf of an origin site

Large-file service with

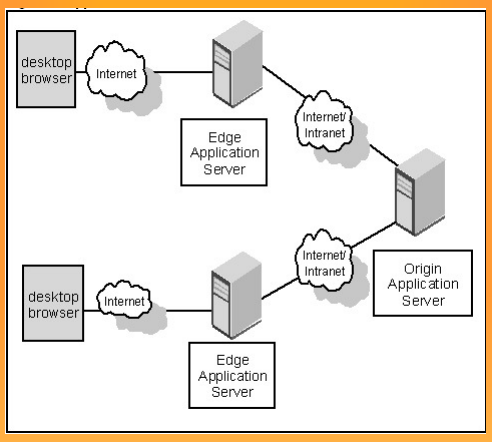
- **No custom client**
- **No custom server**
- **No prepositioning**
- **No rehosting**
- **No manual provisoning**

61

62

Model

- **Application offload (1st generation concern)**



62

63

Content distribution networks

- Client attempts to access the main server site for an application
- It is redirected to one of the other sites
- Each site caches information
  - Avoid going to the main server to get the information/application
- Access a closelly located site
  - Avoid congestion on the path to the main server
- Set of sites used to improve the performance of web-based applications collectivelly
  - Content distribution network

63

Inside a CDN

- Servers are deployed in clusters for reliability
  - Some may be offline
    - Could be due to failure
    - Also could be "suspended" (e.g., to save power or for upgrade)
- Could be multiple clusters per location (e.g., in multiple racks)
- Server locations
  - Well-connected points of presence (PoPs)
  - Inside of ISPs

64

65

Advantages

- Better scalability
- Higher availability
- Improved response time from a centrally managed solution
- Nodes constituting the distribution network are designed to be
  - Self-configuring
  - Self-managing
  - Self-diagnosing
  - Self-healing

to ensure easy management and operational convenience

65

66

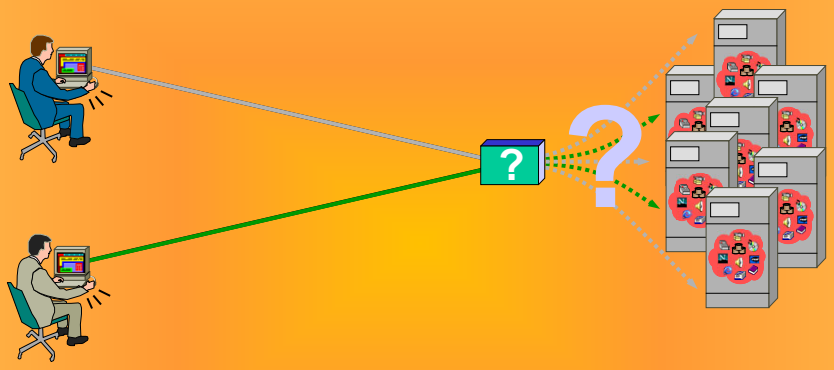
Challenges

- Keep consistency among the enterprise data hosted by the offloaded applications
- Share session state among edge and origin application servers
- Distribution, configuration, and management
- Develop programming models consistent with current industry standards such as J2EE
- Application security.

- There is active research into general frameworks to be used to support distributed applications, as well as prototyping the ideas for specific application instances

66

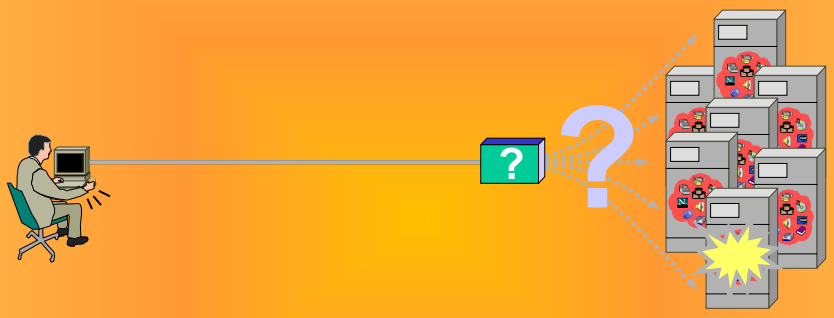
Load-Sharing Content



- Handle requests fairly amongst servers/sites
- Easily add servers/sites to content service
- Adjust connections based on server/site load

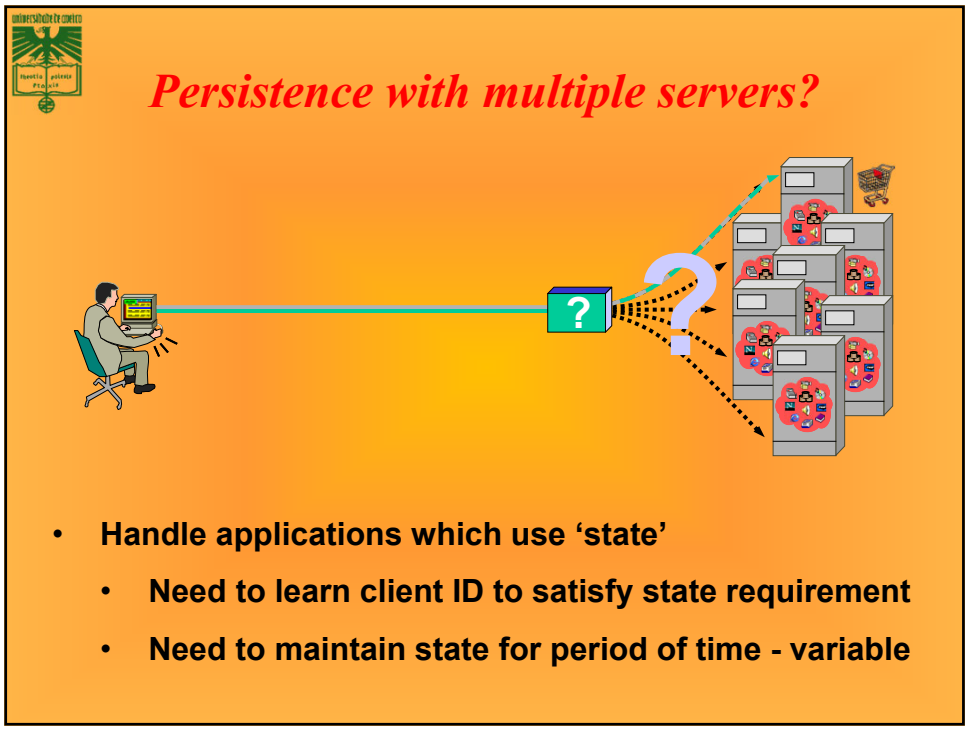
67

Content Availability with multiple servers?



- Synchronize content amongst servers/sites
- Avoid faulty servers/sites
- Faulty servers/sites includes invalid/dated content

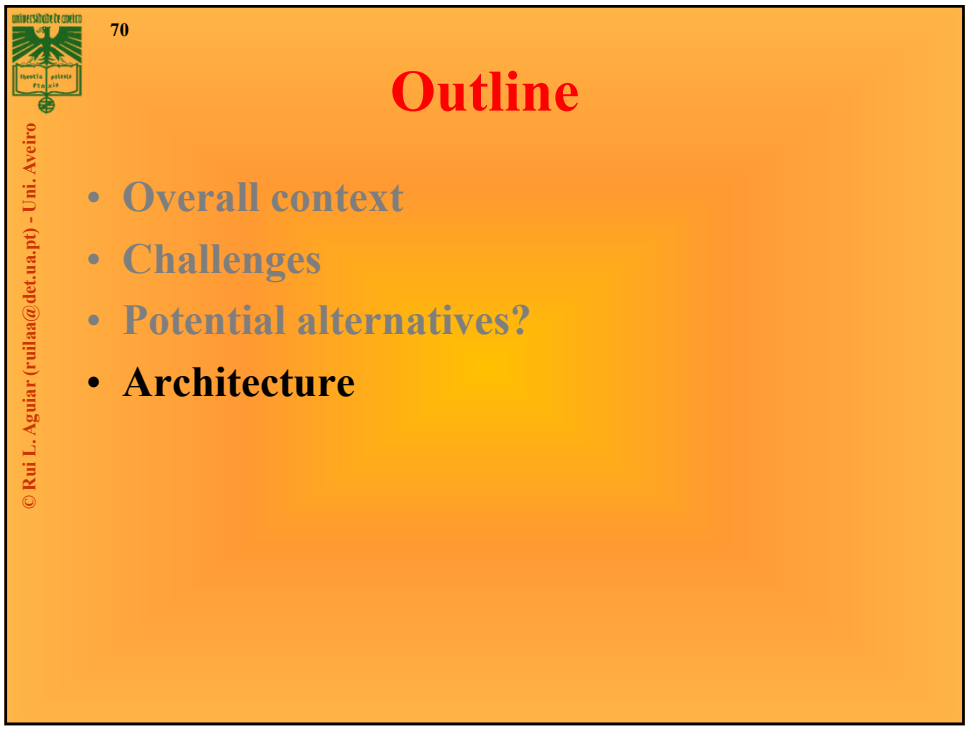
68



Persistence with multiple servers?

- Handle applications which use 'state'
  - Need to learn client ID to satisfy state requirement
  - Need to maintain state for period of time - variable

69



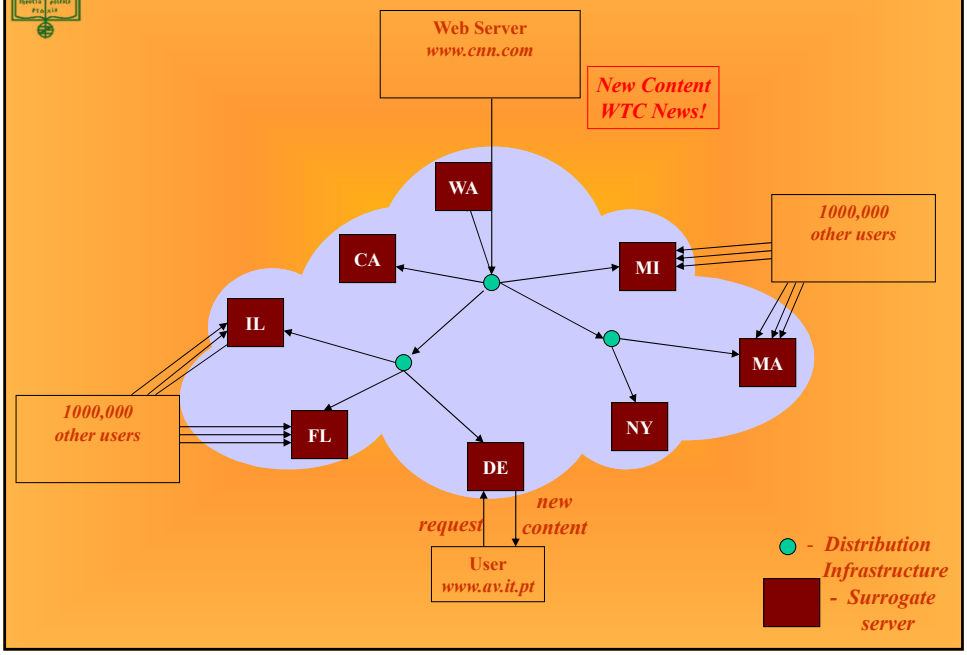
70

Outline

- Overall context
- Challenges
- Potential alternatives?
- **Architecture**

70

CDN, On September 11, 2001



71

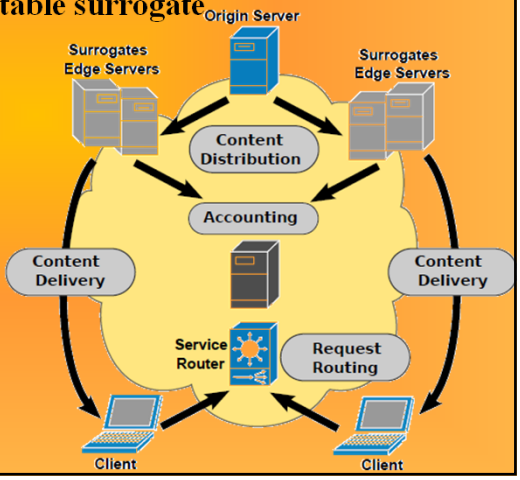
With CDNs

- Overlay network to distribute content from origin servers to users
  - Avoids large amounts of same data repeatedly traversing potentially congested links on the Internet
  - Reduces Web server load
  - Reduces user perceived latency
  - Tries to route around congested networks
- CDN is not a cache!
  - Caches are used by ISPs to reduce bandwidth consumption, CDNs are used by content providers to improve quality of service to end users
  - Caches are reactive, CDNs are proactive
  - Caching proxies cater to their users (web clients) and not to content providers (web servers), CDNs cater to the content providers (web servers) and clients
  - CDNs give control over the content to the content providers, caching proxies do not

72

CDN Components

- *Content Delivery Infrastructure:* Delivering content from producer to clients by surrogates
- *Request Routing Infrastructure:* Steering or directing content request from a client to a suitable surrogate
- *Distribution Infrastructure:* Moving or replicating content from content source (origin server, content provider) to surrogates
- *Accounting Infrastructure:* Logging and reporting of distribution and delivery activities



73

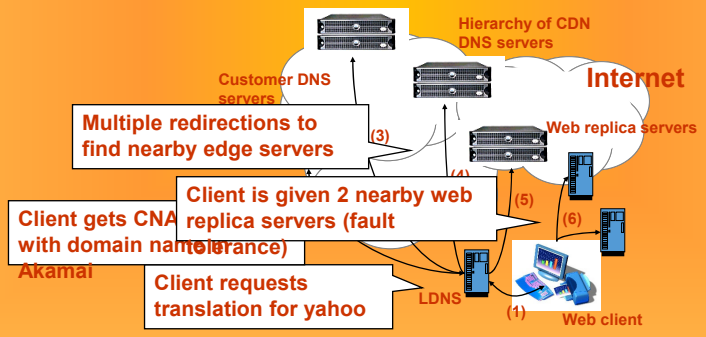
Mapping clients to servers

- CDNs need a way to send clients to the "best" server
  - The best server can change over time
  - And this depends on client location, network conditions, server load, …
  - What existing technology can we use for this?
- DNS-based redirection
  - Clients request www.foo.com
  - DNS server directs client to one or more IPs based on request IP
  - Use short TTL to limit the effect of caching

74

DNS Redirection

- Web client's request redirected to 'close' by server
  - Client gets web site's DNS CNAME entry with domain name in CDN network
  - Hierarchy of CDN's DNS servers direct client to 2 nearby servers



75

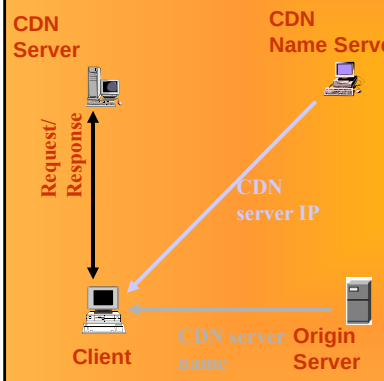
DNS Redirection Considerations

- Advantages
  - Uses existing, scalable DNS infrastructure
  - URLs can stay essentially the same
- Limitations
  - DNS servers see only the DNS server IP
    - Assumes that client and DNS server are close. Is this accurate?
  - Content owner must give up control
  - Unicast addresses can limit reliability

76

77

What other CDN techniques are being used?



- DNS redirection (DR)
  - Full-site delivery
  - Partial-site delivery
- URL rewriting
- Hybrid scheme
  - URL rewriting + DNS redirection
- Manual hyperlink selection
- HTTP redirection
- Layer 4 switching
- Layer 7 switching
- Anycast

77

78

Offloading a portal

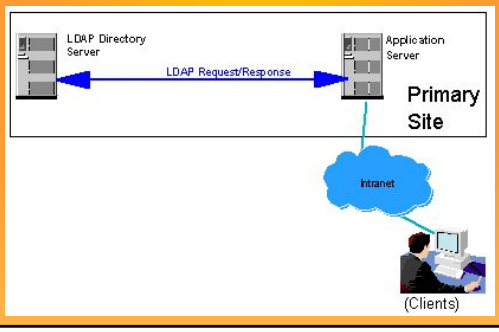
- Portal servers allow users to access content and applications from a single access point
  - Users can create persistent, customized views of applications and content chosen from the set of applications and content by the portal administrators
- Portal server pages are personalized
- Often include dynamic content
- Significant amount of computation required for page assembly
  - Application offload

78

79

Offloading an Enterprise directory

- E.g. a common e-Workplace tool
- The employee data is often stored in a central LDAP directory
  - Separate web-based application providing the interface to the directory

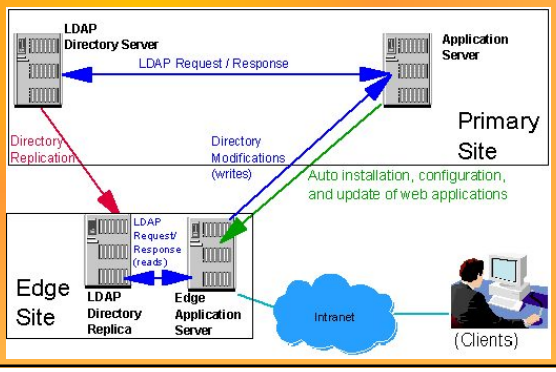


79

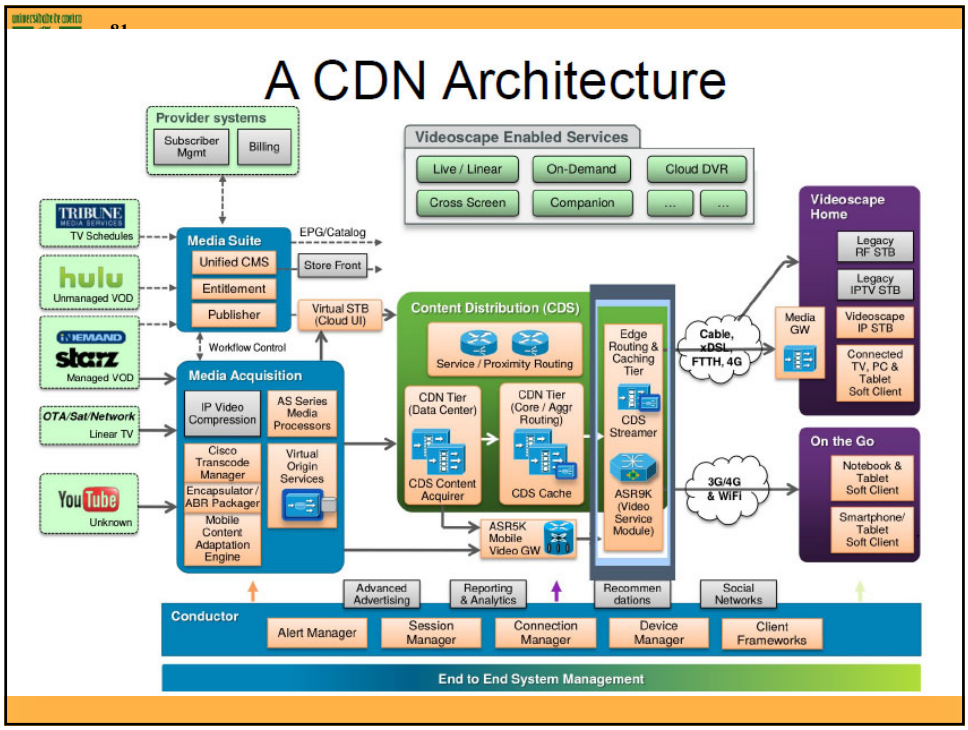
80

Offloading an Enterprise directory

- Centralized directory
  - Convenient to manage
  - Performance for clients accessing the directory from remote sites can be poor
    - E.g. transcontinental network connections suffer from a long delay
- Offloaded version of the application



80


A CDN Architecture
Provider systems
Subscriber Mgmt
Billing
Videoscape Enabled Services
Live / Linear
On-Demand
Cloud DVR
Cross Screen
Companion
...
...
TRIBUNE MEDIA SERVICES
TV Schedules
hulu
Unmanaged VOD
DEMAND
starz
Managed VOD
OTA/Sat/Network
Linear TV
YouTube
Unknown
Media Suite
Unified CMS
Entitlement
Publisher
EPG/Catalog
Store Front
Virtual STB (Cloud UI)
Workflow Control
Media Acquisition
IP Video Compression
AS Series Media Processors
Cisco Transcode Manager
Encapsulator / ABR Packager
Mobile Content Adaptation Engine
Virtual Origin Services
Content Distribution (CDS)
Service / Proximity Routing
CDN Tier (Data Center)
CDS Content Acquirer
CDN Tier (Core / Aggr Routing)
CDS Cache
ASR5K Mobile Video GW
Edge Routing & Caching Tier
CDS Streamer
ASR9K (Video Service Module)
Cable, xDSL, FTTH, 4G
3G/4G & WiFi
Media GW
Videoscape Home
Legacy RF STB
Legacy IPTV STB
Videoscape IP STB
Connected TV, PC & Tablet Soft Client
On the Go
Notebook & Tablet Soft Client
Smartphone/ Tablet Soft Client
Advanced Advertising
Reporting & Analytics
Recommendations
Social Networks
Conductor
Alert Manager
Session Manager
Connection Manager
Device Manager
Client Frameworks
End to End System Management

81