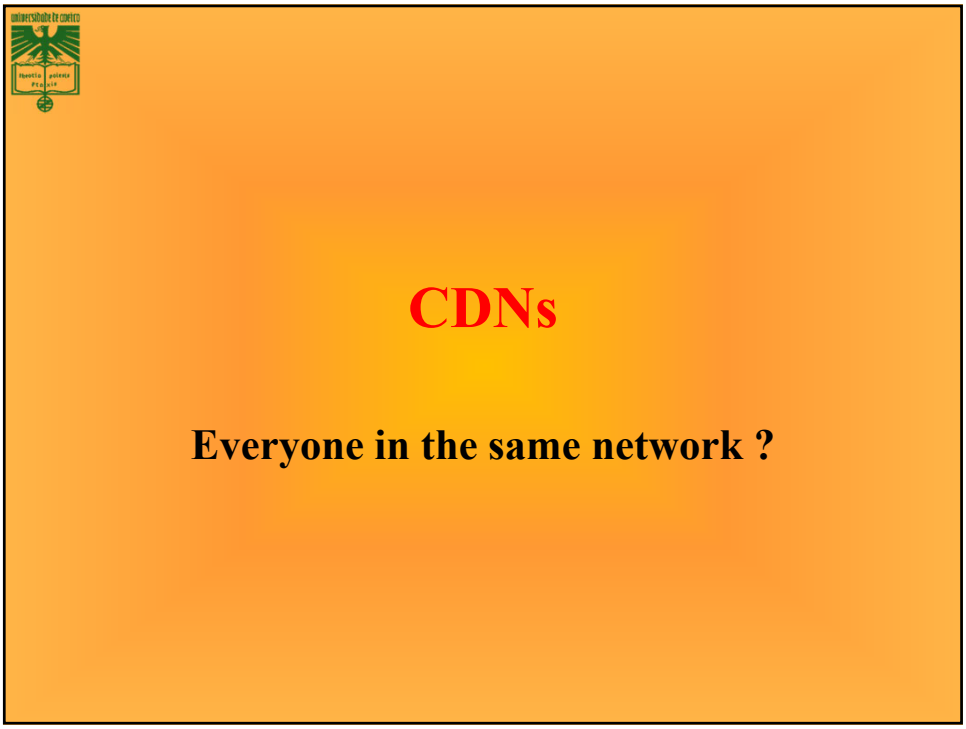
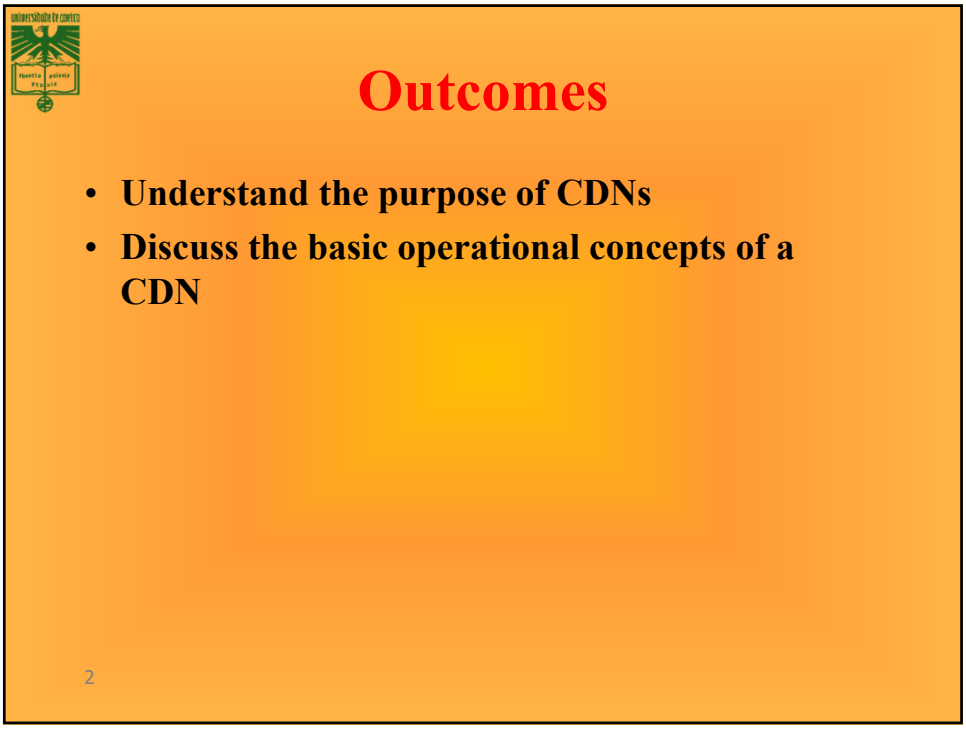
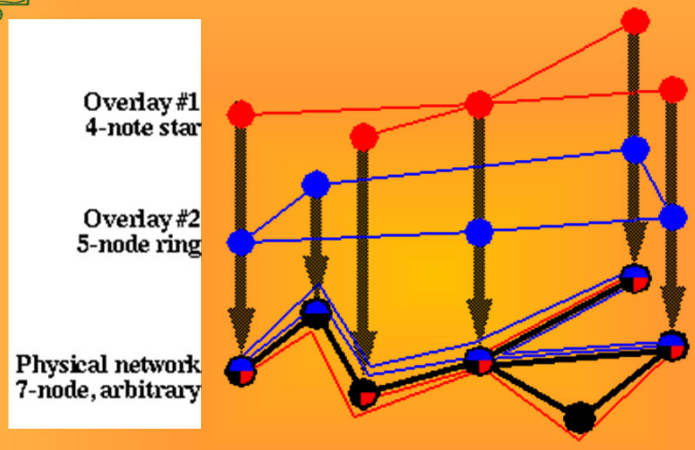

CDNs
Everyone in the same network ?

1


Outcomes
• Understand the purpose of CDNs
• Discuss the basic operational concepts of a CDN
2

2

Recall: what is an Overlay ?



What is the topology of this network?

WHICH network??

www.isi.edu/xbone

3

Content Distribution Networks

Client-Server and distribution models
Caching and load balancing

5

© Rui L. Aguiar (ruilaa@det.ua.pt) - Uni. Aveiro

The Internet in 2023, every minute

7

© Rui L. Aguiar (ruilaa@det.ua.pt) - Uni. Aveiro

Lots of multimedia

TOTAL MOBILE VOLUME

	Category	Total Volume
1	Video	67.60%
2	Social Networking	12.16%
3	Messaging	5.89%
4	Web Browsing	4.51%
5	Marketplace	2.77%
6	Gaming	2.41%
7	File Sharing	1.97%
8	Cloud	1.79%
9	VPN	0.79%
10	Audio	0.11%

DOWNSTREAM VOLUME

	Category	Total Volume
1	Video	70.35%
2	Social Networking	12.27%
3	Messaging	4.78%
4	Web Browsing	3.83%
5	Marketplace	2.86%
6	Gaming	2.43%
7	File Sharing	1.77%
8	Cloud	1.06%
9	VPN	0.53%
10	Audio	0.12%

UPSTREAM VOLUME

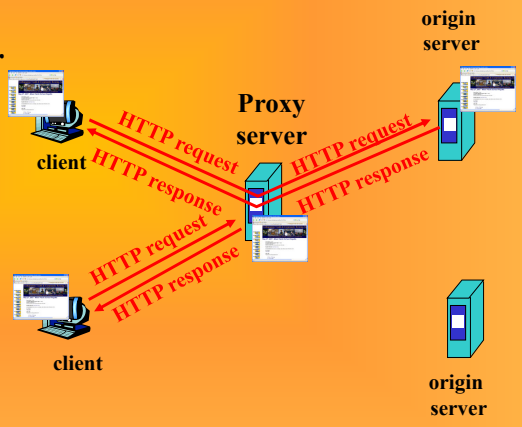
	Category	Total Volume
1	Video	37.11%
2	Messaging	18.23%
3	Web Browsing	11.95%
4	Social Networking	10.96%
5	Cloud	9.81%
6	File Sharing	4.27%
7	VPN	3.65%
8	Gaming	2.11%
9	Marketplace	1.82%
10	Audio	0.10%

8

(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



9

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.

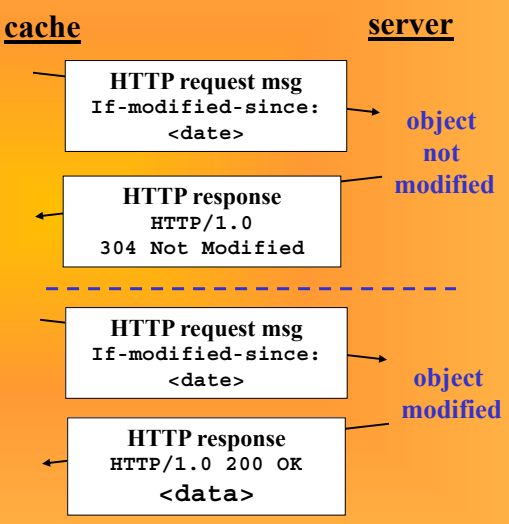
10

Conditional GET

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

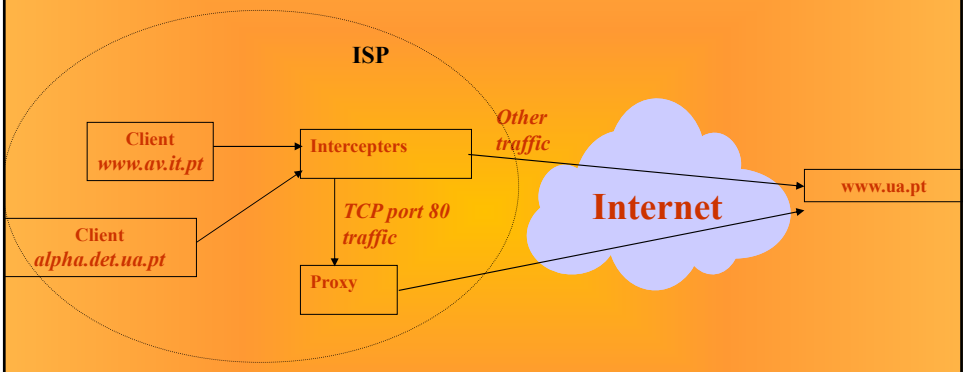
  `If-modified-since: <date>`
- server: response contains no object if cached copy is up-to-date:

  `HTTP/1.0 304 Not Modified`



15

Potential content network structure: Caching Proxies



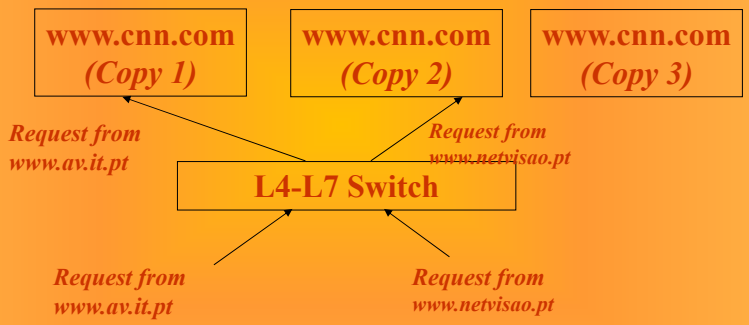
- Mostly motivated by ISP business interests – reduction in bandwidth consumption of ISP from the Internet
  - Reduced network traffic
  - Reduced user perceived latency

This was very popular some years ago in ISPs

20

Potential content network structure : Server Farms

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

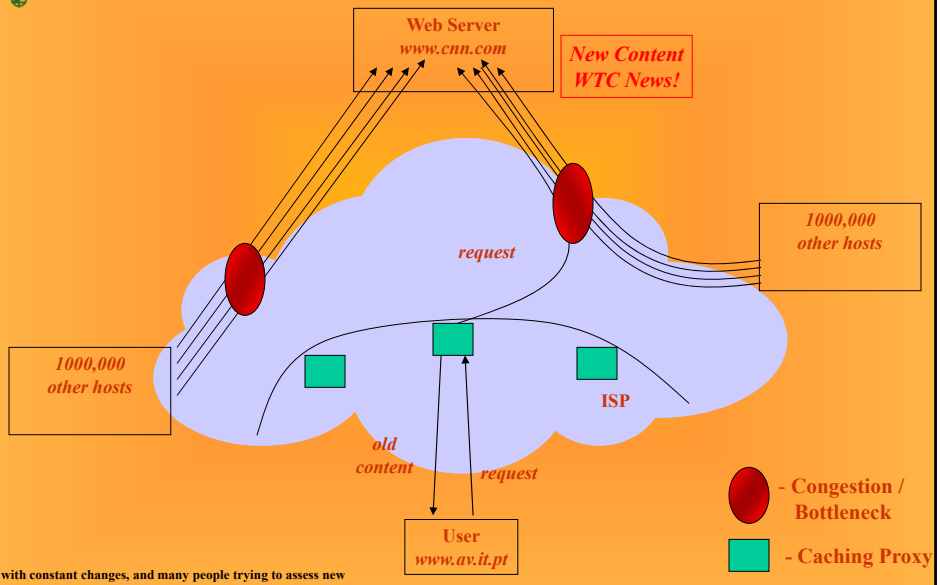


- **Arbitrate client requests to servers using an "intelligent" L4-L7 switch**
  - **Quite used today**

21

Flash Crowds:

Consider, On September 11, 2001



with constant changes, and many people trying to assess new content, the network will saturate. Caches to not cope with these scenarios. This also applies to personalized pages.

22

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 local 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

24

26

© Rui L. Aguiar (ruilaa@det.ua.pt) - Uni. Aveiro

CDNs

26

29

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

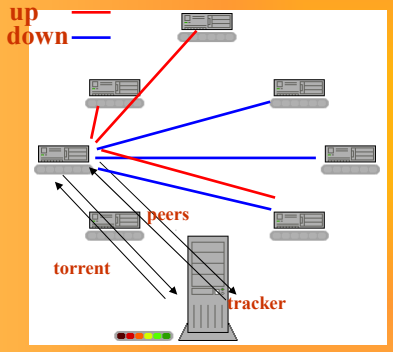
  For a large set of applications, including VIDEO access
- Improving the performance of networked applications
  - Handle highly dynamic pages or constant updates.
  - Use many sites at different points within the network
    - Stand alone servers
    - Routers

29

30

Peer-to-Peer?

- BitTorrent has peaked at ~30% Internet BW



1. Download a "torrent" file
2. Contact the tracker
3. Enter the "swarm" network
4. Chunk exchange policy
   - Rarest chunk first or random
   - Tit-for-tat: incentive to upload
   - Optimistic unchoking
5. Validate the checksums

Benefit: extremely good use of resources!

30

31

Could we use Peer-to-Peer to address the content explosion?

P2P has multiple problems:

× **Requires custom software**
- **Deployment is a must**
- **Configurations needed**

× **Companies usually may want managed service**
- **Handles flash crowds**
- **Handles long-lived objects**

× **Brings performance problem**
- **Hard to guarantee the service quality**
- **Delay in updating the content**

31

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 *(after 2nd generation)***
    - Executed at an edge server close to the user
    - Executed on a traditional application server

32

© Rui L. Aguiar (ruilaa@det.ua.pt) - Uni. Aveiro

33

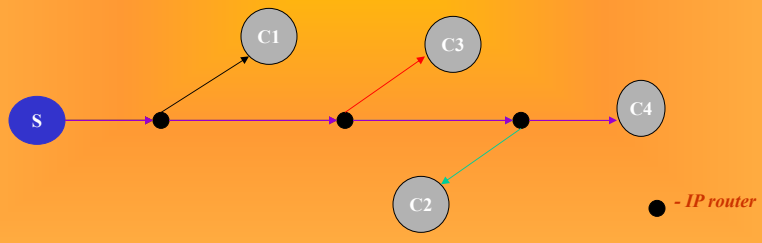
CDN Generations

- First generation (early 90ies): static caching
  - Accelerate the performance of web sites
  - Support increasing volumes of traffic
    - Key disruption event: 9/11
    - Akamai technologies created
- Second generation (early 2000ies): dynamic content
  - Support high volumes of multimedia traffic
  - Audio/video intensive networks
    - All ISPs developed/used CDNs
- Third generation (2010ies): cloud integration, SDN CDNs
  - Cloud computing
    - Amazon cloud (2008)
  - UGC (user generated content)
  - P2P and interactivity
    - AT&T distributed data centers (2011)
  - Mobile support, and device adapted content
- Fourth generation (2020+ *still ongoing*): edge and federation
  - Integrates with mobile environments
  - Edge devices (e.g. MEC concepts)
  - Multioperator federation

33

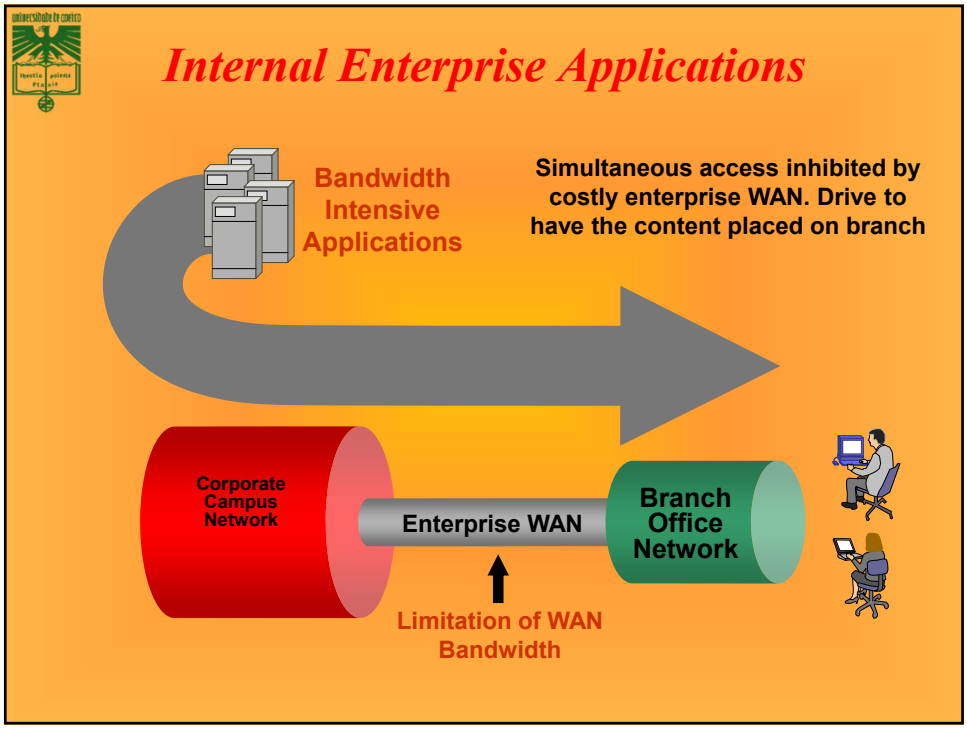
Early Motivations for Content Networks (1st generation)

- 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)*

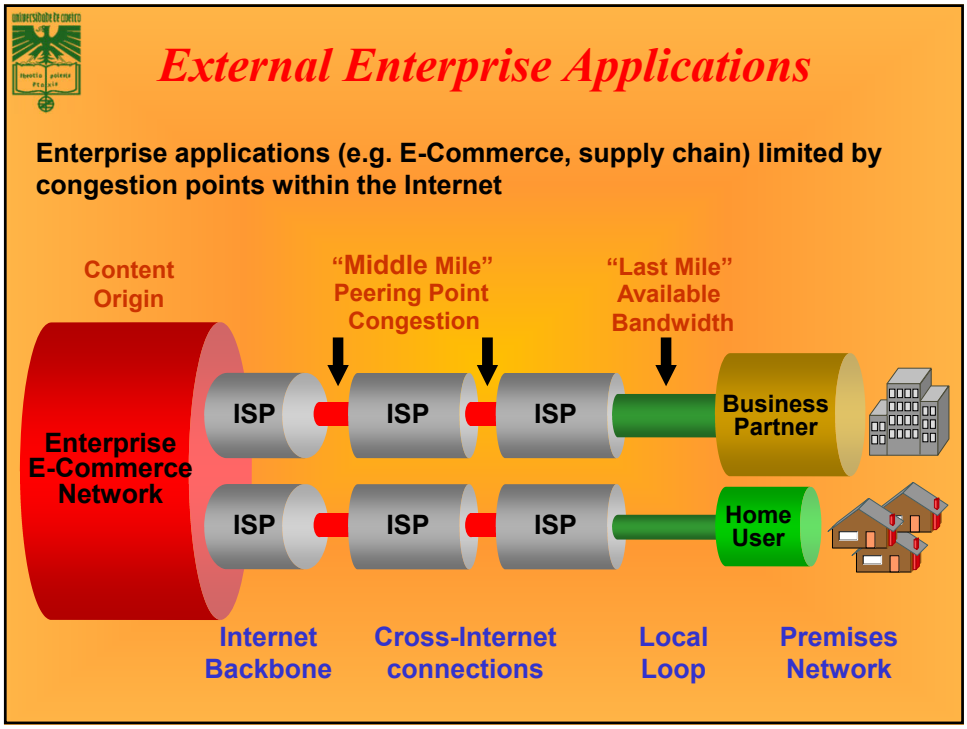


This was the simpler initial concept; immediatly network applications became of concern

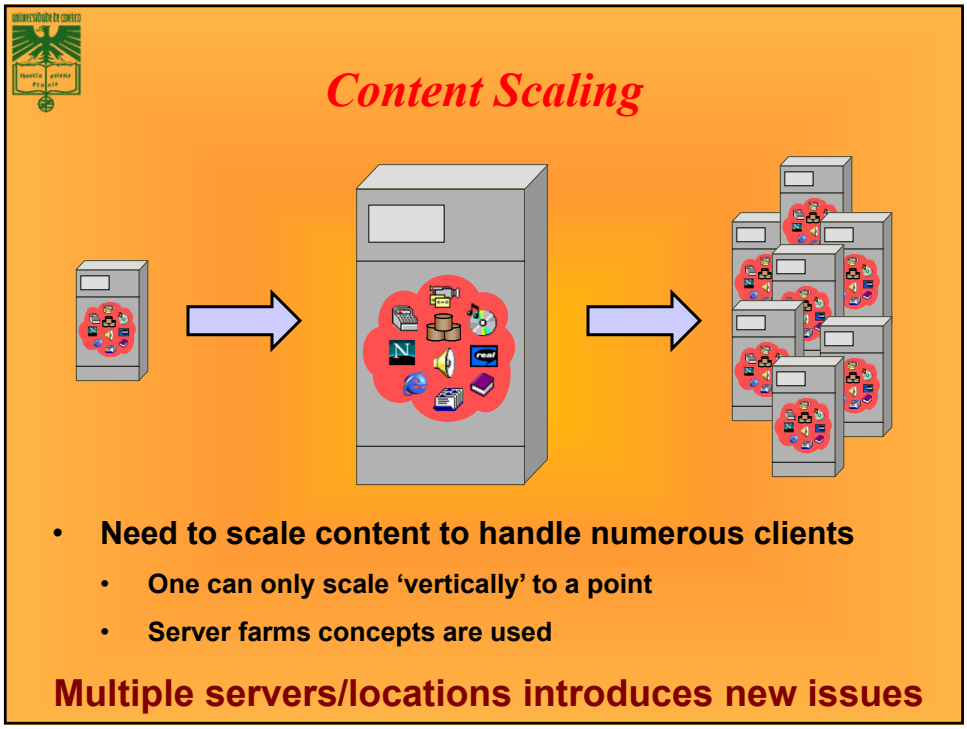
34


Internal Enterprise Applications
Bandwidth Intensive Applications
Simultaneous access inhibited by costly enterprise WAN. Drive to have the content placed on branch
Corporate Campus Network
Enterprise WAN
Branch Office Network
Limitation of WAN Bandwidth

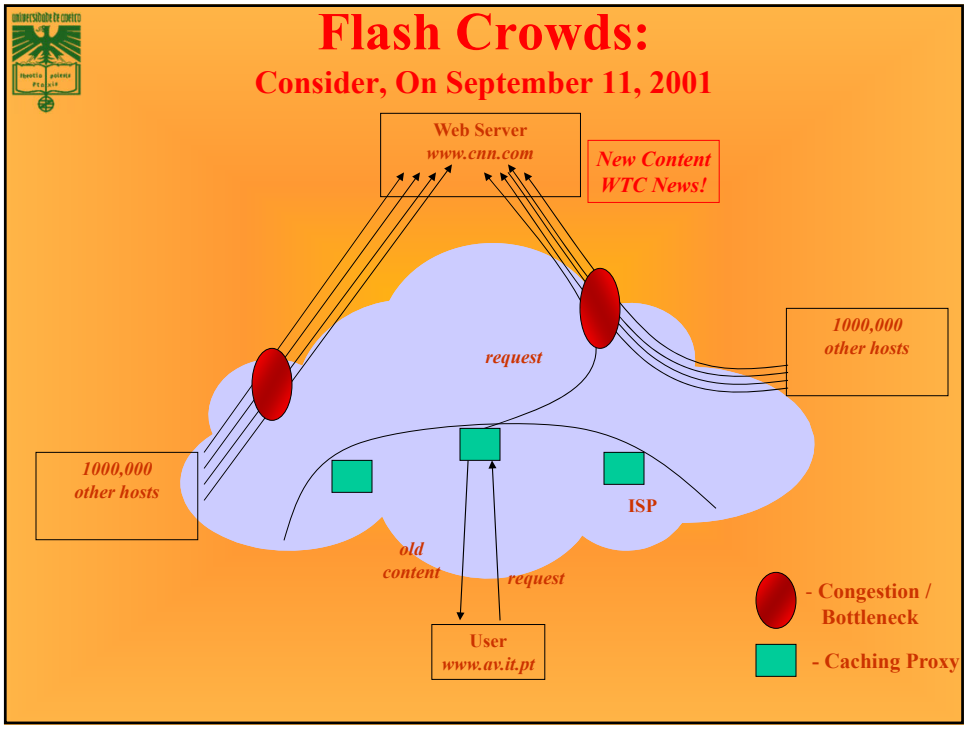
35


External Enterprise Applications
Enterprise applications (e.g. E-Commerce, supply chain) limited by congestion points within the Internet
Content Origin
"Middle Mile" Peering Point Congestion
"Last Mile" Available Bandwidth
Enterprise E-Commerce Network
ISP
ISP
ISP
Business Partner
ISP
ISP
ISP
Home User
Internet Backbone
Cross-Internet connections
Local Loop
Premises Network

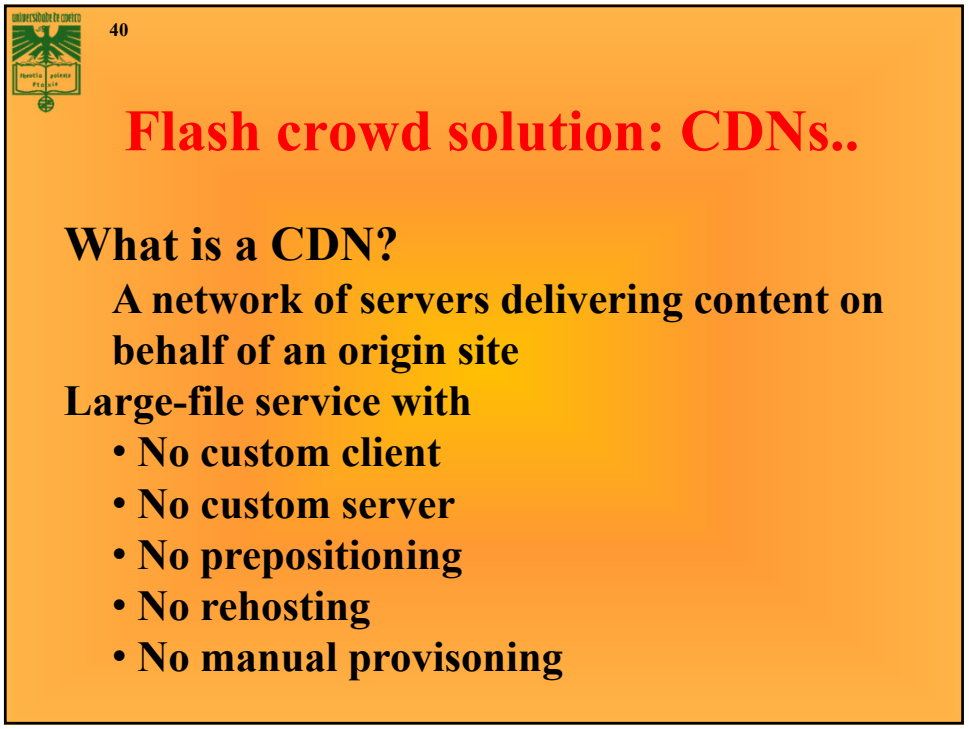
36


Content Scaling
• Need to scale content to handle numerous clients
• One can only scale 'vertically' to a point
• Server farms concepts are used
Multiple servers/locations introduces new issues

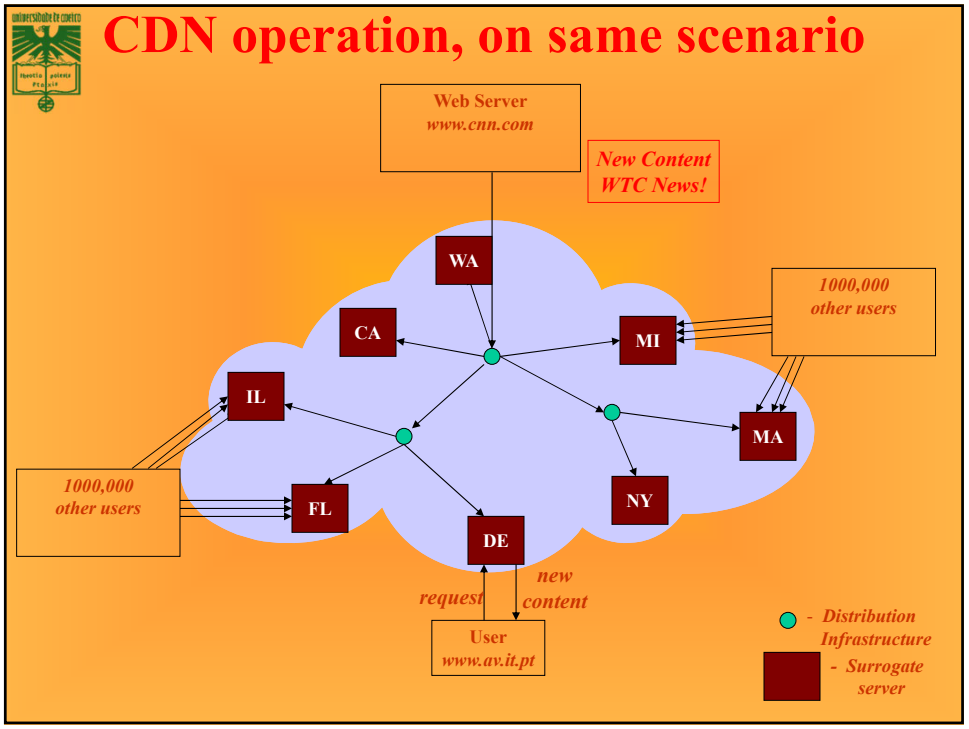
37


Flash Crowds:
Consider, On September 11, 2001
Web Server
www.cnn.com
New Content
WTC News!
1000,000
other hosts
request
1000,000
other hosts
ISP
old
content
request
User
www.av.it.pt
- Congestion /
Bottleneck
- Caching Proxy

39


40
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

40


CDN operation, on same scenario
Web Server
www.cnn.com
New Content
WTC News!
WA
CA
MI
1000,000
other users
IL
MA
1000,000
other users
FL
NY
DE
new
content
request
User
www.av.it.pt
- Distribution Infrastructure
- Surrogate server

41

42

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**

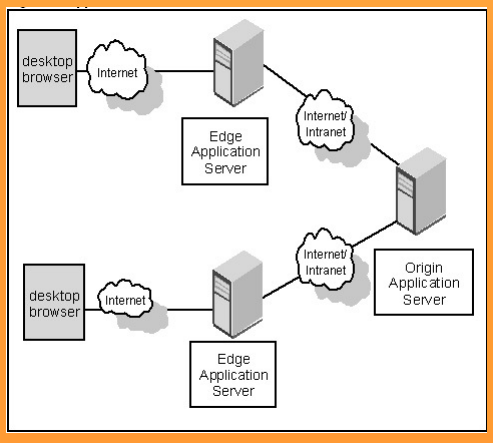
to ensure easy management and operational convenience.

The CDN has its own operation team, that needs to cover multiple places, even in different continents.

42

43

Model

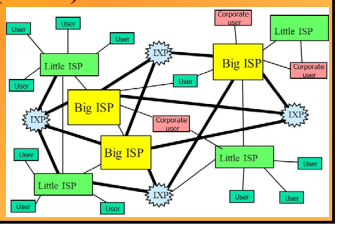


Application offload was the 1st generation unsolved challenge.

43

Inside a CDN

- Servers are deployed in clusters for reliability
  - Some may be (occasionally) 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 farms")
- Server locations
  - Well-connected points of presence (PoPs)
  - Inside of ISPs
  - IXP – internet exchange points



44

45

Content distribution networks

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

45

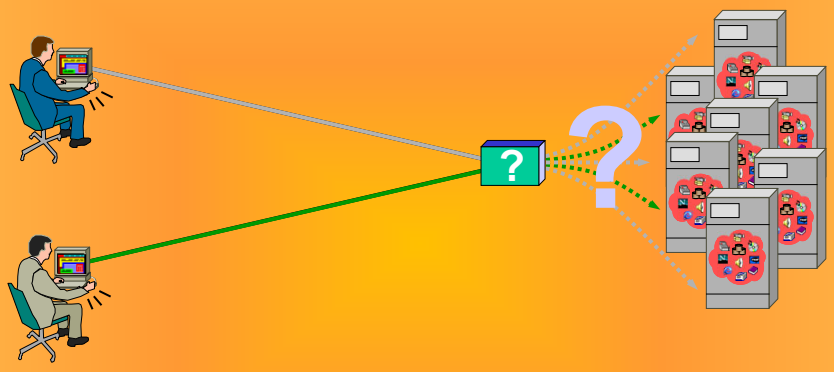
Challenges

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

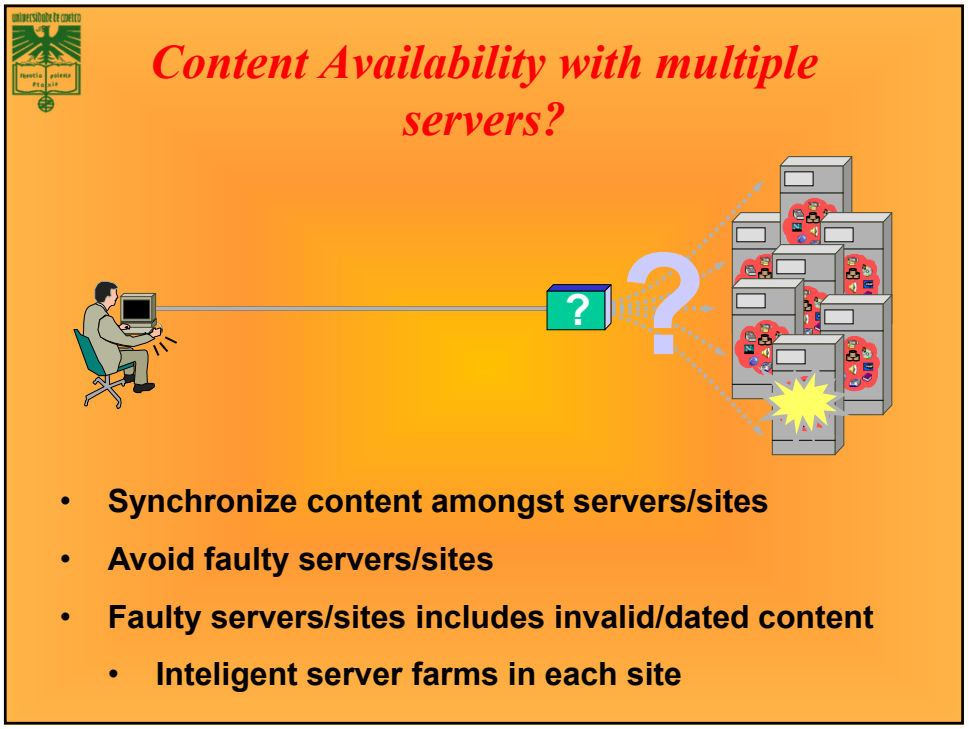
46

Load-Sharing Content

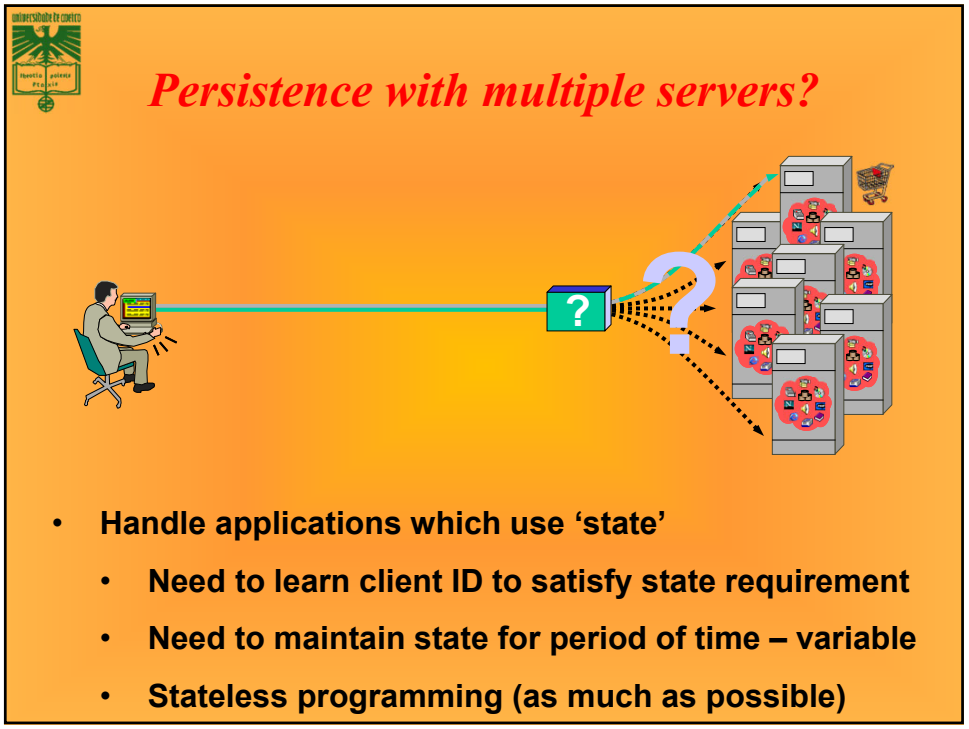


- Handle requests fairly amongst servers/sites
- Easily add servers/sites to content service
- Adjust connections based on server/site load

47


Content Availability with multiple servers?
?
Synchronize content amongst servers/sites
Avoid faulty servers/sites
Faulty servers/sites includes invalid/dated content
Inteligent server farms in each site

48


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
Stateless programming (as much as possible)

49

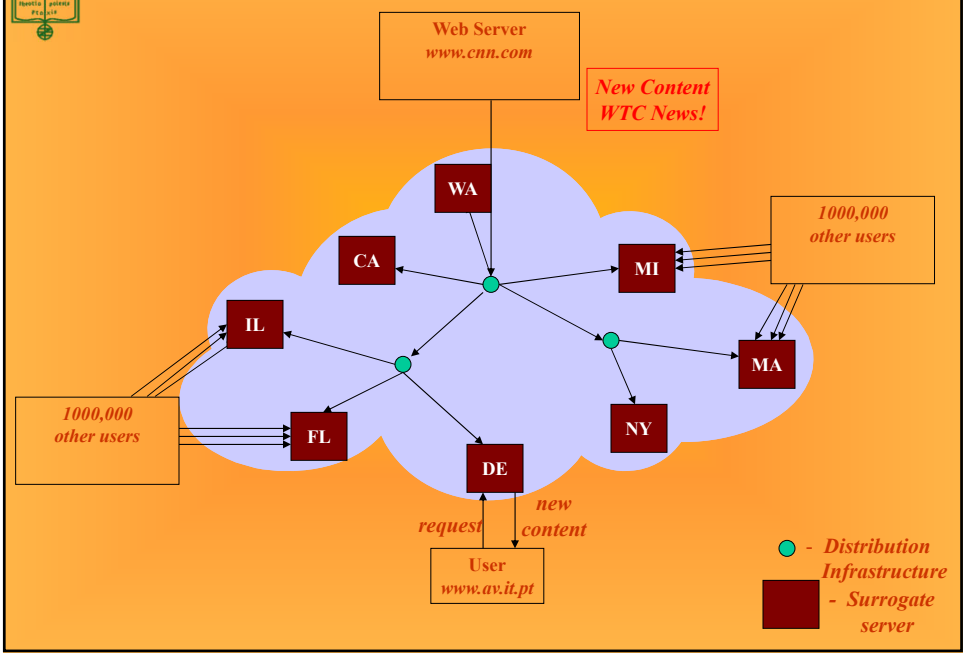
© Rui L. Aguiar (ruilaa@det.ua.pt) - Uni. Aveiro

Outline

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

50

CDN, On September 11, 2001



51

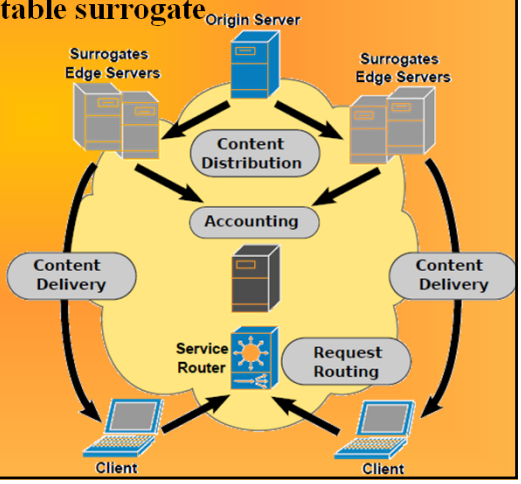
With CDNs

- 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

52

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



53

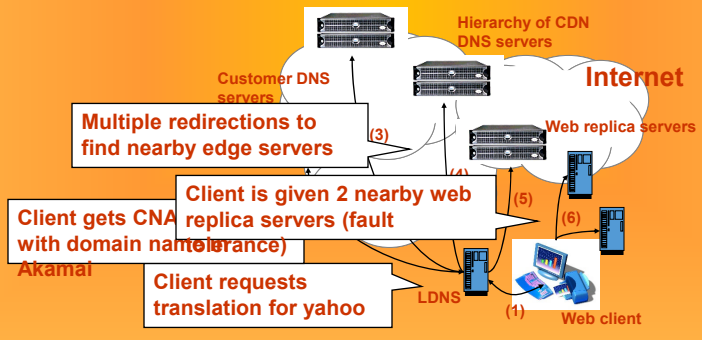
Mapping clients to servers

- 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

54

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



55

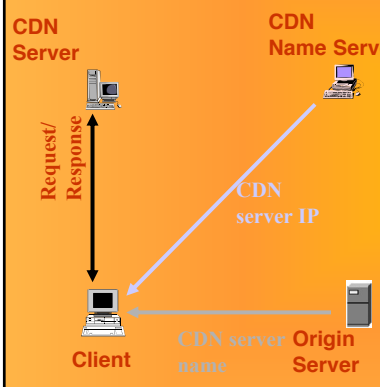
DNS Redirection Considerations

- **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

56

57

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

57

58

Offloading a portal

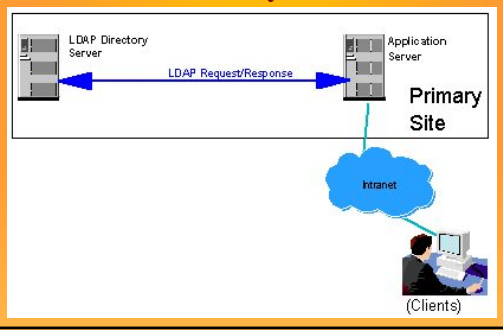
- Portal servers allow users to access content and applications from a single access point
  - Users can create persistent, customized views of applications and content chosen from the set of applications and content by the portal administrators
- Portal server pages are personalized
- Often include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

58

59

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

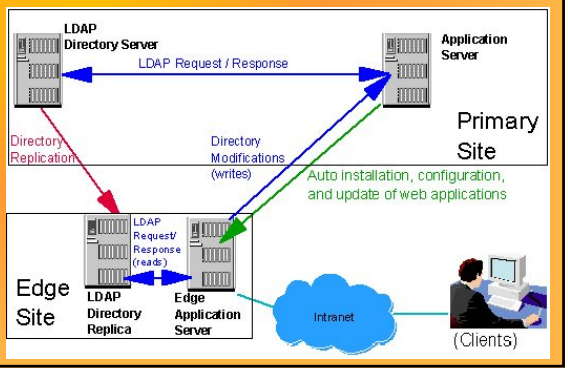


59

60

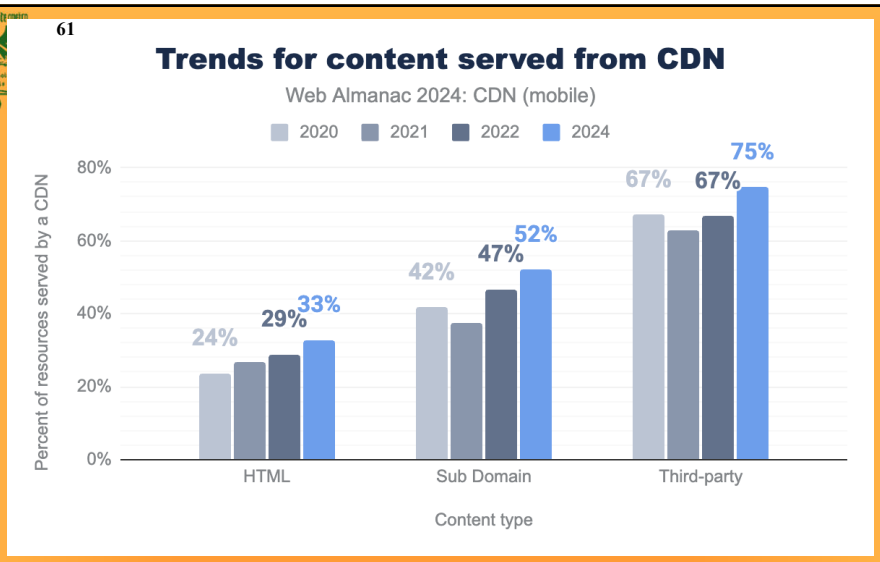
Offloading an Enterprise directory

- **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**
- **Challenge: aspects as *password update*. How to do it?**



60

61



Pages (~applications) containing different types of elements, with different ability to be CDNed

1. Base HTML page content
2. Embedded first-party content (such as images, css, fonts and javascript files on the main domain)
3. Third-party content (for example, Google Analytics, advertisements)

© Rui L. Aguiar (ruilaa@det.ua.pt) - Uni. Aveiro

61