Cache Management for TelcoCDNs

Daphné Tuncer
Department of Electronic & Electrical Engineering
University College London (UK)
d.tuncer@ee.ucl.ac.uk

Agenda

1. Internet traffic: trends and evolution
2. Content delivery models
3. Stakeholders: cooperation and challenges
4. ISP caches
5. Cache management strategies

Internet traffic forecast

- Based on Cisco VNI 2016 [1]
  - Consumer Internet video traffic to represent 82 percent of all consumer Internet traffic in 2020 (70 percent in 2015).
  - Internet video to TV doubled in 2015 and to fourfold by 2020.
  - Consumer VoD traffic to double by 2020 (equivalent to 7.2 billion DVDs per month).
  - Emergence and rapid growth of advanced video services such as Internet video surveillance and virtual reality traffic.
  - Traffic from wireless and mobile devices will exceed traffic from wired devices by 2019 (48% in 2015 and 66% in 2020).

Internet traffic in volume

- Traffic volume in petabytes (per month)

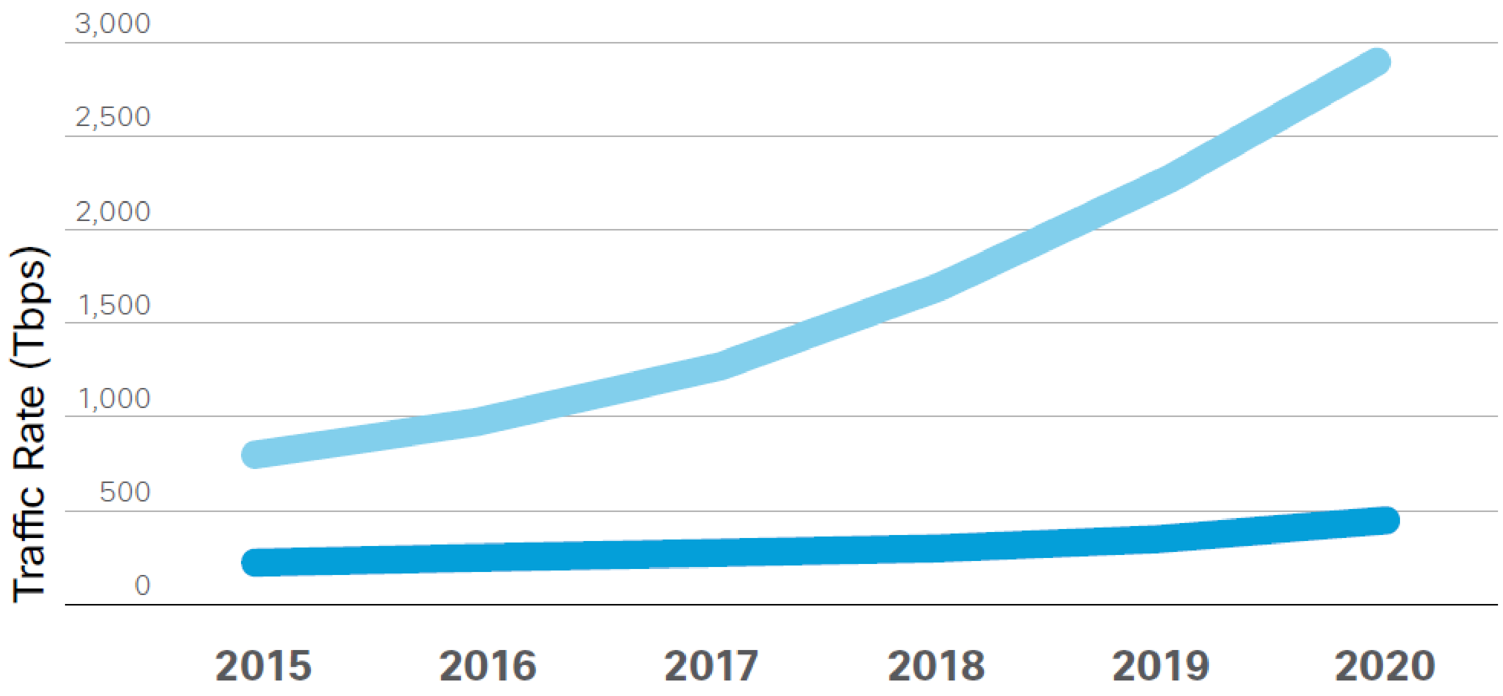
	2015	2020	Compound annual growth rate
Video	28 768	109 907	+31%
Web, email, data	7 558	17 006	+18%
File sharing	5 965	5 974	0%
Online gaming	82	568	+47%

Source: Cisco VNI 2016 [1]

Note: 1PB = 10^15 bytes

Bandwidth requirements

- Busy-hour compared with average Internet traffic growth



Source: Cisco VNI: The Zettabyte Era - Trends and Analysis, July 2016 [2]

Content delivery network

- Content distribution mainly relies on Content Delivery Networks (CDNs)
  - A CDN can be defined as "*a large, geographically distributed network of specialized servers that accelerate the delivery of web content and rich media to internet-connected devices*", Akamai [3].

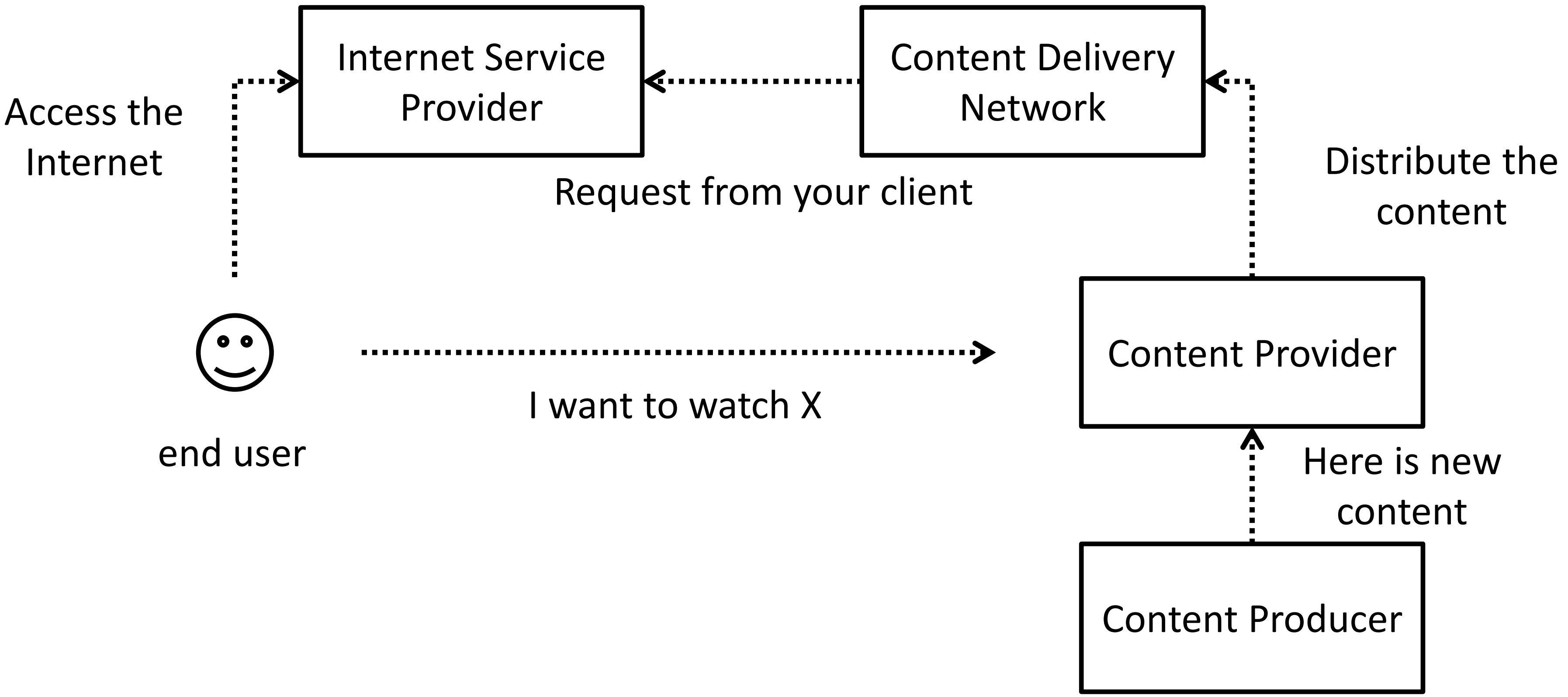
- Example of Akamai
  - More than 175,000 servers in more than 100 countries

- Content delivery network traffic will deliver three-fourth of all Internet video traffic by 2020 [1].

Content distribution solutions

- Commercial CDNs
  - *ex:* Akamai Technologies, Limelight Networks, Fastly, *etc.*
- ISP-operated CDNs
  - *ex:* AT&T Inc., Level 3 Communications, Deutsche Telekom, NTT, Telefonica, *etc.*
- Content provider-operated CDNs
  - *ex:* Netflix
- Peer-to-peer CDNs
  - *ex:* Coral Content Distribution Network

Stakeholders



CDN management operations

- Content placement
  - ➢ Decide on the distribution of content items in the different server locations.
- Server selection
  - ➢ Decide how to serve client requests.
- Usually taken without or with only limited knowledge of the underlying network conditions
  - ➢ Exert enormous strain of ISP networks

Impact for the ISP

- External costs
  - Internet tie costs
  - Decreasing trend but still significant given volume of traffic carried by CDNs

- Internal costs
  - Internal network upgrades
  - Upgrading a single router can amount in the order of tens of thousand dollars

Quality of Experience degradation

- Degradation of the Quality of Experience (QoE)
- Congestion and network failure lead to video playback issues (slow start, pixilation *etc.*) and buffering
- Severe effects on user experience
- The end user is more likely to contact his/her ISP than Netflix!

User (in)tolerance and QoE expectation

- Effect of poor resolution and/or frequent interruption on user

Tolerance (in min)	Percentage of abandonment
0 min	33%
1-4 min	43%
5-10 min	14%
11-30 min	5%
30+ min	3%

Source: Conviva 2015 [5]

ISP network caches

- Two solutions [4]
  - Partner caching
  - Transparent caching

Partner caches

- The Content Provider (CP) installs caches in the ISP's network.
- Caches are owned and maintained by the CP.
- Reduction of traffic on interconnect links.
- Internal traffic reduction strongly depends on the number of partner caches.
- Example: Netflix via OpenConnect

Transparent caches

- The ISP deploys its own caches used to locally cache most popular content items.
- Caching decision based on content popularity.
- Control messages between the client and the CP
  - Video statistics, ad views etc.
  - Essential for the CP's business
- Example: Mediacom using Qwilt
- Legal implications associated with caching third party content.

Partner caches vs. transparent caches (1/2)

	Partner caches	Transparent caches
Equipment cost	Free for the ISP	Investment needed by the ISP
Content coverage	• Can only cache content of specific CP • Good option only if one CP dominates	• Transparent to the CPs • Best option if many CPs of equal importance

Partner caches vs. transparent caches (2/2)

	Partner caches	Transparent caches
Source of revenue	No additional source of revenue for the ISP	New models involving the ISP
External and internal costs	Address external cost only (transit cost)	Address both external and internal upgrade costs but added complexity for the ISP

Other solutions

- Collaborative models such as CDNI (Content Delivery Networks Interconnection)
- Cloud-based services
- Towards ISP-operated CDNs?

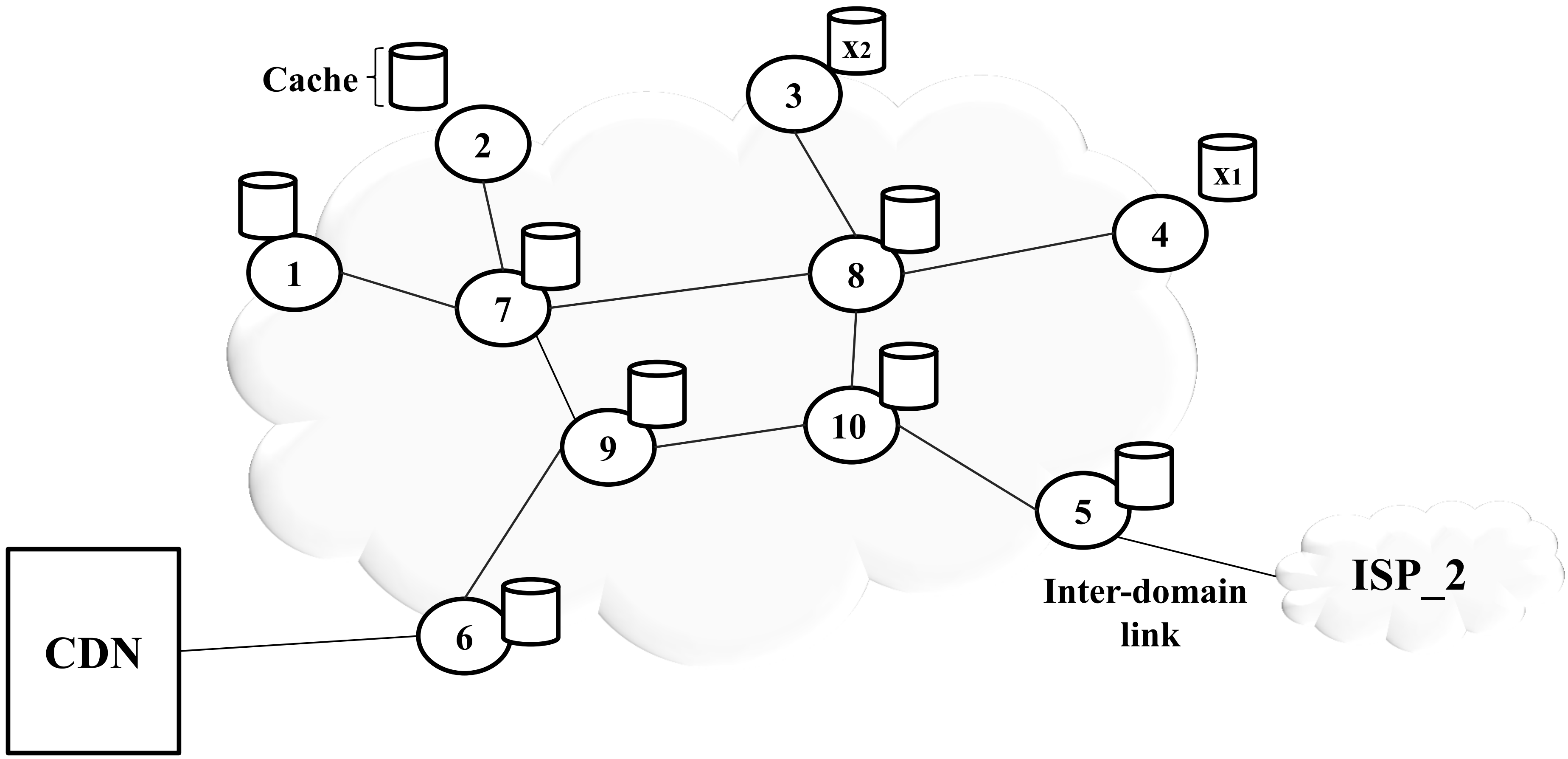
New technological opportunities

- Decreasing cost storage module
  - Enable network devices (i.e. access point, set-top boxes *etc.*) to be equipped with storage modules

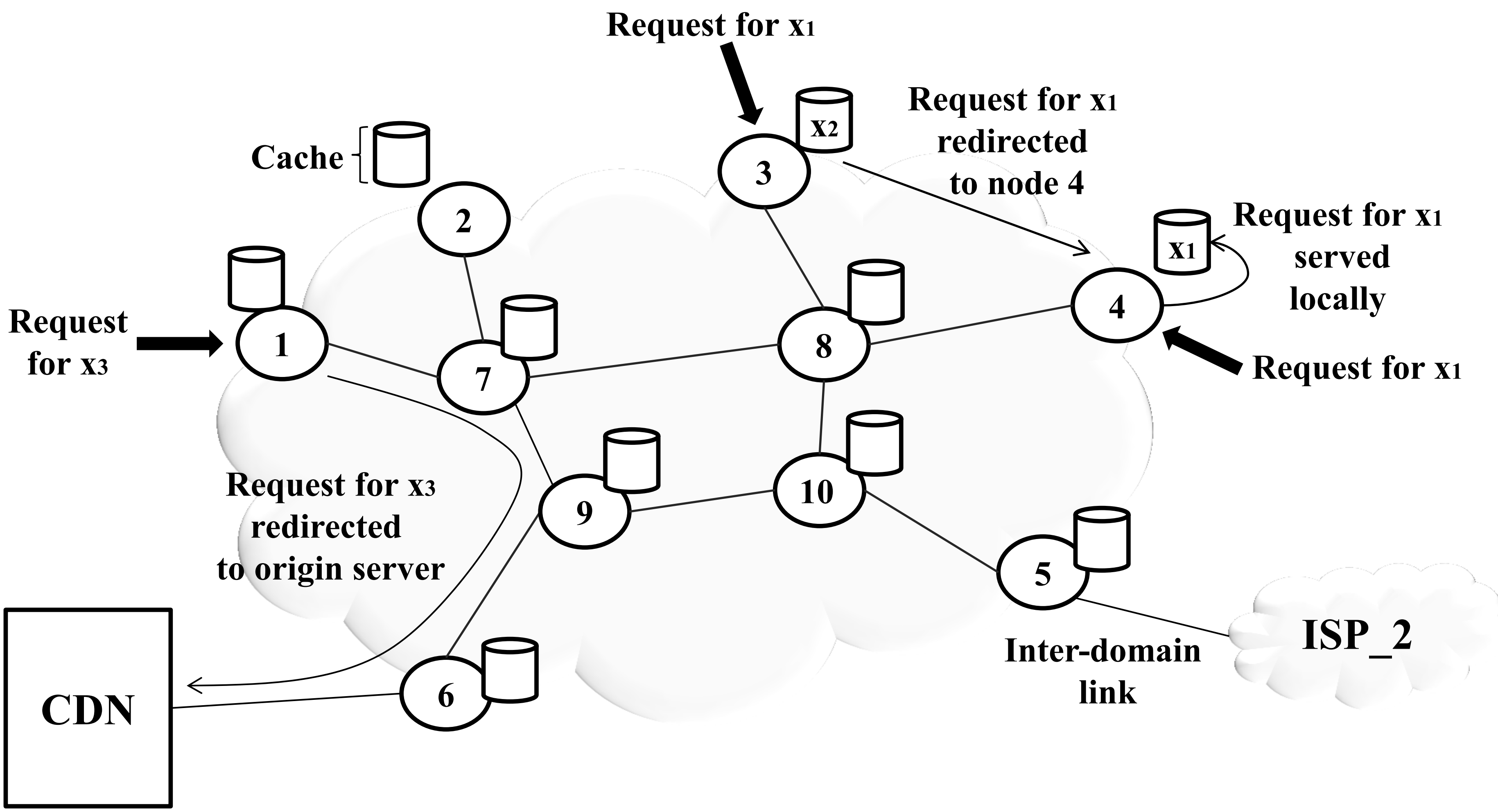
- Programming interfaces to network devices

- Virtualisation
  - Not only compute and storage resources but also network resources
  - Offer flexibility in managing the resources

Cache management strategies



Cache management strategies



Management operations

- Content placement
- Server selection

Content placement

- How to distribute the content items in the different cache locations?
  - Constrained by the available caching capacity
  - Traffic cost equal zero if infinite capacity (unrealistic!!)
- Optimisation/Performance objective(s)
  - Reduce user perceived delay
  - Optimise use of internal resources
  - Reduce transit cost
  - *etc.*
- Reactive vs. proactive strategies

Reactive content placement (1/2)

- Each cache autonomously decides on the content items to (re)place.
- Two components:
  - Placement strategy
  - Replacement policy (ex: LFU, LRU)
- Dynamic system
  - Apply insertion and eviction decisions based on the content popularity evolution at each location
- Approach used by Facebook on its edge servers

Reactive content placement (2/2)

- Advantages
  - Very low complexity
  - Uncoordinated and local decisions
  - Relatively good cache hit ratio (i.e. number of requests server locally)

- Drawbacks
  - Can have an impact on network cost (i.e. link utilisation)
  - Cannot avoid few cache misses when a content becomes suddenly popular

Proactive content placement (1/2)

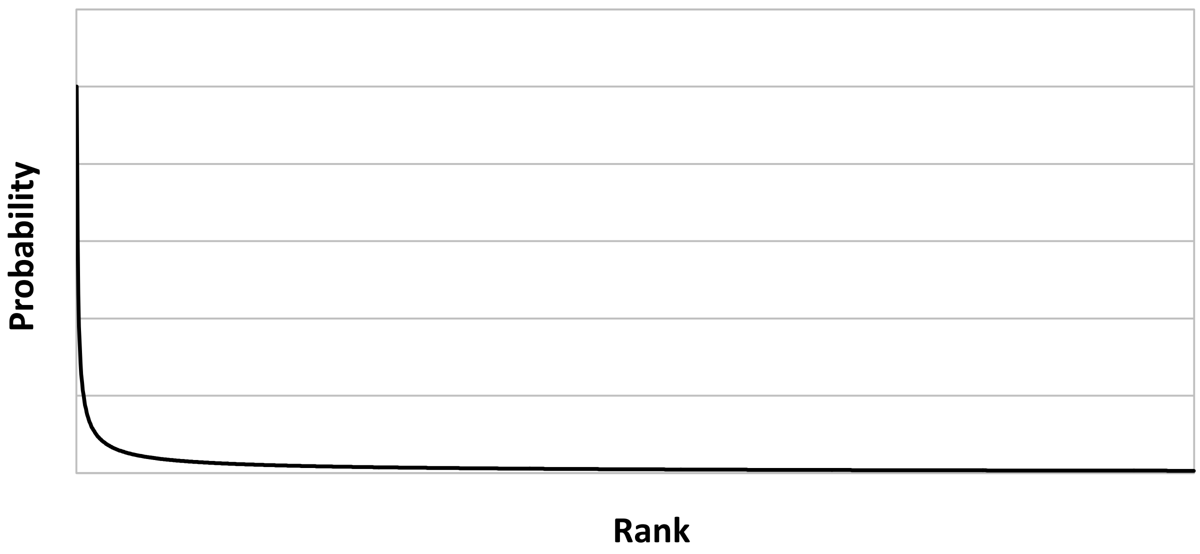
- The operator periodically decides on the location of the content items in the available caching location.
- The placement decisions are taken based on the prediction of content popularity for the next configuration period.
- New configurations are applied at medium to long timescale (in the order of few hours)
  - Generally once a day at night time during period of low resource utilisation
- Solution used by Netflix

Proactive content placement (2/2)

- Advantages
  - Fewer cache misses by provisioning the caches in anticipation to surge in popularity
  - The network cost can be taken as an optimisation parameter in the placement algorithm
- Drawbacks
  - The performance depends on the accuracy of prediction strategy
  - Higher management complexity
  - Migration overhead when provisioning the caches

Content popularity

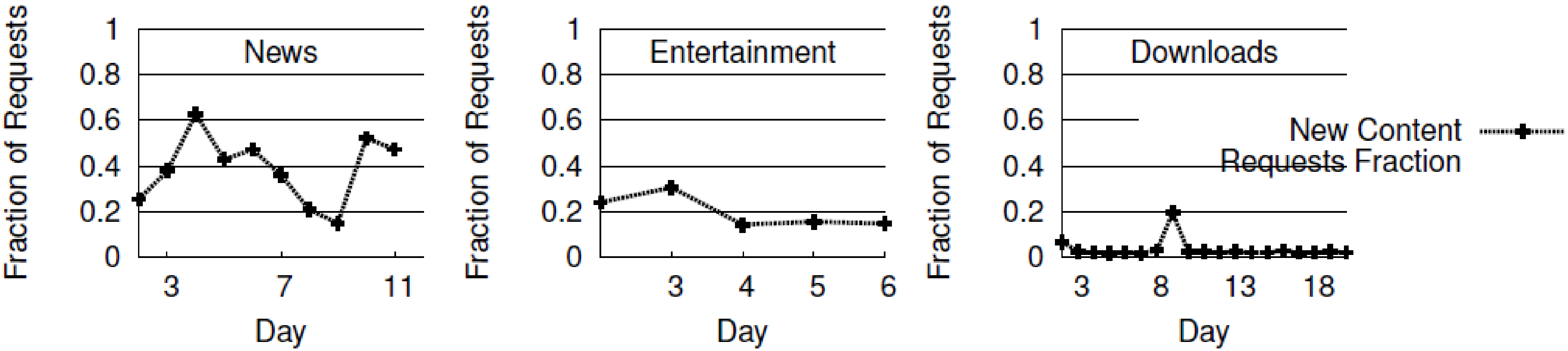
- The popularity is defined both temporally and spatially
  - ➢ Number of requests per content item (long tail distributed)



  - ➢ Content items requested at each location

Content popularity evolution

- The evolution of the popularity of an item over time strongly depends on the content type.



Source: A. Sharma et al. " Distributing Content Simplifies ISP Traffic Engineering, " SIGMETRICS'13 [6].

Example of series

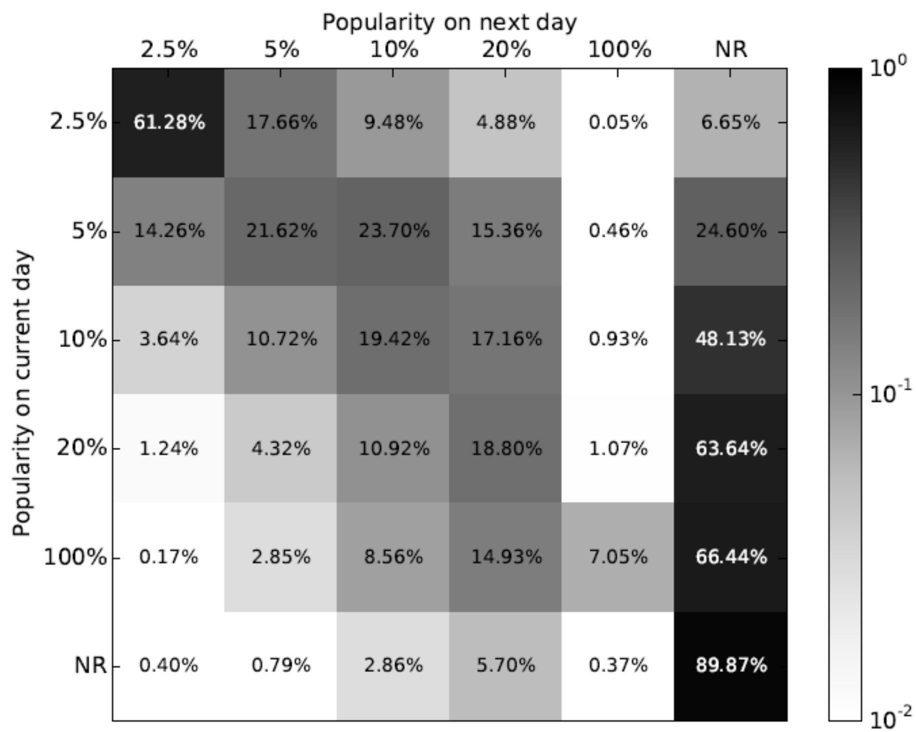
- To which extent do series viewers stick to a series?
- Behaviour of the viewers of series 1 (S1) when series 2 (S2) is released

Viewer behaviour	Percentage
Watch S1 and 2 together	59%
Put S1 on hold	25%
S2 replaces S1 if S2 is great	11%
Abandon S1	4%

Source: Conviva 2015 [5]

Predicting content popularity

- Limit of any prediction strategies
  - Some contents are inherently unpredictable
- Example on a real VoD trace



Source: M. Claeys et al. "Hybrid Multi-tenant Cache Management for Virtualized ISP Networks," JNCA 2016 [7]

Proactive approaches (1/2)

- Problem formulation

Given a set of M caches and a set of X contents, determine

  - *the number of copies of each content item to store in the network*
  - *the location of each copy*

in order to optimise some objective.

- Family of facility location problems

Proactive approaches (2/2)

- Different options to solve the problem
  - ➢ Integer Linear Programming (ILP)-based approaches
    + Optimal solution for the input parameters
    - Does not scale well
  - ➢ Heuristics (*e.g.* greedy approaches)
    + Computationally more efficient than ILP approaches
    - Sub-optimal solutions

- CDNs usually apply proprietary algorithms (*e.g.* Akamai, Netflix)

Server selection (1/2)

- To decide on the best server location to serve client requests
  - For scalability decisions are taken at the group of clients level.
- Different redirection mechanisms can be implemented
  - DNS-based
  - HTTP-based
  - Use of smart intermediaries
- DNS-based mechanisms remain the preferred method of industry leader, *e.g.* Akamai.

Server selection (2/2)

- Server selected based on different factors
  - Performance indicators, *e.g.* latency, packet loss, server load *etc.*
  - Business and regulatory restrictions

- Large scale monitoring systems required to build up-to-date map of the conditions.

- Decisions recomputed at the minute level.

Performance metrics (1/2)

<u>At the resource level</u>

- Network metrics
  - Network load
  - Link utilisation
  - Retrieval latency

- Cache metrics
  - Cache hit ratio
  - Cache occupancy ratio
  - Content replication degree

Performance metrics (2/2)

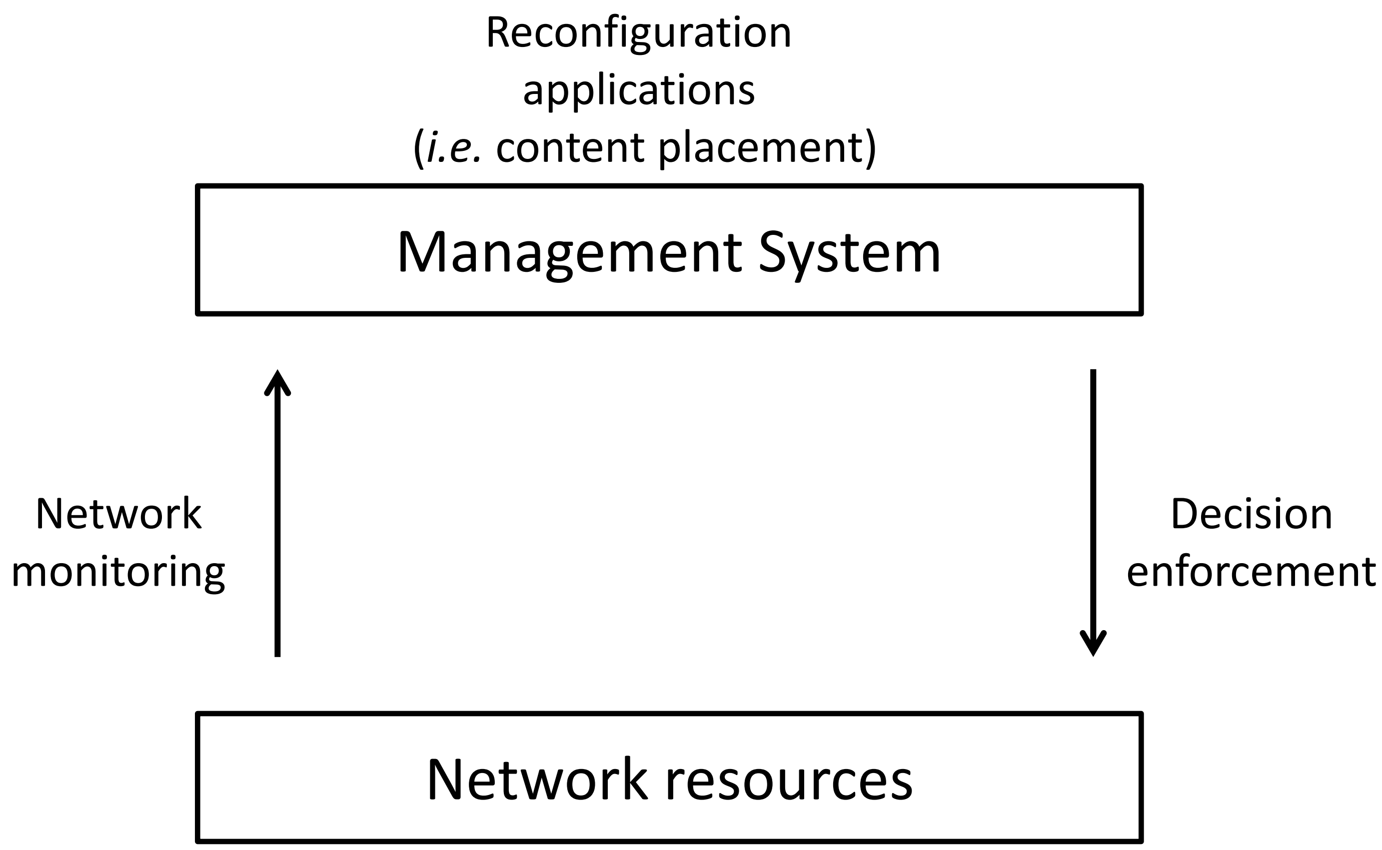
- Management costs
  - Signalling and monitoring overhead
  - Migration overhead
  - Algorithm complexity

- User metrics reflecting the QoE
  - Buffering ratio, start-up latency, average bitrate, frequency and duration of interruptions during playback *etc.*

Management system

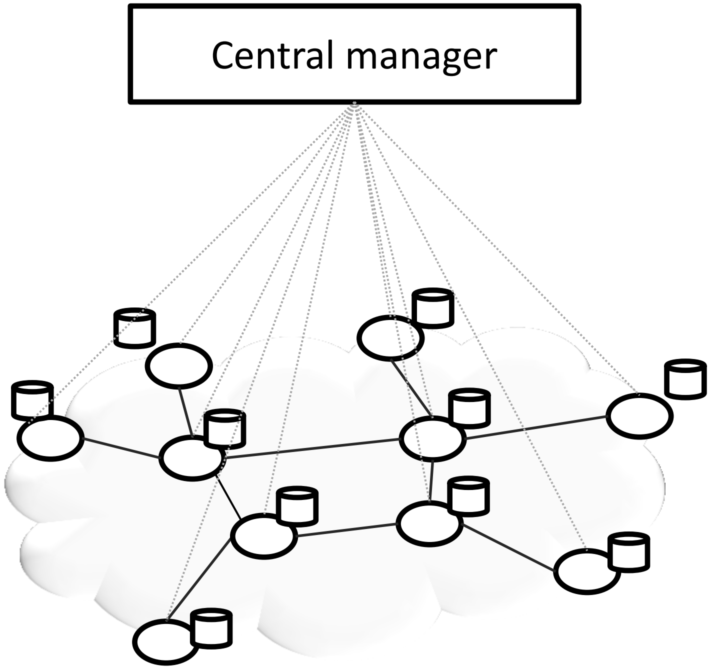
- How to implement cache management applications?

Management system model

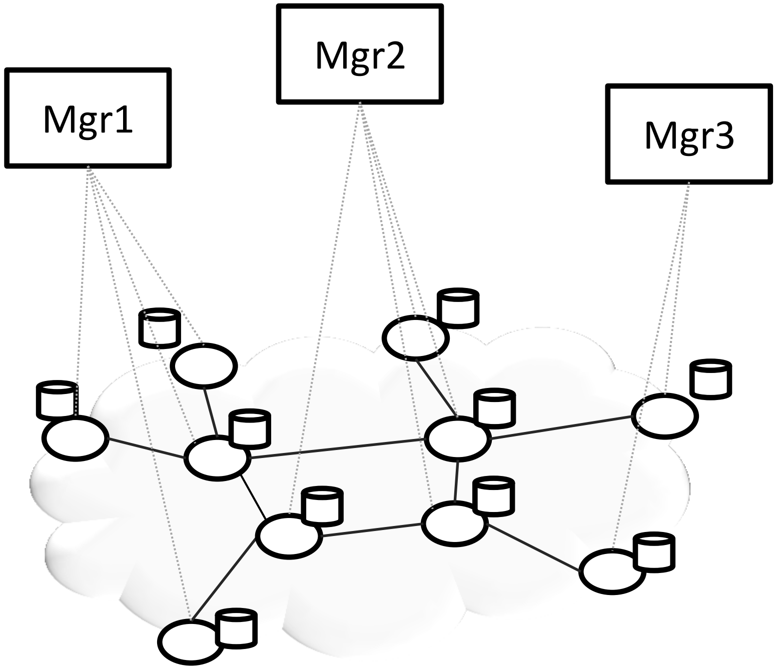


Centralised vs. distributed management (1/2)

Centralised system



Distributed system



Centralised vs. distributed management (2/2)

	Advantages	**Limitations**
Centralised management	Easy to implement Optimal solution	Single point of failure Does not scale well Not appropriate for dynamic system
Distributed management	Scale well Suitable for dynamic system	Higher implementation complexity Coordination

References

[1] *Cisco Visual Networking Index: Forecast and Methodology*, 2015-2020, June 2016, White Paper

[2] *Cisco Visual Networking Index: The Zettabyte Era -Trends and Analysis*, July 2016, White Paper

[3] *Akamai Technologies*, https://www.akamai.com/us/en/resources/content-distribution-network.jsp

[4] Colin Dixon, " *Handling the explosion of online video: why caching is the key to containing costs*, " October 2013, nScreenMedia

[5] Conviva.com, *Binge Watching, The New Currency of Video Economics*, 2015

[6] A. Sharma et al., "*Distributing Content Simplifies ISP Traffic Engineering*, " in proc. ACM SIGMETRICS ’13, 2013, pp. 229–242.

[7] M. Claeys et al., "*Hybrid Multi-tenant Cache Management for Virtualized ISP Networks*," Journal of Network and Computer Applications (JNCA), Volume 68, pp. 28-41, June 2016.