Andrea Araldo
www.mit.edu/~araldo

19 Avril 2018
Paris

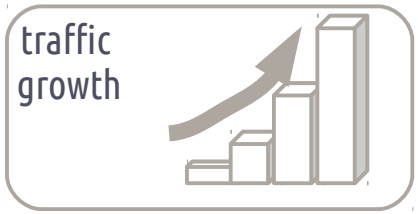
Distribution de contenu sur Internet : coût opérationnel, transmission multimédia et communication chiffrée

Massachusetts
Institute of
Technology

about me

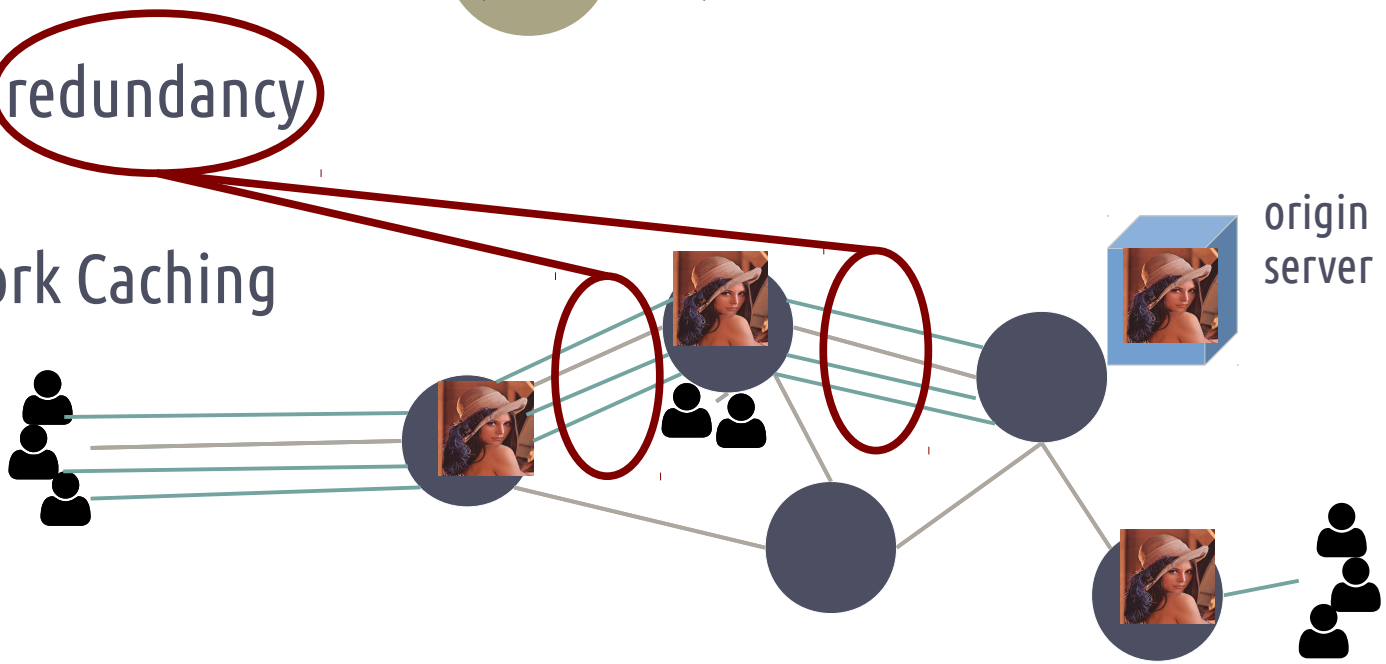
Where	What	Topic	Software
	MSc and BSc in Computer Engineering		
	Research Engineer	SDN and ICN	CoNET
	Internship	Internet Measurement	Bufferbloat Dissector
  	PhD + postdoc	Content Distribution in the Internet	ccnSim
	Postdoc	Intelligent Transportation Systems	. DynaMIT . SimMobility

traffic explosion and caching



- Internet traffic deluge [Cisco]

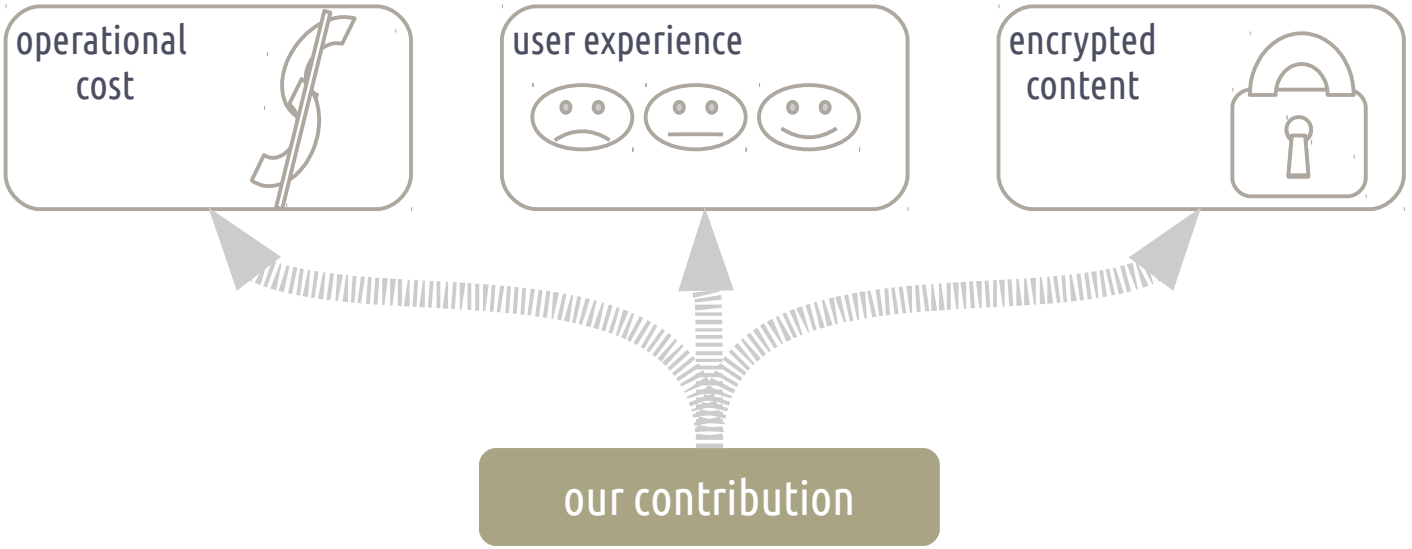
- Traffic has high redundancy
- Solution: Network Caching

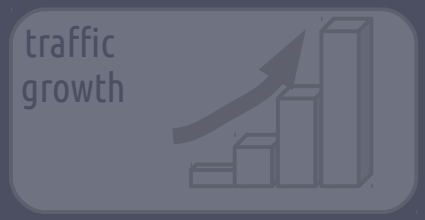


- Hit ratio: fraction of requests satisfied by the cache

[Cisco] Cisco Visual Networking Index: Forecast and Methodology, 2014-2019. (2012).

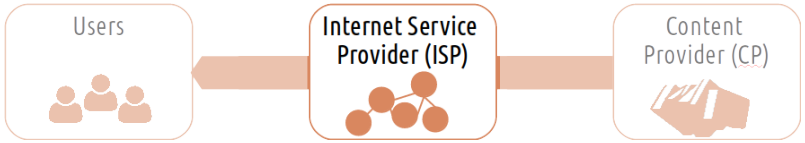
the challenges



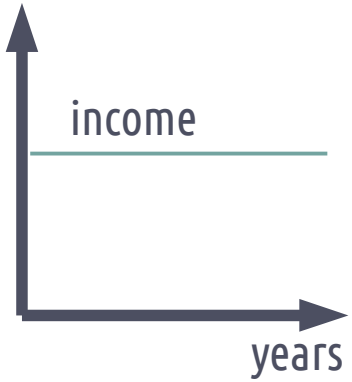


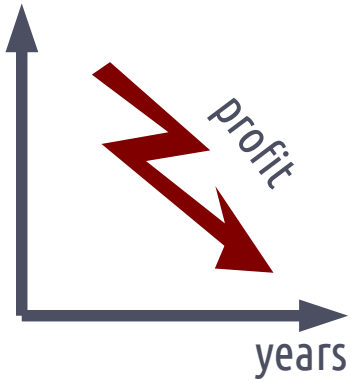
> cost-aware caching

the challenges



Flat rate fee + traffic increase => profit decrease [WB]





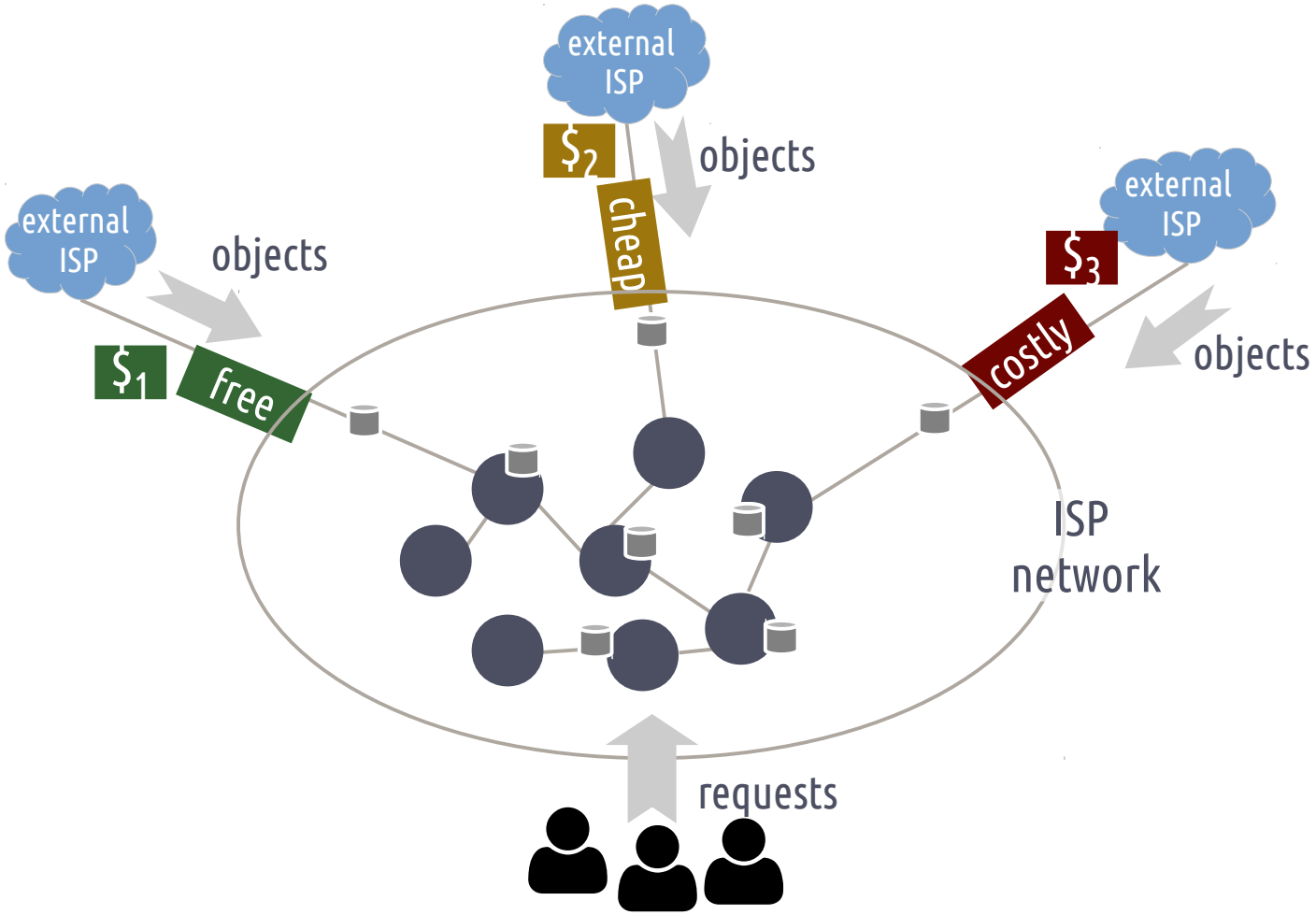
Caching to reduce ISP operational cost

[WB] The Little Data Book on Information and Communication Technology. (2015). The World Bank.

Internet Service Provider (ISP) cost

~~+ hit ratio => - cost~~

Price heterogeneity

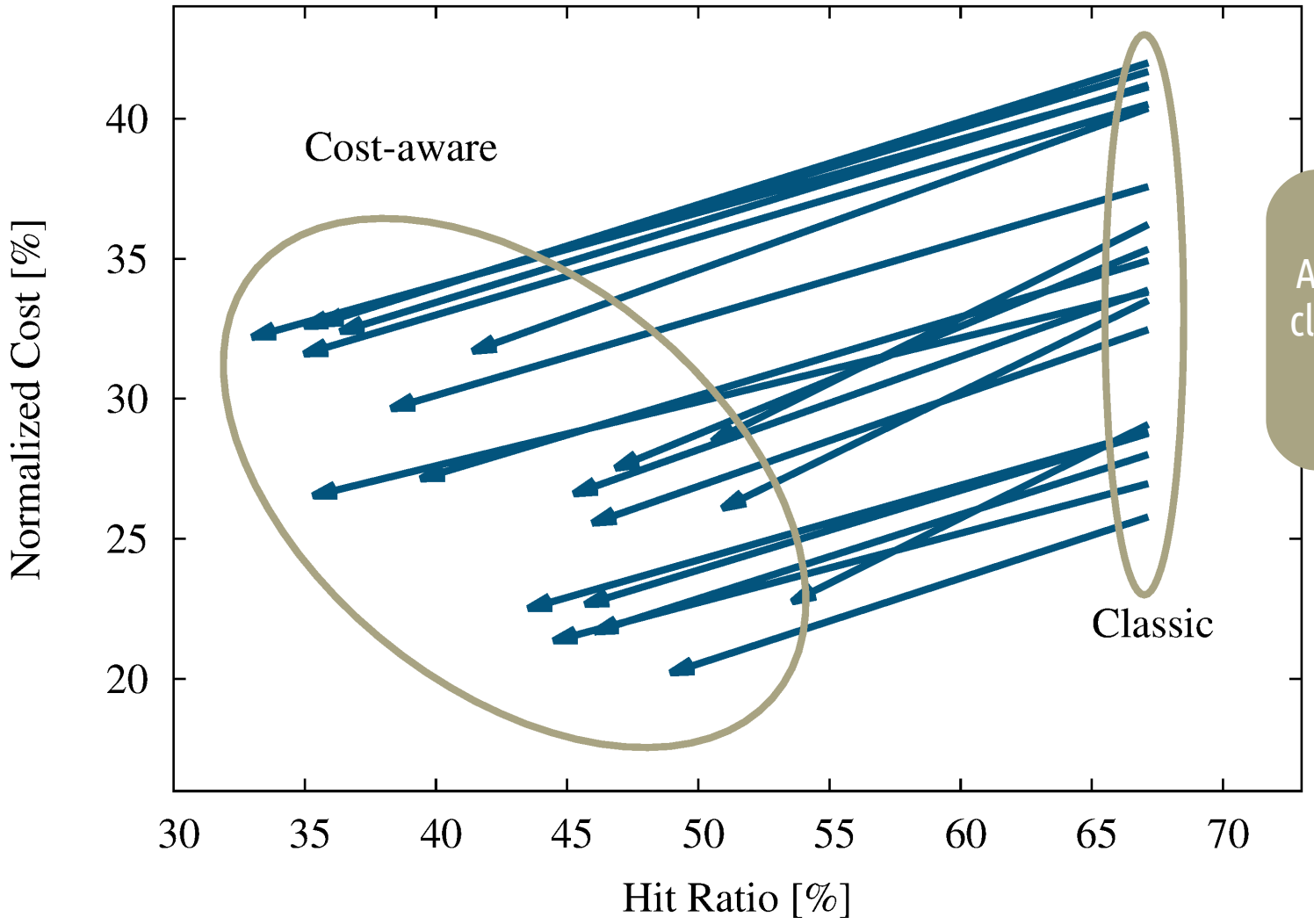


S : cache space
p_o : popularity of object o
$\$_o$: price of object o

Classic caching:
Store the S objects with the highest p_o

Cost-aware caching:
Store the S objects with the highest $p_o \$_o$

conflicting goals



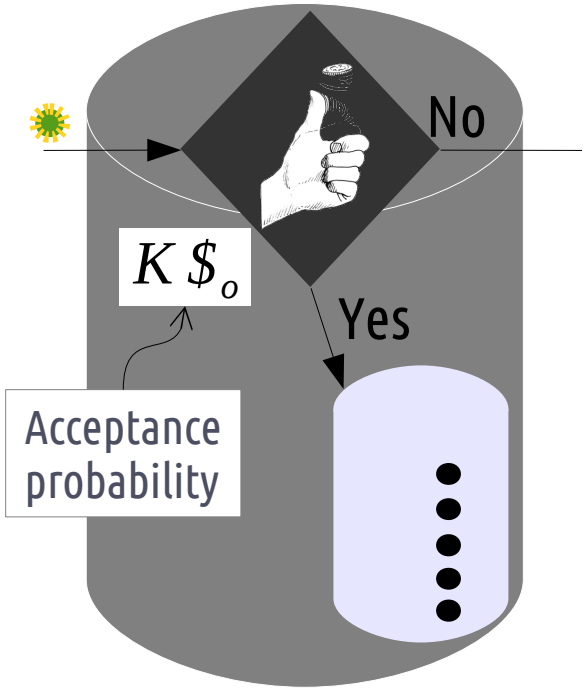
An ISP which performs classic caching is losing in terms of cost

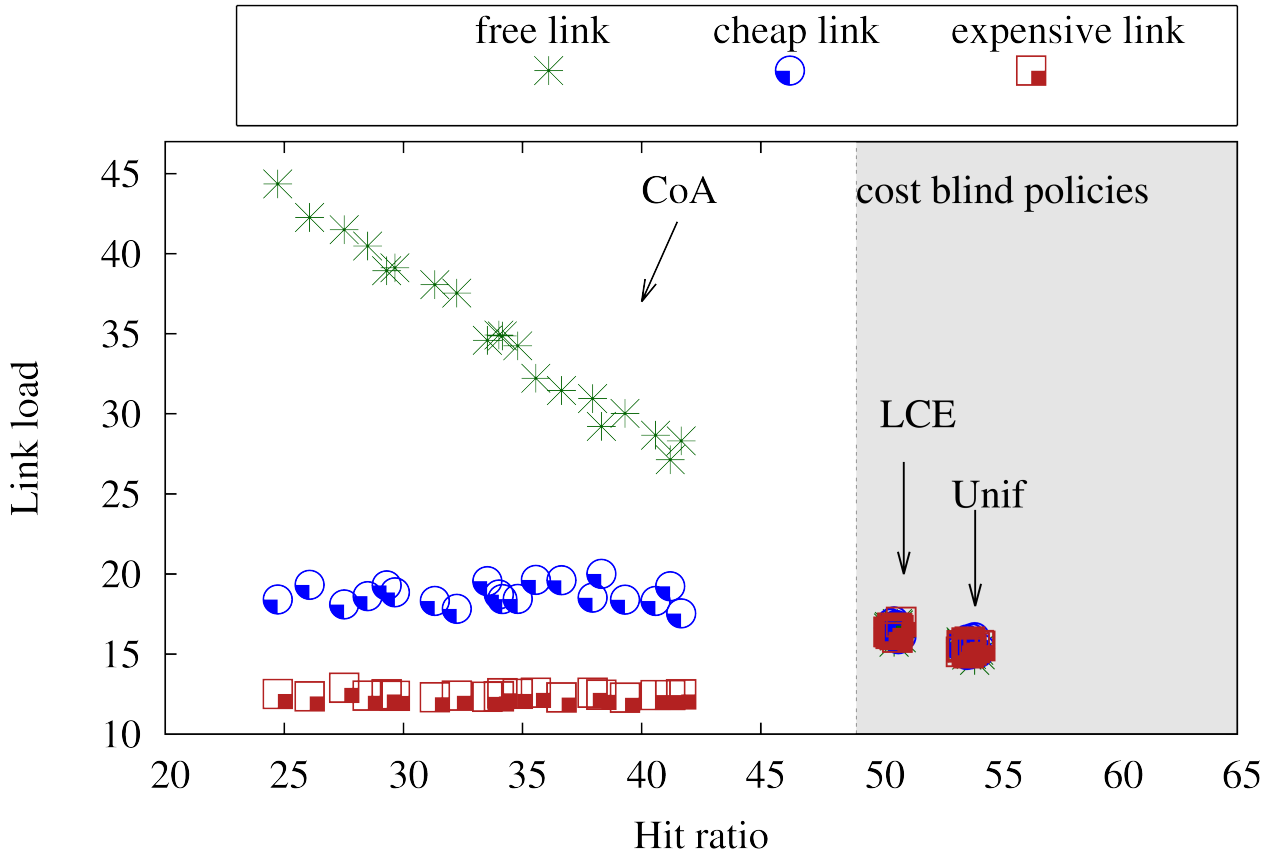
Scenario: catalog 10^5 objects, cache space 10^3 objects, Zipf exponent 1, prices: ($\$_1$,$\$_2$,$\$_3$) = (0,1,10)

online distributed policy

- In theory we should cache the $|S|$ objects o with highest value $p_o \bullet \$_o$
- But we do not know p_o in advance

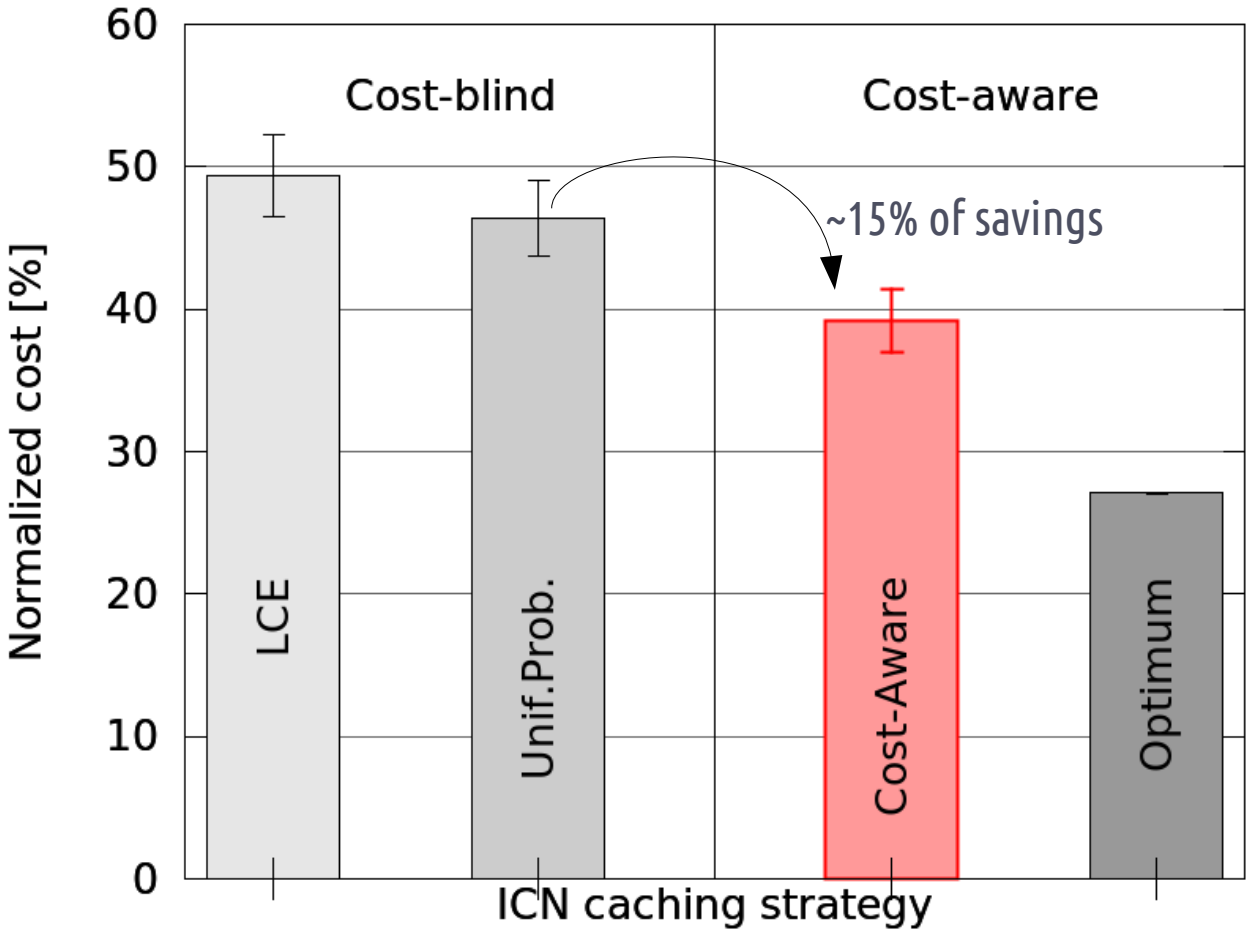
Our Cost-Aware caching preferentially caches expensive objects





Load: Downloads from that link / total requests

online strategy



Operational cost saving accumulates over years

A. Araldo, D. Rossi, F. Martignon, "Cost-aware caching: Caching more (costly items) for less (ISPs operational expenditures)", IEEE Transactions on Parallel and Distributed Systems (TPDS) 2016

Scenario: catalog 10^5 objects, cache space 10^3 objects, Zipf exponent 1, prices: ($\$_1$,$\$_2$,$\$_3$) = (0,1,10)

traffic growth

operational cost

user experience

secrecy

> caching and video delivery

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir
université PARIS-SACLAY

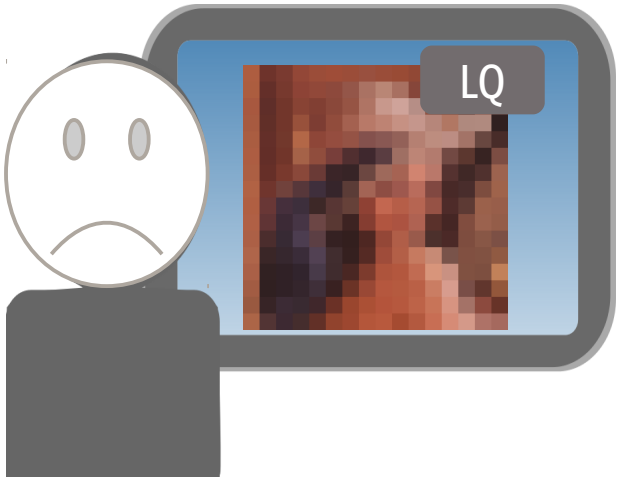
TELECOM ParisTech

cines

video caching

- 80% of traffic will be video [Cisco]
- Despite video is highly cacheable …
- Caching is not suitable for video

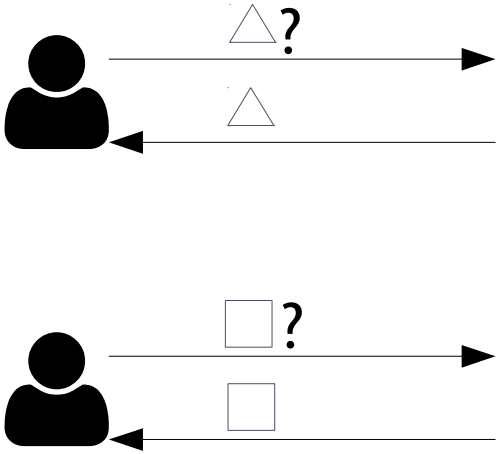
Caching to improve user experience, adapted to video delivery



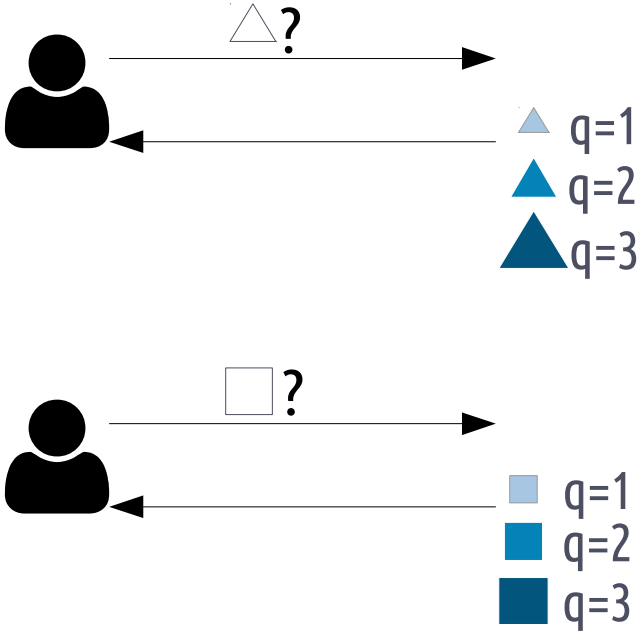
[Cisco] Cisco Visual Networking Index: Forecast and Methodology, 2014-2019. (2012).

a new dimension to the caching problem

- Classic caching
  - One request / one file



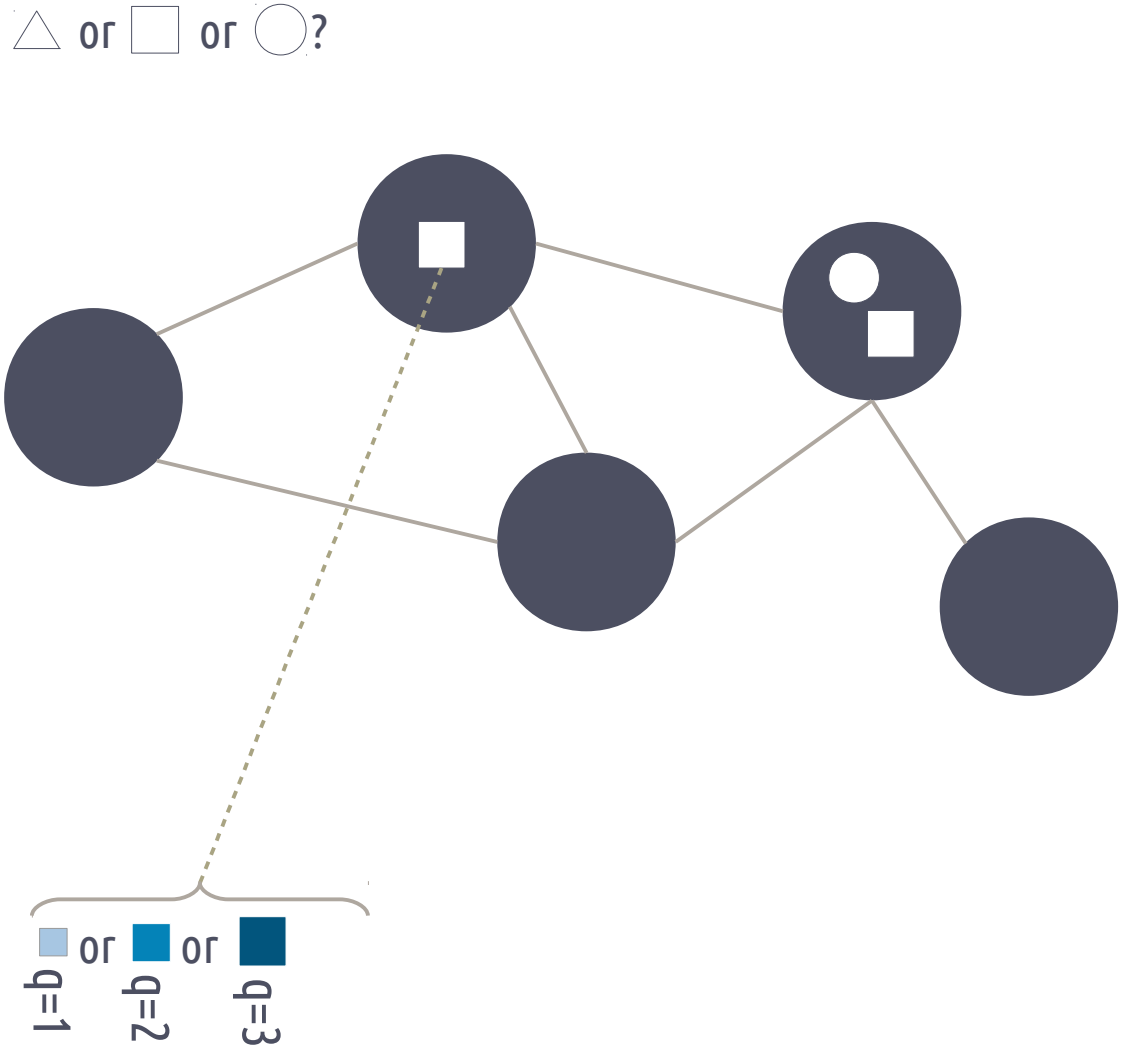
- Video
  - Different quality representations



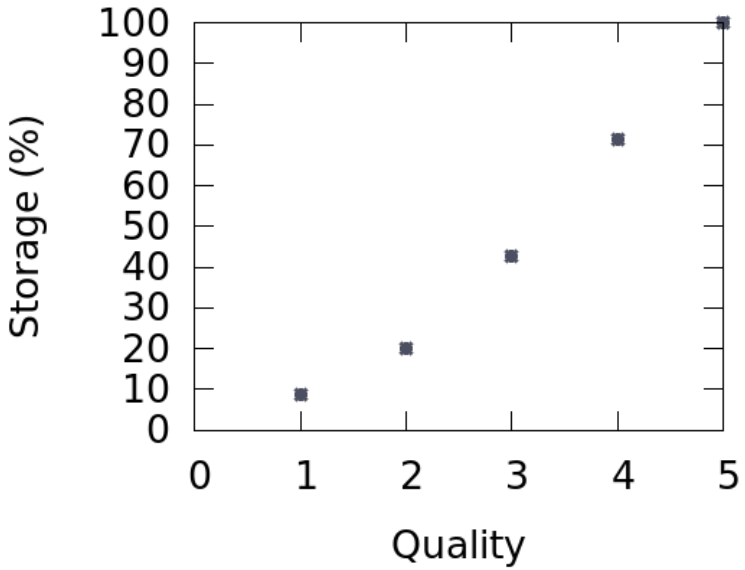
[2] Lee, D. H., Dovrolis, C., & Begen, A. C. (2014). Caching in HTTP Adaptive Streaming : Friend or Foe ? In ACM NOSSDAV.

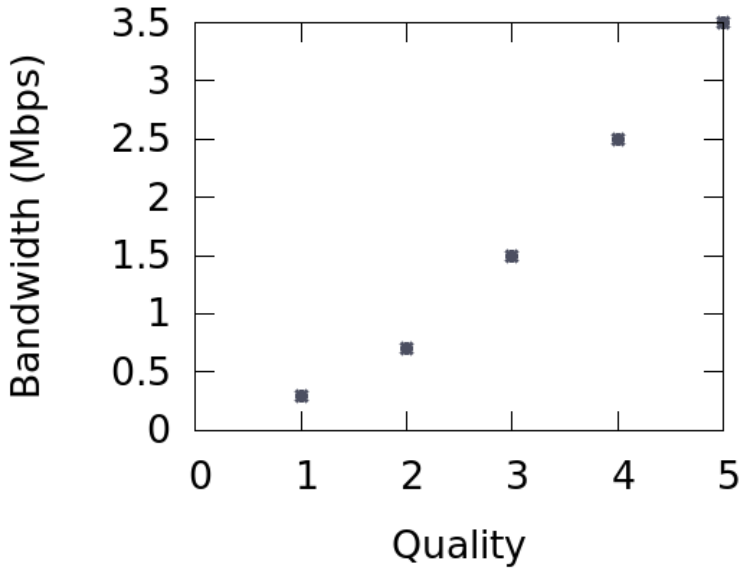
a new dimension to the caching problem

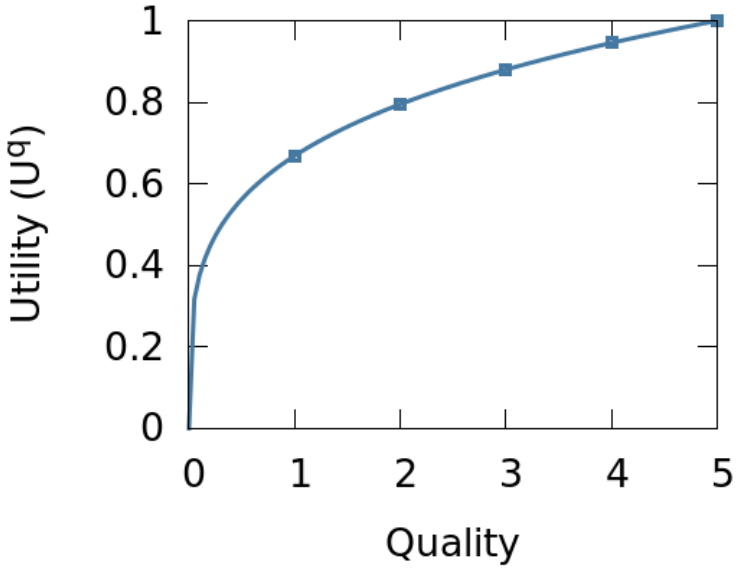
- Object selection
- Replica placement
- Quality selection



different quality representations

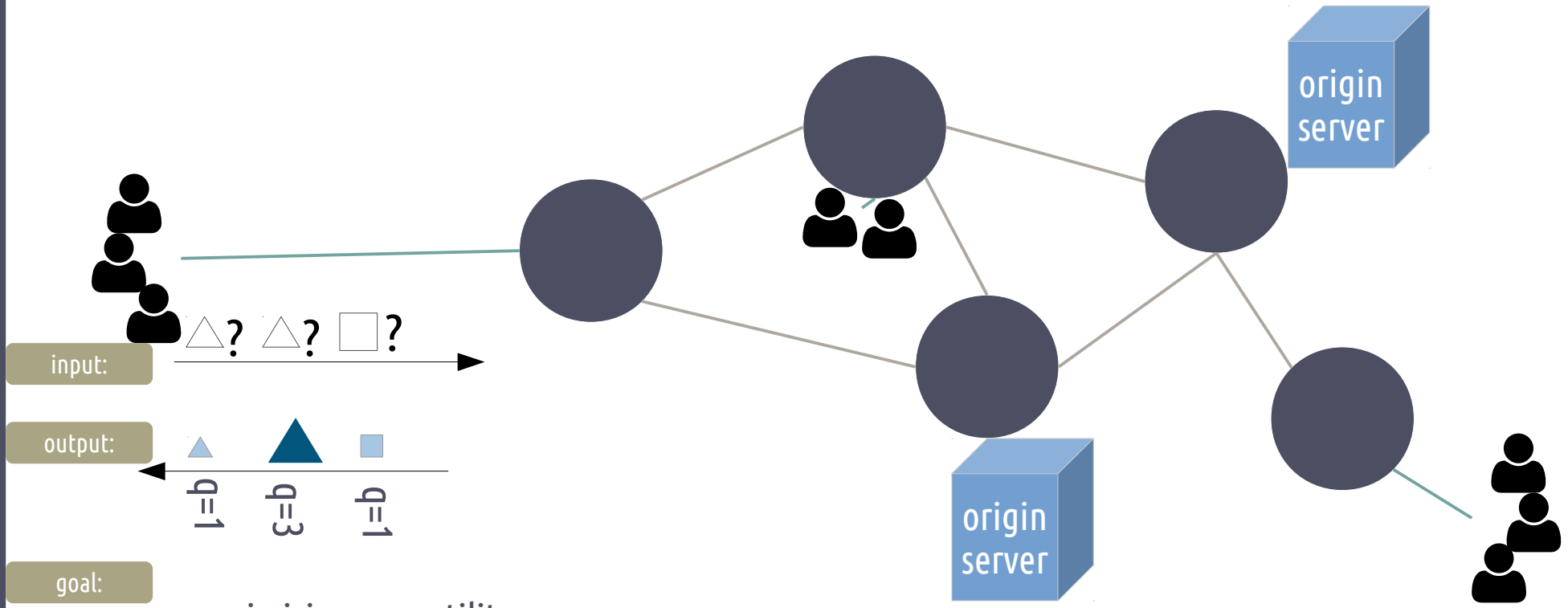






L. De Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo. ELASTIC:a Client-side Controller for Dynamic Adaptive Streaming over HTTP(DASH). In IEEE Packet Video Workshop (PV), 2013.

objective

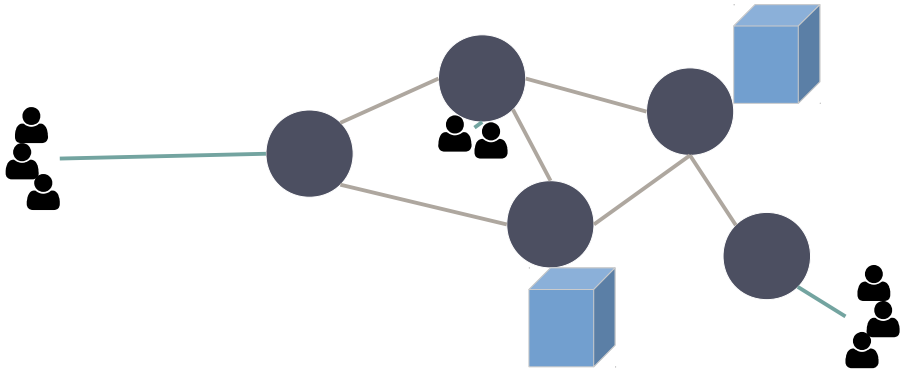


maximizing user utility

$$\max \sum_{o \in O} \sum_{q \in Q} \sum_{v \in V} n_v^{o,q} U^q$$

Utility perceived at quality q

MILP



$$\max \sum_{o \in O} \sum_{q \in Q} \sum_{v \in V} n_v^{o,q} U^q$$

subject to:

$$\sum_{q \in Q} n_v^{o,q} = n_v^o \qquad \forall o \in O, v \in V \quad (2)$$

$$d^{o,q,v_d} = n_{v_d}^{o,q} \cdot r^q \qquad \forall o \in O, q \in Q, v_d \in V \quad (3)$$

$$d^{o,q,v_d} = z_{v_d}^{o,q,v_d} + w_{v_d}^{o,q,v_d} + \sum_{e \in BS(v_d)} y_e^{o,q,v_d} - \sum_{e \in FS(v_d)} y_e^{o,q,v_d} \qquad \forall o \in O, q \in Q, v_d \in V \quad (4)$$

$$z_{v_s}^{o,q,v_d} + w_{v_s}^{o,q,v_d} + \sum_{e \in BS(v_s)} y_e^{o,q,v_d} = \sum_{e \in FS(v_s)} y_e^{o,q,v_d} \qquad \forall o \in O, q \in Q, v_s \in V, v_d \in V, v_s \neq v_d \quad (5)$$

$$\sum_{o \in O} \sum_{q \in Q} \sum_{v_d \in V} y_e^{o,q,v_d} \leq b_e \qquad \forall e \in A \quad (6)$$

$$\sum_{v_d \in V} z_{v_s}^{o,q,v_d} \leq p_{v_s}^{o,q} \cdot bw_{v_s} \qquad \forall o \in O, q \in Q, v_s \in V \quad (7)$$

CONSTRAINTS

$$\sum_{o \in O} \sum_{q \in Q} x_{v_s}^{o,q} \cdot s^q \leq S_{v_s} \qquad \forall v_s \in V \quad (9)$$

$$\sum_{o \in O} \sum_{q \in Q} \sum_{v_s \in V} x_{v_s}^{o,q} \cdot s^q \leq S_{TOT} \quad (10)$$

$$x_v^{o,q} \in \{0,1\} \qquad \forall o \in O, q \in Q, v \in V \quad (11)$$

$$n_v^{o,q} \in \mathbb{Z}^+ \qquad \forall o \in O, q \in Q, v \in V \quad (12)$$

$$y_e^{o,q,v_d} \in \mathbb{R}^+ \qquad \forall o \in O, q \in Q, v_d \in V, e \in A \quad (13)$$

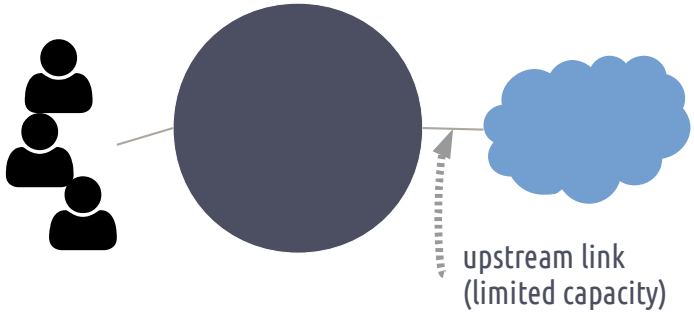
$$d^{o,q,v_d} \in \mathbb{R}^+ \qquad \forall o \in O, q \in Q, v_d \in V \quad (14)$$

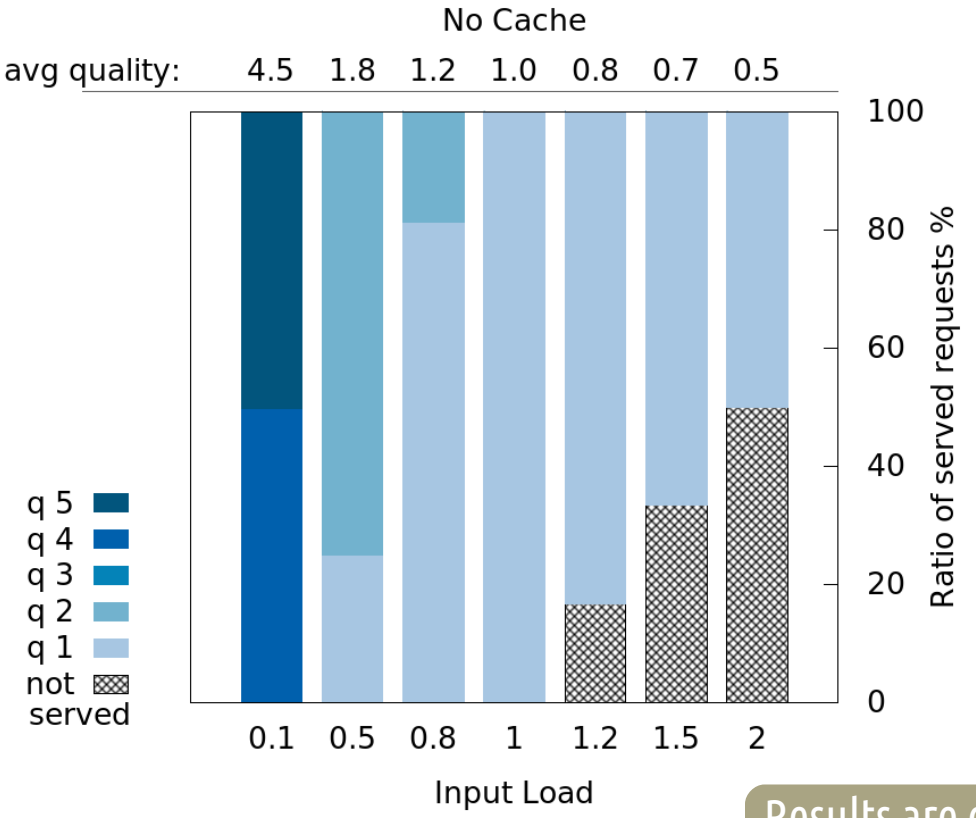
$$z_{v_s}^{o,q,v_d}, w_{v_s}^{o,q,v_d} \in \mathbb{R}^+ \qquad \forall o \in O, q \in Q, v_d \in V, v_s \in V. \quad (15)$$

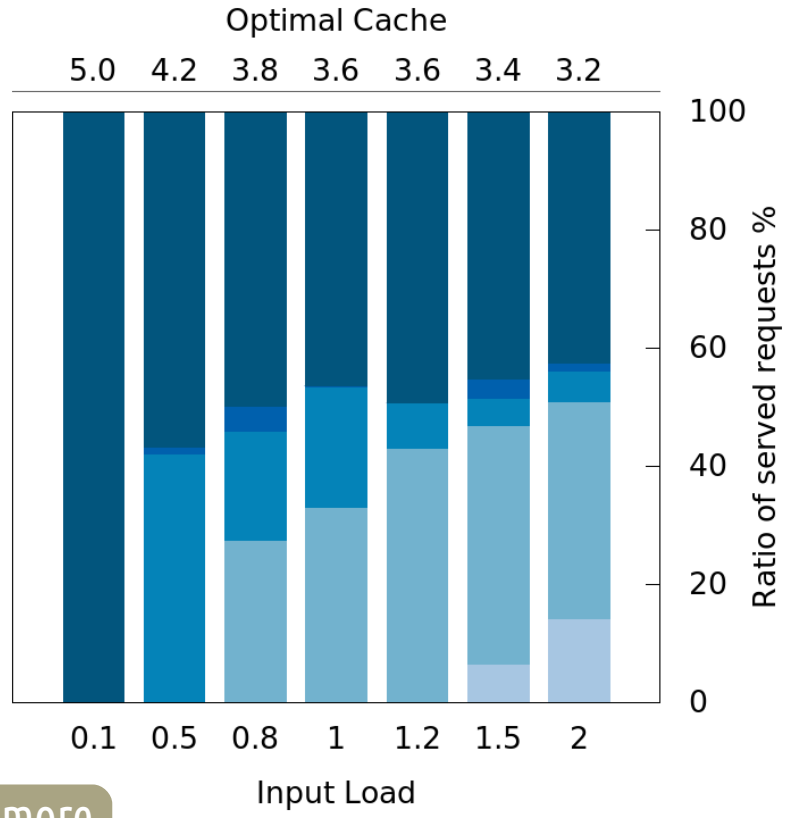
DECISIONS:

- Object Selection
- Replica Placement
- Quality Selection
- Routing

served quality



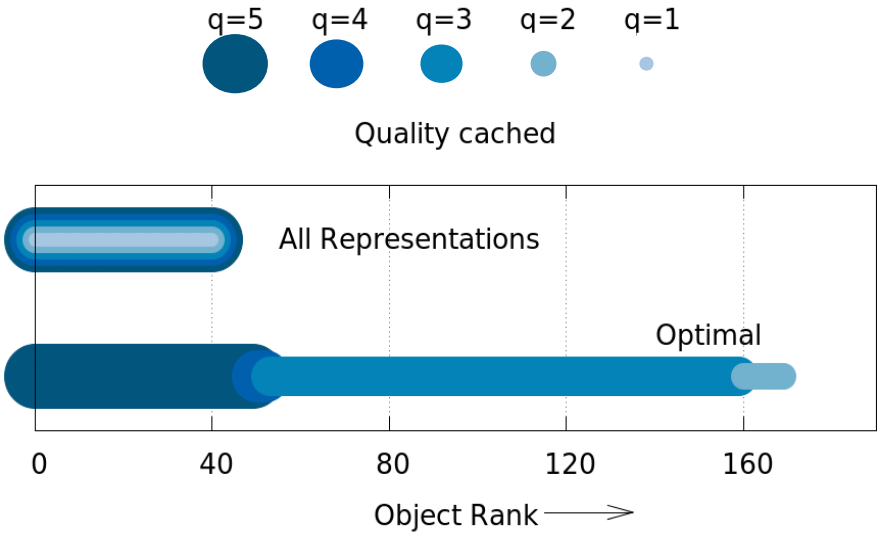




Results are confirmed in more complex topologies

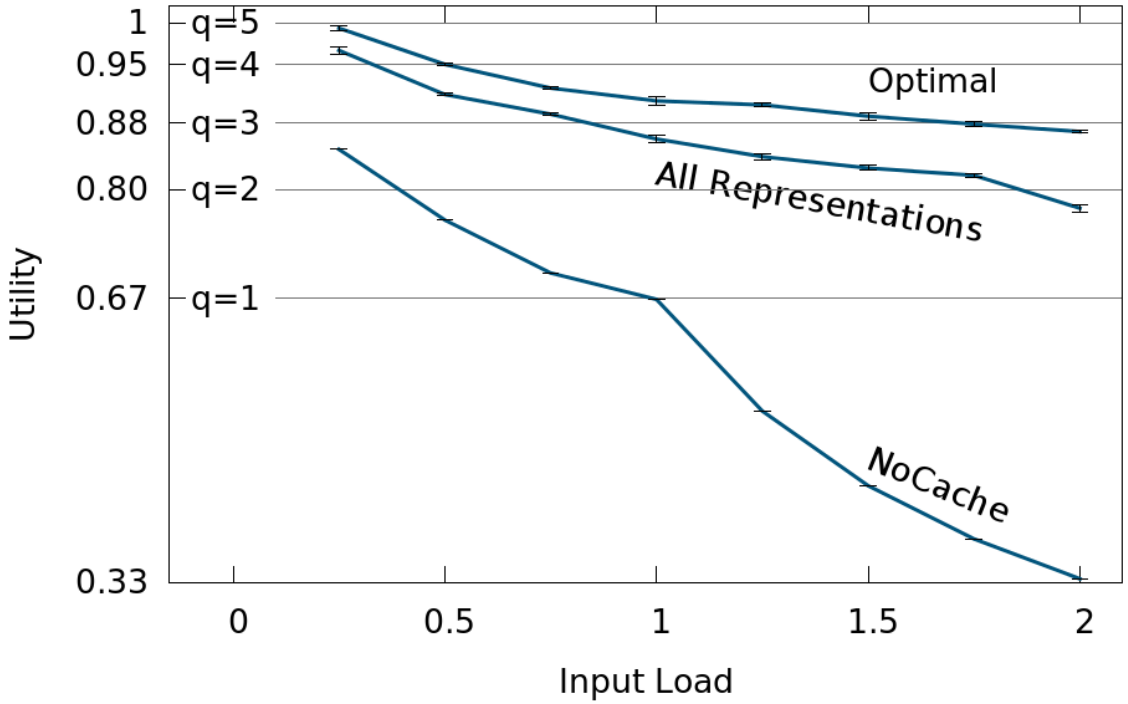
Scenario: catalog 10^4, popularity Zipf α=1, cache space sufficient to store 1/100 of the catalog at the lowest quality

optimum



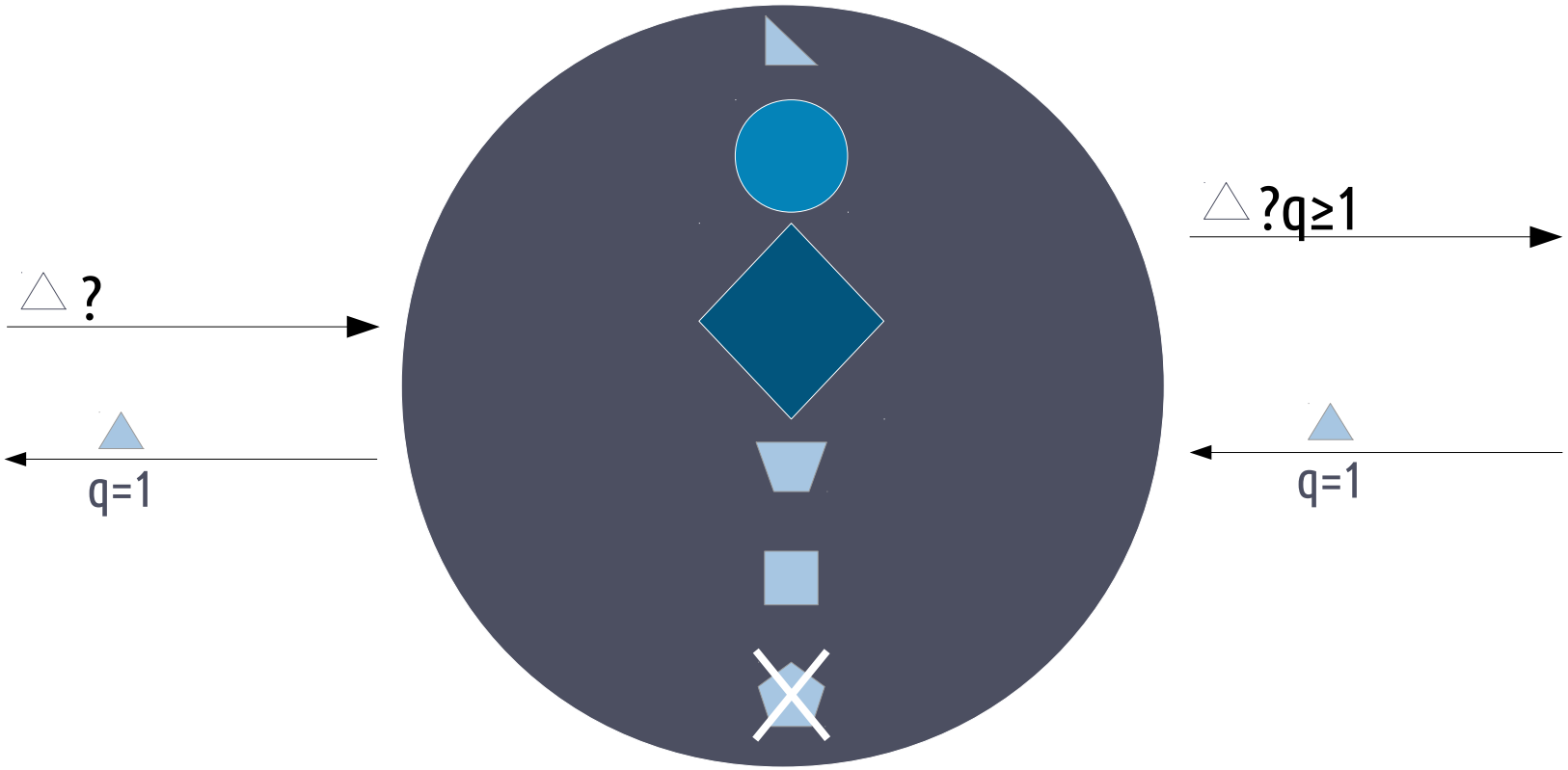
Cache at high quality the popular objects

Do not cache all the representations, but only the right ones

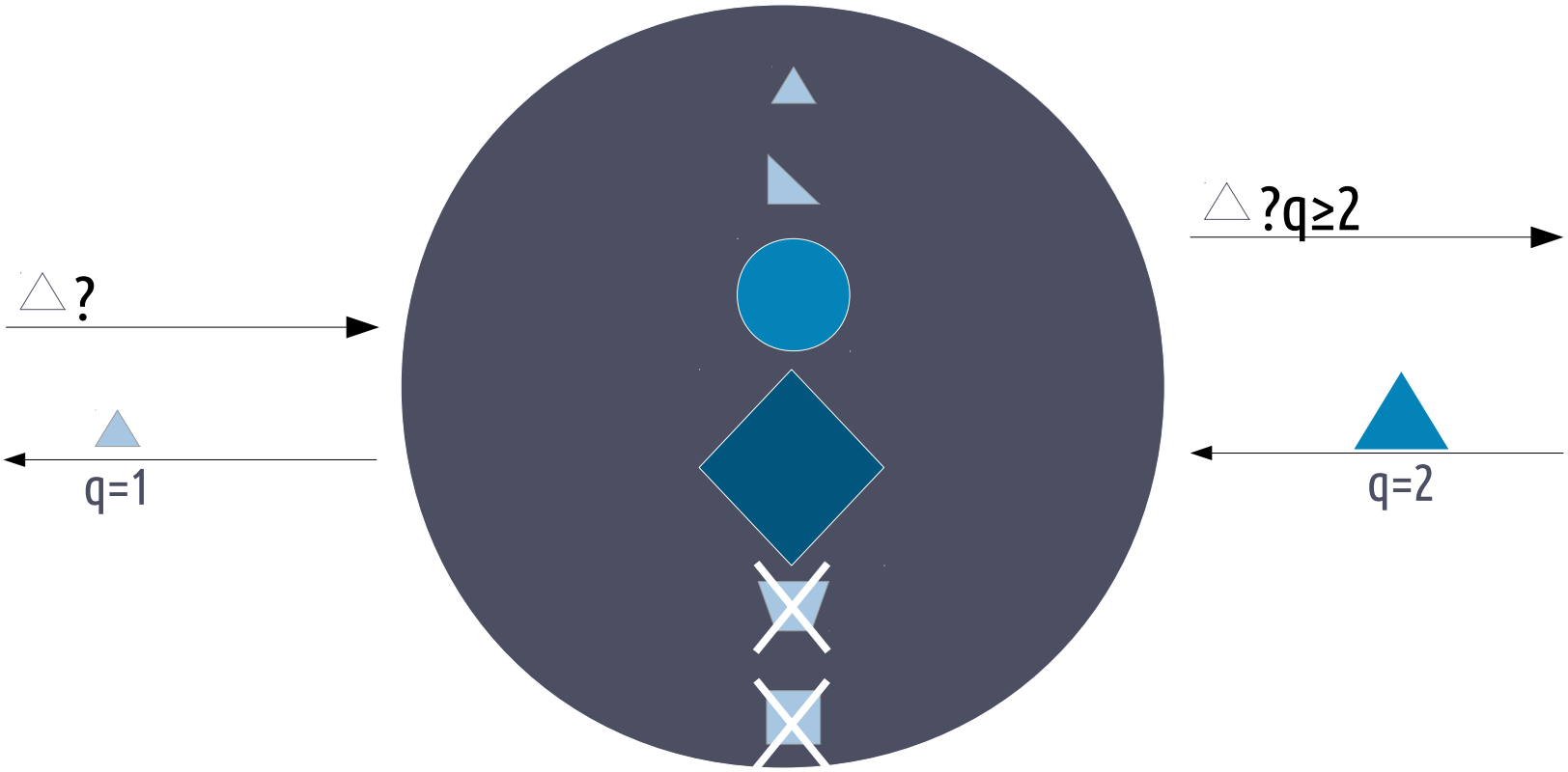


Scenario: catalog 10^4, popularity Zipf α=1, cache space sufficient to store 1/100 of the catalog at the lowest quality

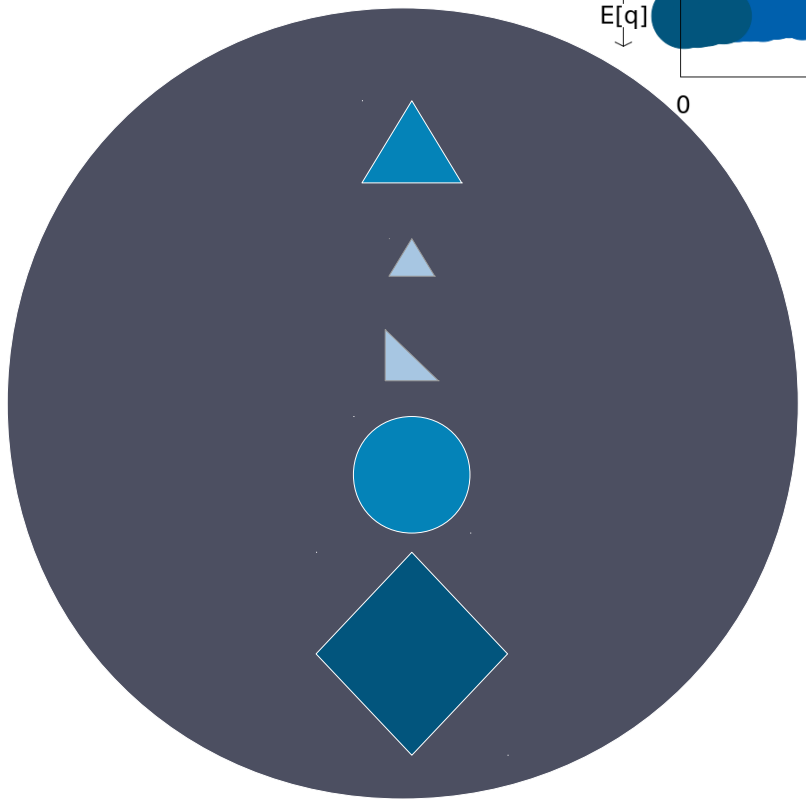
online distributed policy

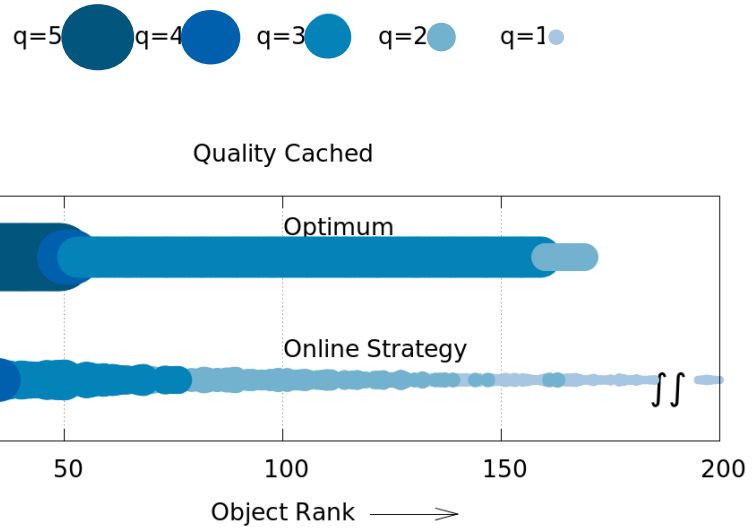


online distributed policy



online distributed policy





- Large scale simulation on realistic topologies

Scenario: catalog 10^4, popularity Zipf α=1, cache space sufficient to store 1/100 of the catalog at the lowest quality
Simulations run on ccnSim,

traffic growth

operational cost

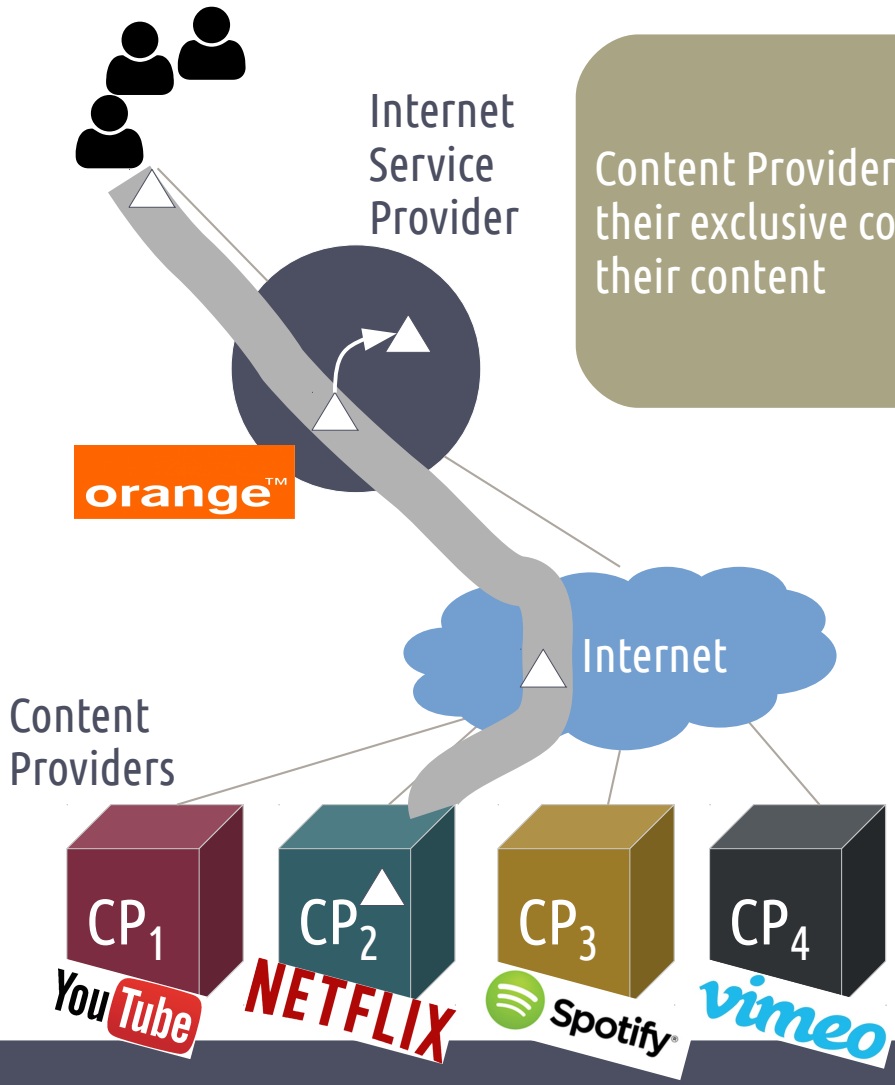
user experience

secrecy

> caching and encryption

unfeasibility of transparent caching

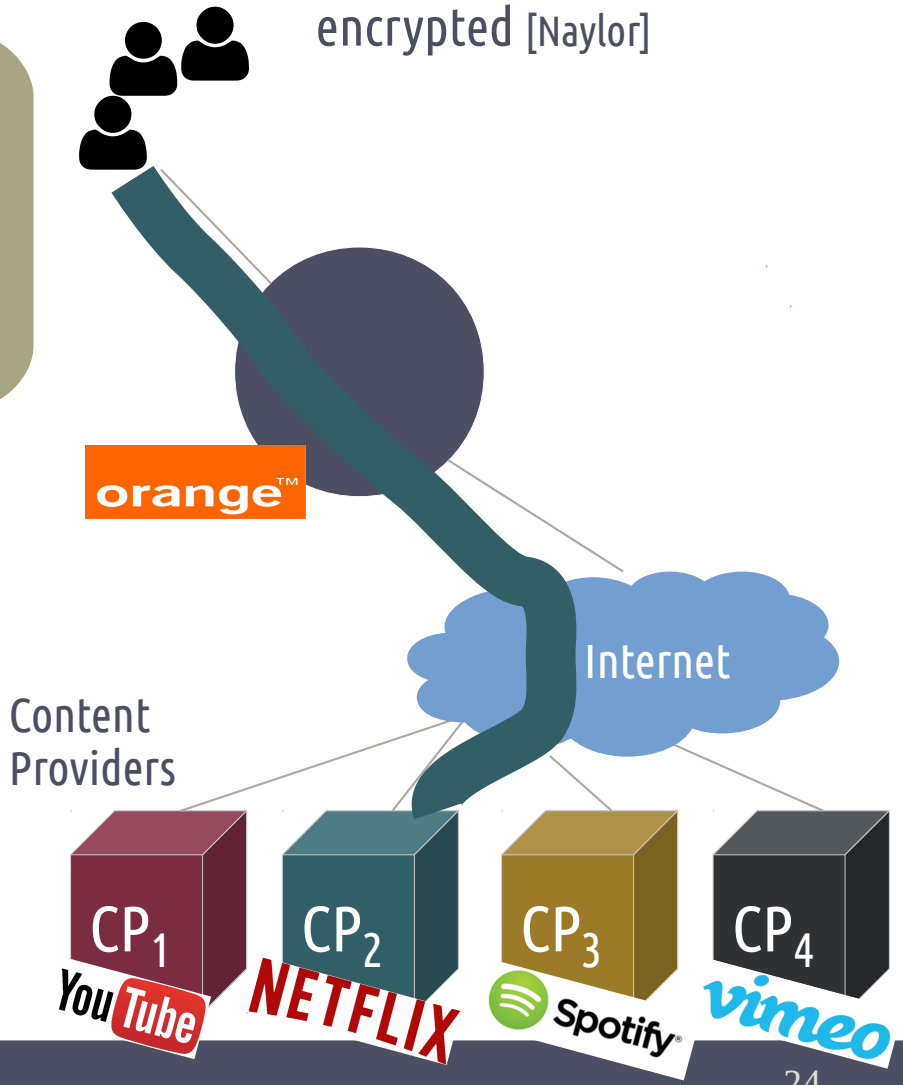
- Classic ISP transparent caching



Content Providers cannot keep their exclusive control over their content

- ISPs cache is prevented by encryption

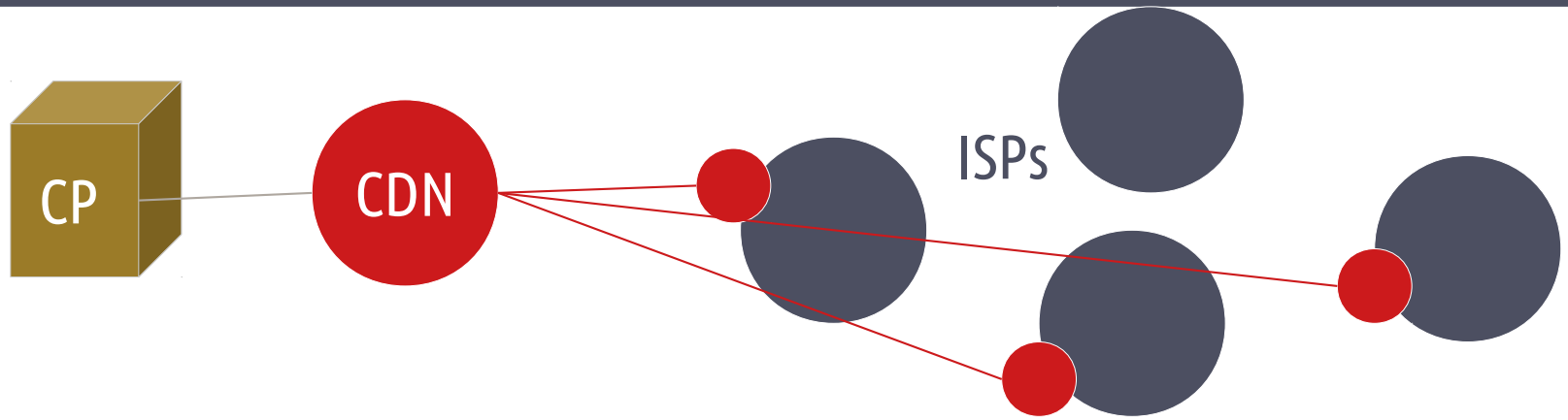
>50% of HTTP traffic is encrypted [Naylor]



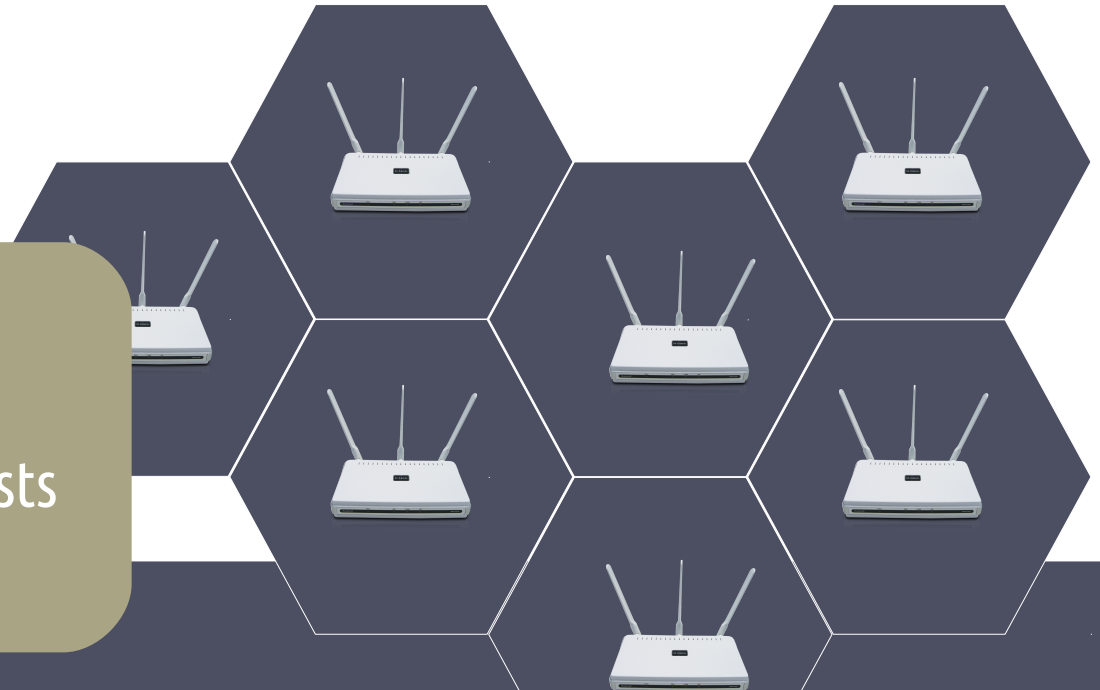
[Naylor] D. Naylor, A. Finamore et al., "The Cost of the "S" in HTTPS," in ACM CoNEXT, 2014

ISPs vs. Content Providers

limits of Content Delivery Networks (CDNs)

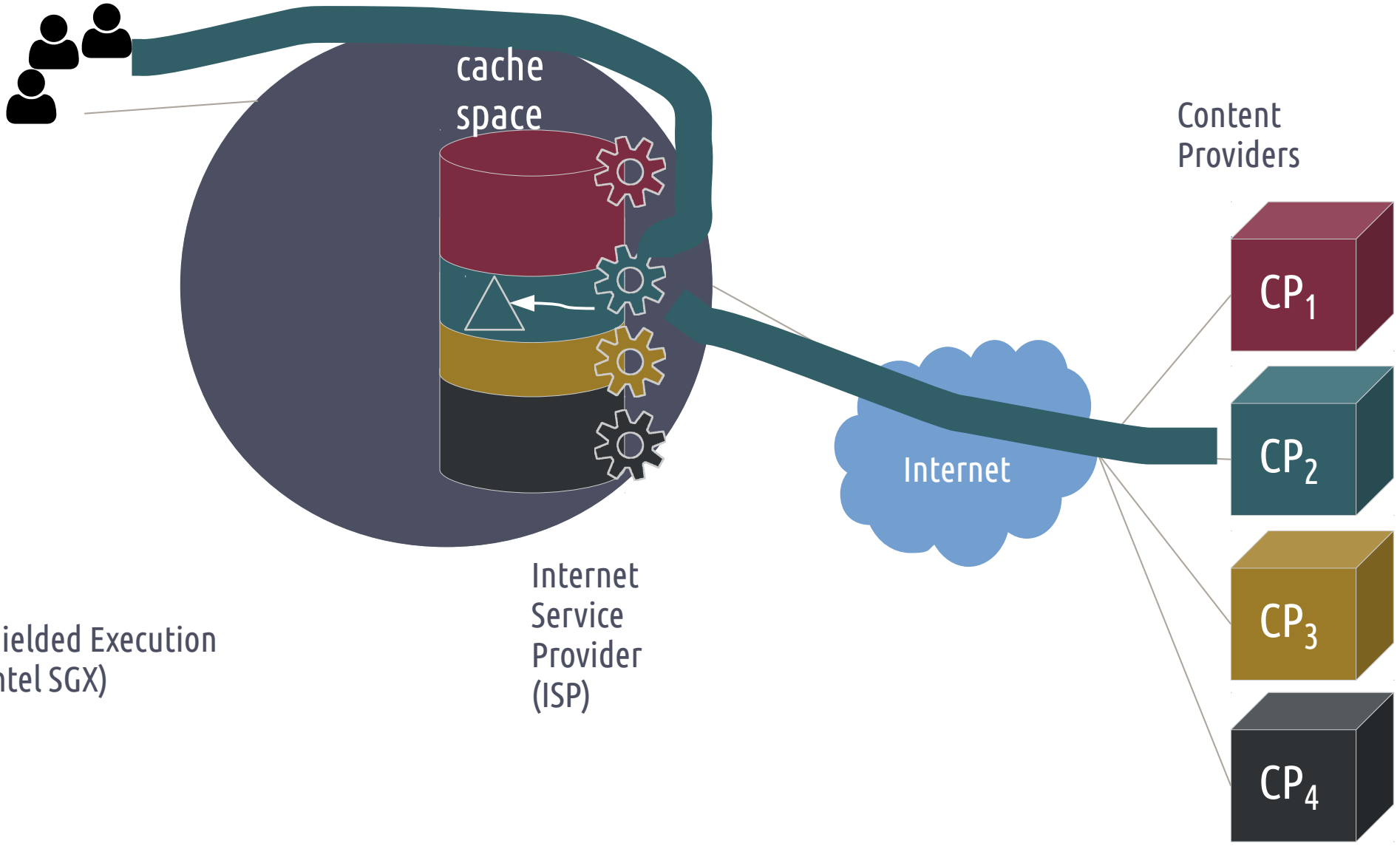


- Content Providers loose the exclusive control on their content
- Content Providers have to pay third party CDNs
- Limited permeation
  - femtocaching can only be implemented by ISPs

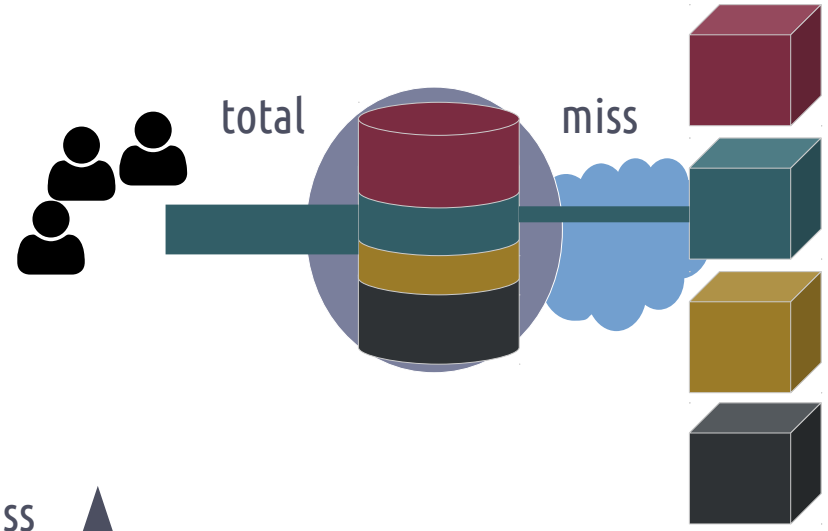


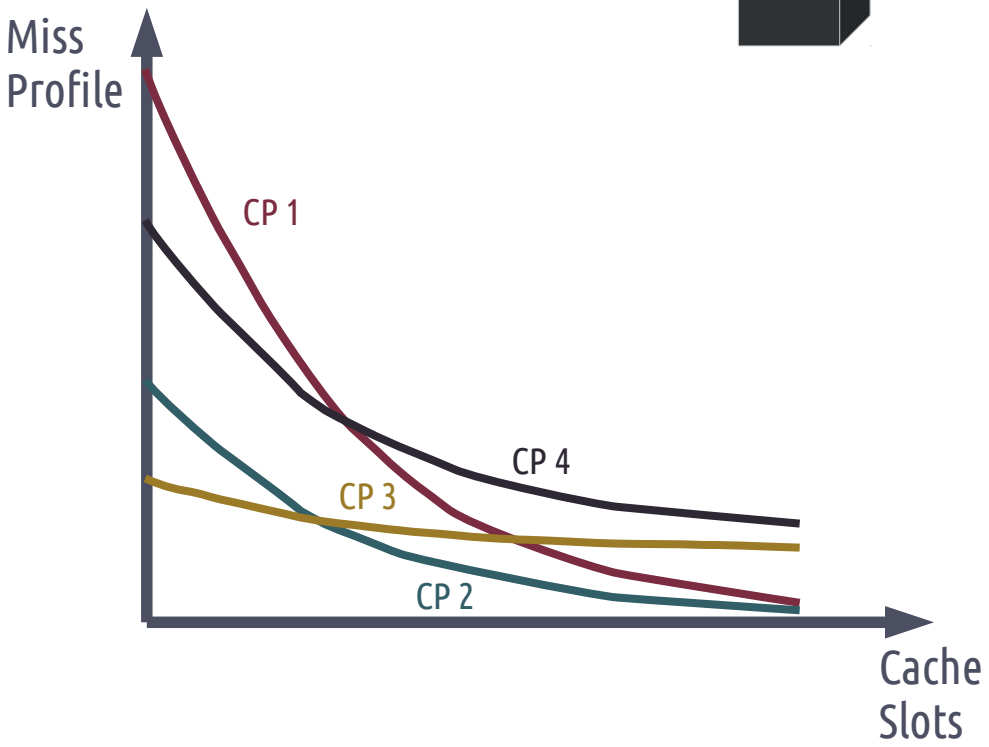
GOAL:
. Allowing ISPs to cache while …
. Preserving Content Providers interests
. Caching must be Content Oblivious

caching as a service

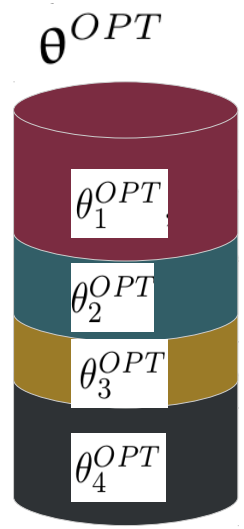


allocation





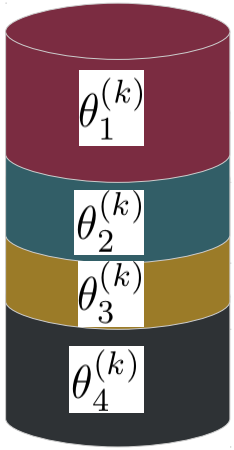
- GOAL:
  find θ^{OPT} which minimizes the overall miss stream L



Miss profiles are unknown
=> We have to infer them

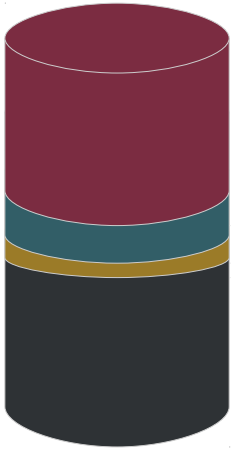
partitioning algorithm

(1) Current allocation vector $\boldsymbol{\theta}^{(k)}$

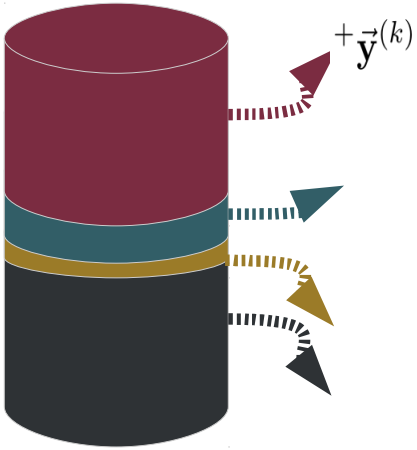


(2) Generate random perturbation $D^{(k)}$

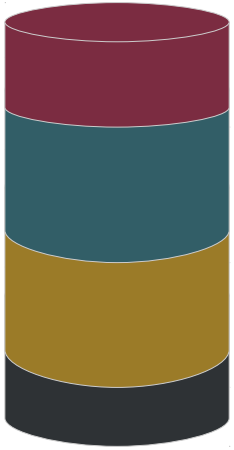
(3) Add the perturbation



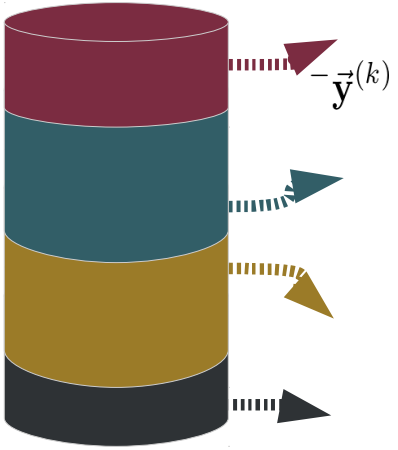
(4) Measure the miss-streams for a time *T*

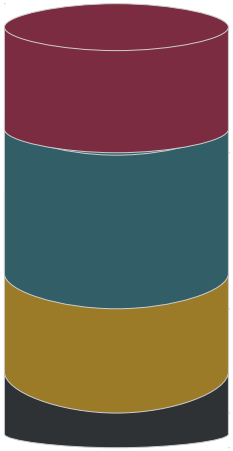


(5) Subtract the perturbation



(6) Measure the miss-streams for a time *T*

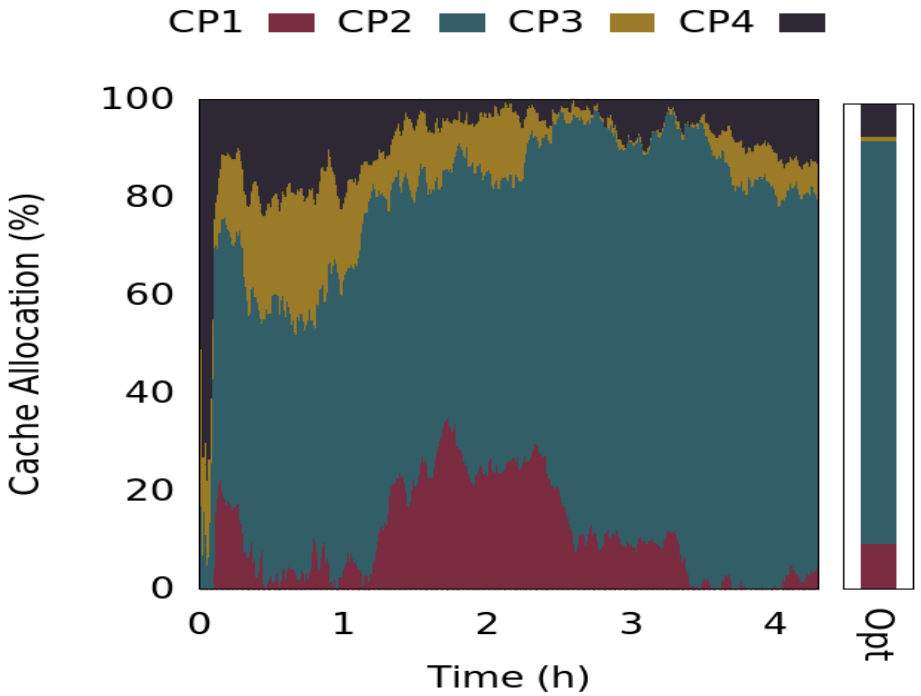




rewarding the *good* Content Providers

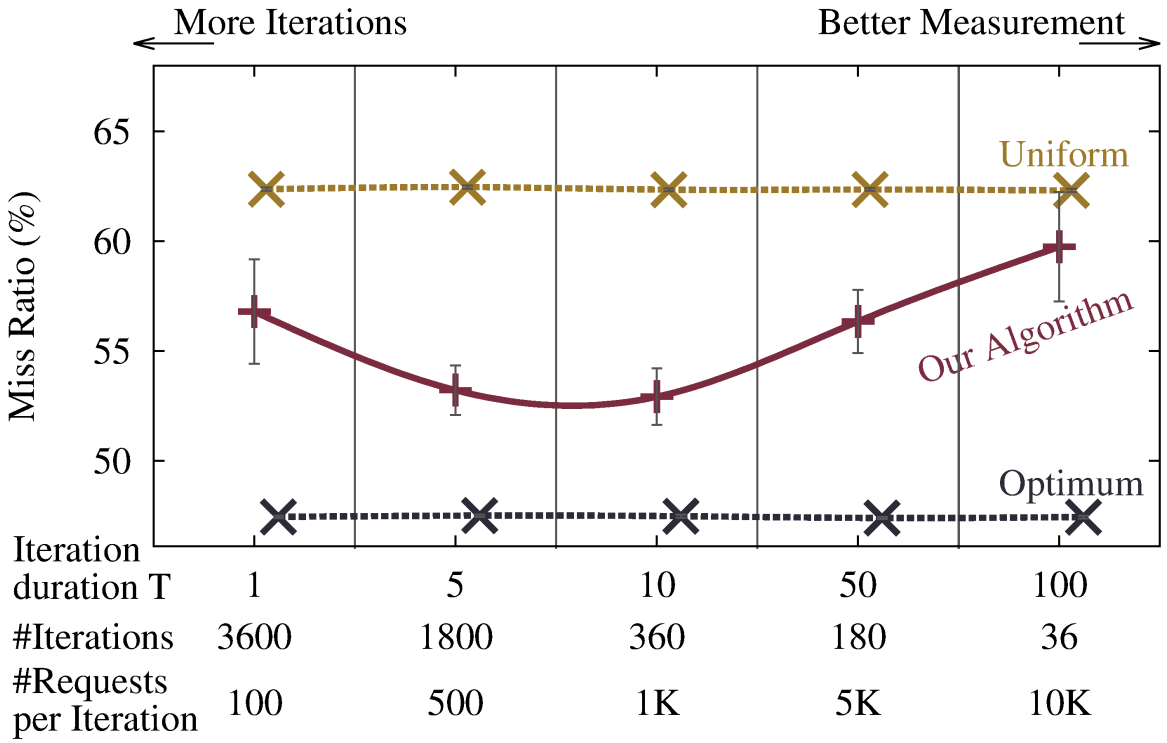
convergence

- Theorem:
  - Hp:
    - Stationary content popularity
    - Expected miss streams are decreasing and convex w.r.t. the cache slots
  - Th:
    - Despite noisy miss measurement, we convergence boundedly close to the optimum



- Overall req rate λ=100req/s, catalog size 10^8 objects, total cache 10^5, request percentage 13%,75%,2%,10%, Zipf α=0.8
- Bound 3/2 P

observation period T



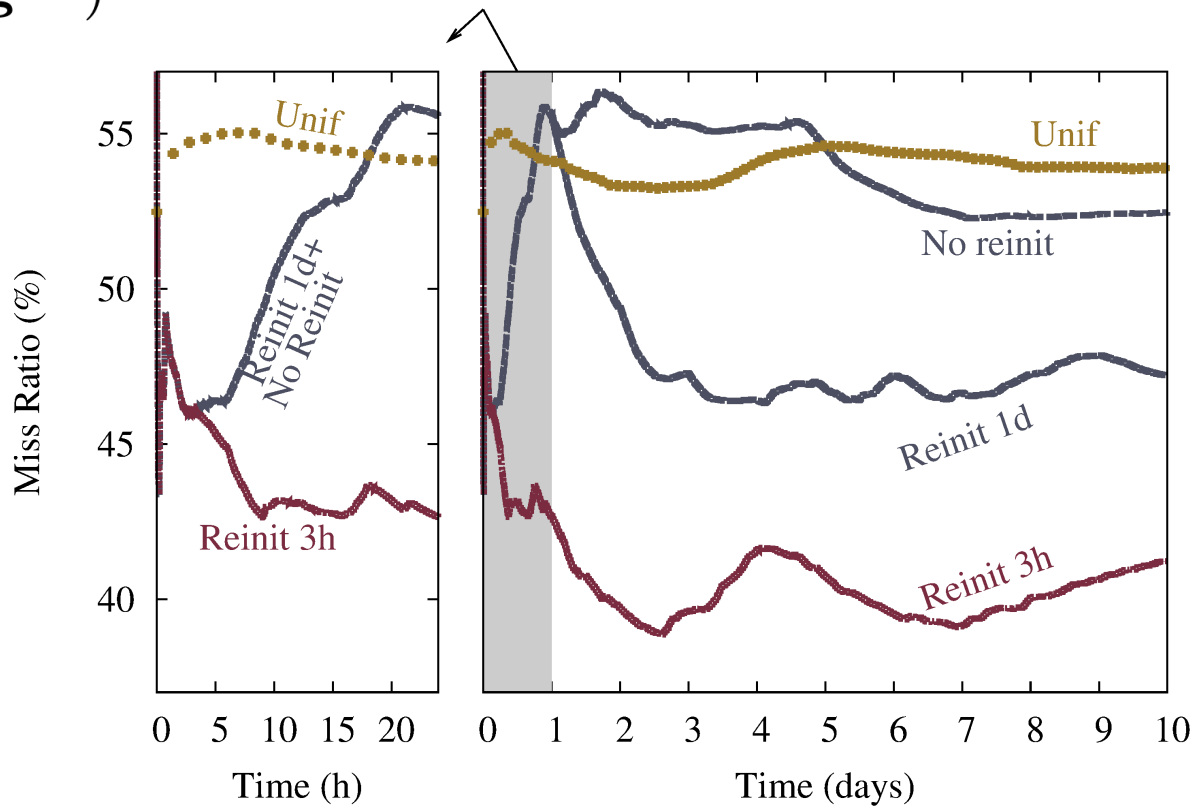
- overall req rate λ=100req/s. overall catalog 10^8, overall cache space $K=10^6$, Zipf α=0.8

time evolution of content

step size reinitialization

$$\theta^{(k+1)} = \varphi(\theta^{(k)} - a^{(k)}\hat{\mathbf{g}}^{(k)})$$

Popularity changes inspired from [Leonardi15]



In order to follow the popularity change, reinitialize step sequence every 3 h

[Leonardi15] M. Garetto, E. Leonardi and S. Traverso, "Efficient analysis of caching strategies under dynamic content popularity," IEEE INFOCOM, 2015

- overall req rate λ=100req/s. overall catalog $3.5 \cdot 10^6$, overall cache space $K=10^4$, object ON/OFF average state duration 1/9 days

conclusions

- Despite its long history, network caching still offers new research challenges
- We should look beyond the classic hit ratio maximization
- ISP Cost
  - Cache more expensive object for less OPEX
- Video
  - Maximizing user utility = Representation selection problem
- Encryption
  - Allow ISPs to cache encrypted traffic and CPs to keep their sensitive information private
- Internet is not simply a computer network!
  - Technical solutions must be refined to take into account the rule of the economic ecosystem