THE SCIENCE BEHIND CACHE HITS AND MISSES

Optimizing streaming systems with CDNs

Yuriy Reznik

Brightcove, Inc.

Content Delivery Summit

Networks, Edges, & Orchestration

May 17, 2023, Boston, MA

OUTLINE

INTRODUCTION

- Adaptive streaming (before CDNs)
- CDN-assisted streaming models (HLS, DASH)
- The disconnect

FORMULAS

- Popularity models
- Cache misses in ideal cache
- Cache misses in a system with multiple formats
- Cache misses in a system with multiple renditions

APPLICATIONS

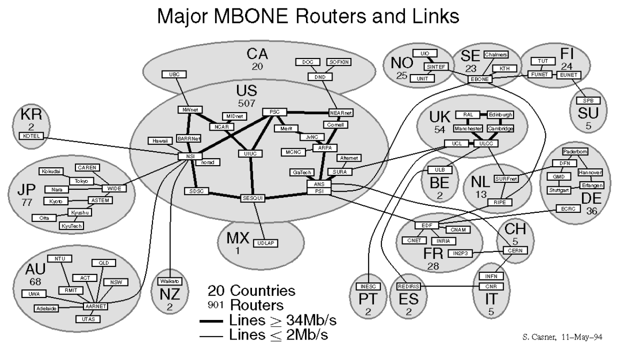
- Multi-format systems
- ABR ladders designs
- Multi-CDN delivery systems

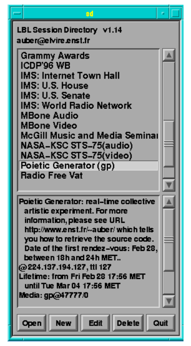
BRIGHTCOVE

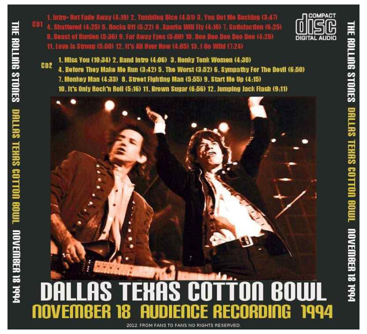
EARLY DAYS OF STREAMING

1993: MBONE

- Virtual multicast network connecting several universities & ISPs
- RTP-based video conferencing tool (vic) is used to send videos
- 1994 Rolling Stones concert – first major event streamed online







1995: RealAudio, 1997: RealVideo

- First commercially successful mass-scale streaming system
- Proprietary protocols, codecs: PNA, RealAudio, RealVideo
- Worked over UDP, TCP, and HTTP ("cloaking" mode)
- First major broadcast: 1995 Seattle Mariners vs New York Yankees

1995+: VDOnet, Vivo, NetShow, VXtream, …

- Many vendors have tried to compete in streaming space initially
- Vivo & Xing got acquired by Real, VXtreme by Microsoft
- By 1998, 3 main vendors remained: Real, Microsoft and Apple

1998: RealSystem G2

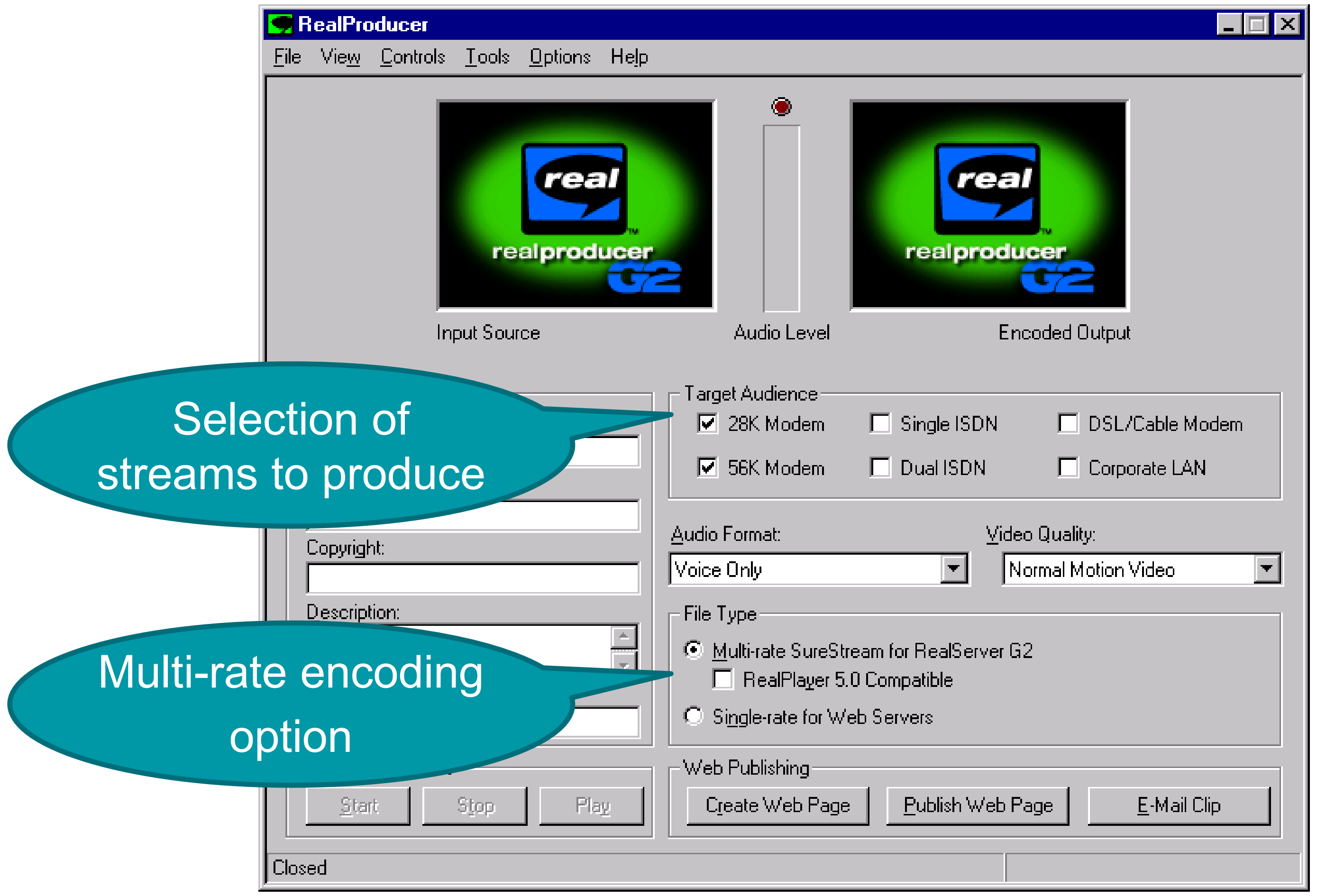
- First ABR streaming system

FIRST ABR SYSTEM

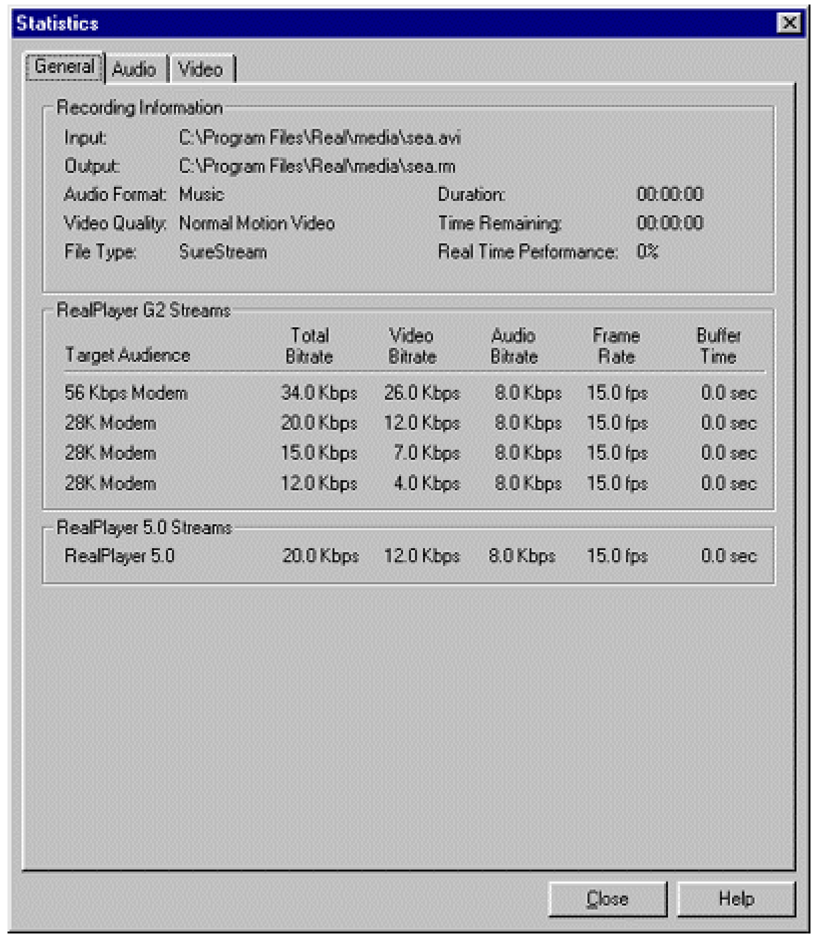
1998: RealSystem G2: "SureStream"

First commercially successful ABR streaming system

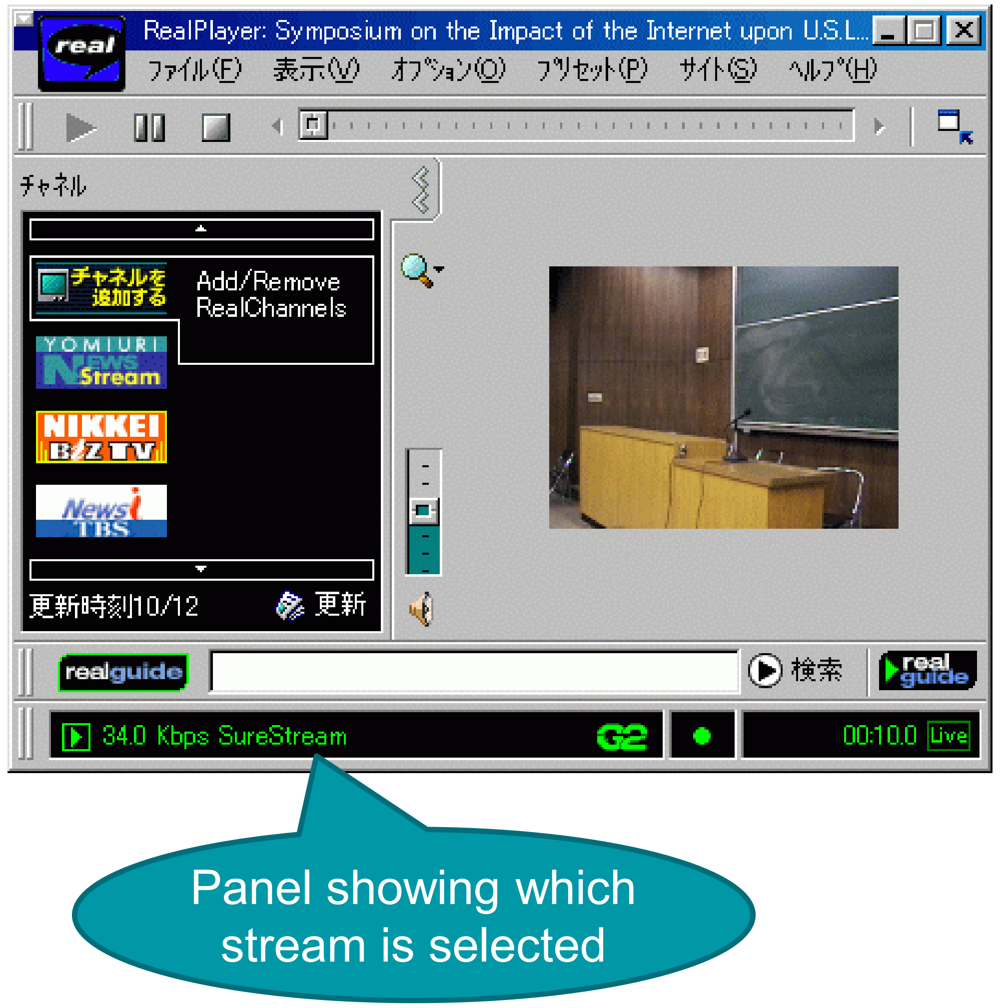
Encoder:



Encoded streams



Player



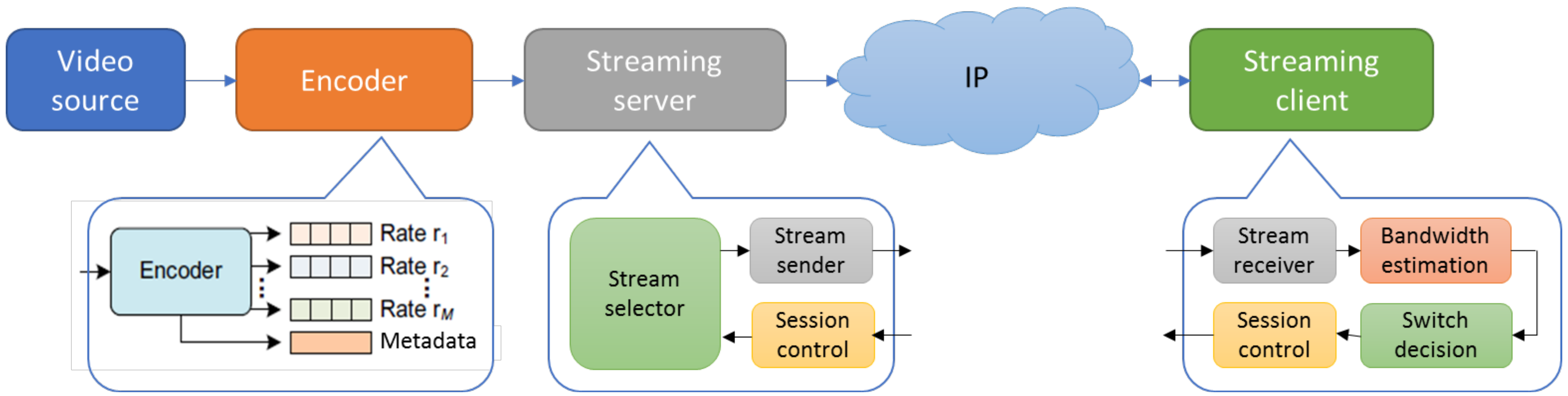
Some references

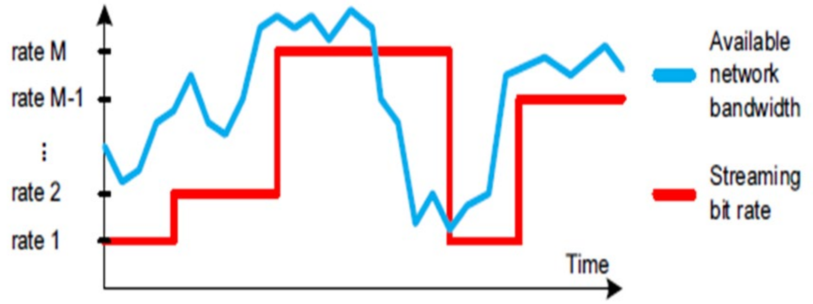
- B. Girod, et al, "Scalable codec architectures for Internet video-on-demand," ACSSC, pp. 357 – 361, 1997.
- Y. Reznik, et al, "Video Coding for Streaming Media Delivery on the Internet," *TCSVT*, 11 (3), pp. 20-34, 2001.
- US Patents: 6314466, 6480541, 7075986, 7885340, …

BRIGHTCOVE

ORIGINAL ABR MODEL

RTSP/UDP-based streaming architecture:





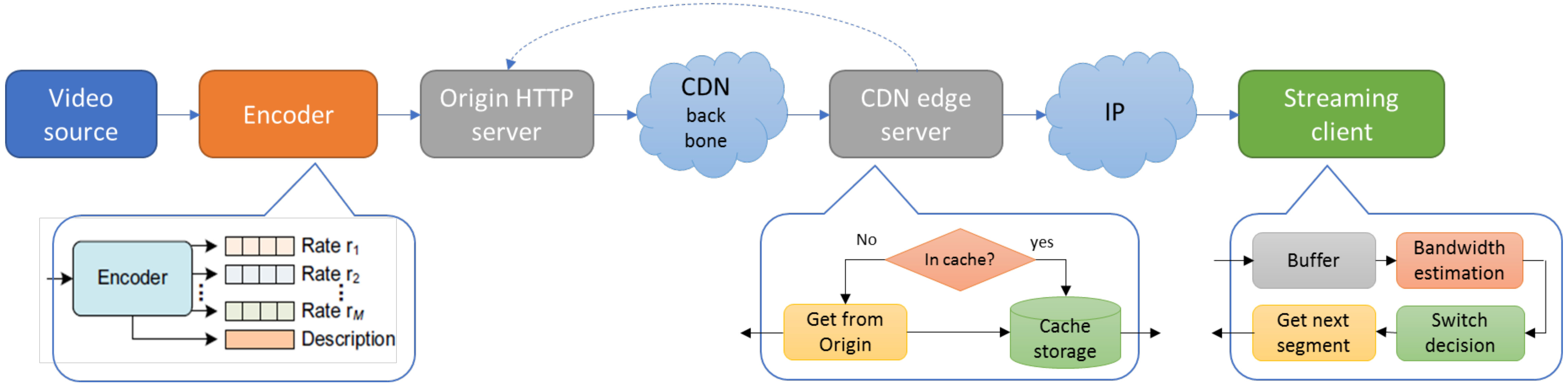
- Public internet is used for delivery
- RTSP protocol was used for session control, and UDP (plus RTP or proprietary transport) were used for sending the data
- Stream adaptation was done by server, but with most clients – it was client-driven: client was sending requests to switch
- Server was also responsible for retransmissions, injecting extra FEC packets, etc.
- Everything was sent in “packets”

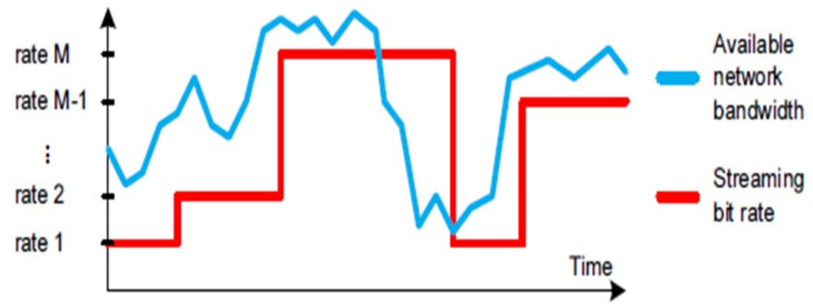
Important design elements:

- Only one stream was sent of over IP from the server to each client
- Multiple renditions were stored only on the original server, no transmissions of “stacks of streams” to other servers
- This was all before CDNs and relay networks for streaming!

HTTP-BASED ABR MODEL

Modern-era HLS/DASH architecture:





Key differences from RTSP/UDP streaming:

- Instead of streaming server, a regular HTTP server is used as origin
- Stream switching is trivialized to HTTP GET operations originating from streaming client
- The scaling and delivery is delegated to CDN, which caches content on the edge servers, reducing the load on the origin…

Important new factors:

- This works well when the “content” is popular and it becomes stored in the edge cache
- If content is not popular, and not stored at the edge cache – it becomes pulled from the origin server
- In other words, CDN helps to improve delivery, but only when some content is popular.

THE DISCONNECT

Observations

- ABR Streaming need several encoded versions of the content:
  - Multiple streams are needed to achieve better network adaptation and minimize the visibility of stream switches.
  - Multiple streams are also needed to support different delivery formats (HLS, DASH, MSS, etc.) and DRM systems.
  - Support for multiple video codecs (H.264, HEVC, AV1, and VVC) also results in a creation of multiple streams

- However, with CDNs, such streams start “competing” for the CDN edge cache disk space
  - This results in more CDN cache misses, and higher load on origin server.
  - This also increases delivery costs and makes whole system less reliable, less scalable, etc.

The disconnect

- ABR systems need “more” streams to deliver various functionalities, while
- CDNs need “fewer” streams to be most effective

Objectives of this talk

- Offer few models quantifying the impact of multiple streams/representations on CDN performance
- Offer recommendations for the design of streaming systems to make them more efficient from the CDN performance point of view

THE MODELS

CONTENT POPULARITY MODELS

Let us assume that

- We have a set of items (e.g. videos or segments)

$$S = \{s_x, x \geq 1\}$$

- The requests to their retrieval can be modeled by an iid source with pmf p(x)

Zeta-distribution model

- Let us also assume that probabilities of retrieval of these items follow Zeta distribution:

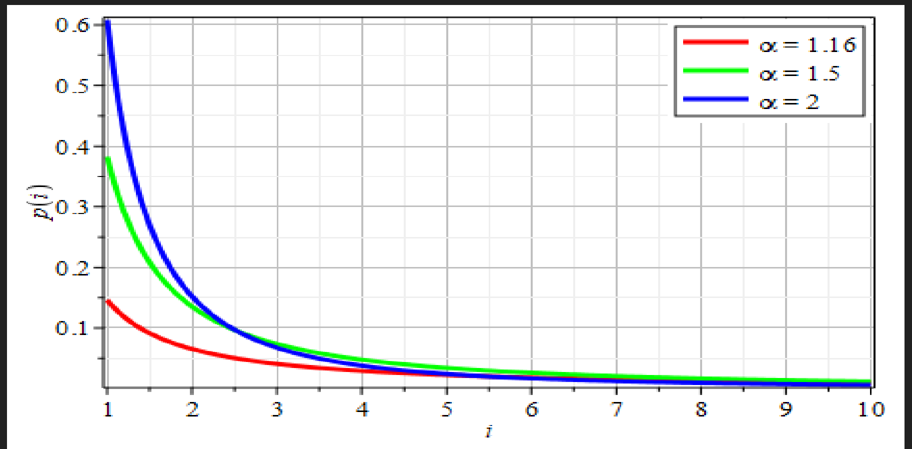
$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha)}$$

where where α is a shape parameter, and $\zeta(.)$ is a Riemann Zeta function.

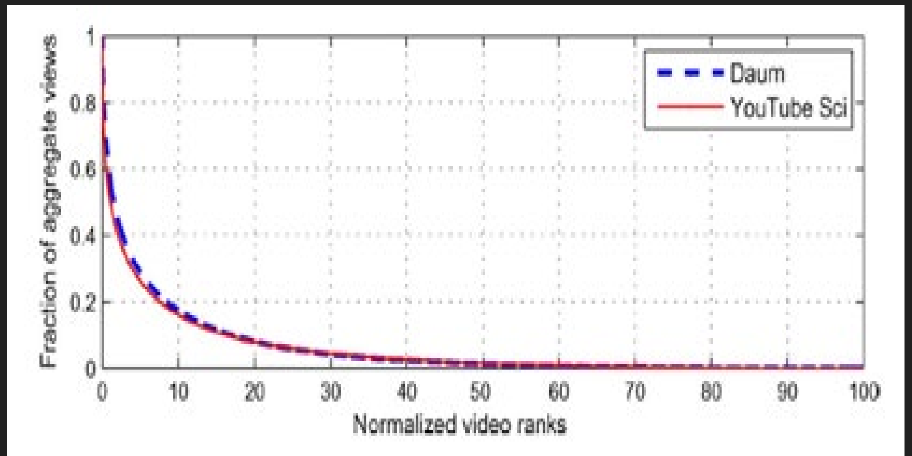
- This is a classic discrete distribution model, with many known examples of its use in similar contexts. E.g. it is known to provide a good approximation of popularity of videos in YouTube.

- In bounded case, when $x \leq N$, it turns into a well-known Zipf's distribution.

Zeta distribution



Ranks of videos in YouTube



M. Cha, et al, "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems," IEEE/ACM Trans. Networks, vol. 17, 2009, pp 1357-1370.

N. Kamiyama and M. Murata, "Reproducing Popularity Distribution of YouTube Videos," in IEEE Transactions on Network and Service Management, vol. 16, no. 3, 2019, pp. 1100-1112.

BRIGHTCOVE

IDEAL CACHE MODEL

Let us assume that

- We have a cache with capacity of C items
- And this cache is "ideal":
  - it knows exact probabilities of all items
  - it only stores C items with highest probabilities of occurrence

Then

- By considering that input items follow Zeta distribution:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha)}$$

- The probability that any randomly selected item x falls outside of cache of size C will be

$$p_{miss}(C,\alpha) = \mathbf{1} - \sum_{x=1}^{C} p(x) = \mathbf{1} - \frac{H_{C,\alpha}}{\zeta(\alpha)}$$

where $H_{C,\alpha} = \sum_{x=1}^{C} x^{-\alpha}$ is a generalized Harmonic number

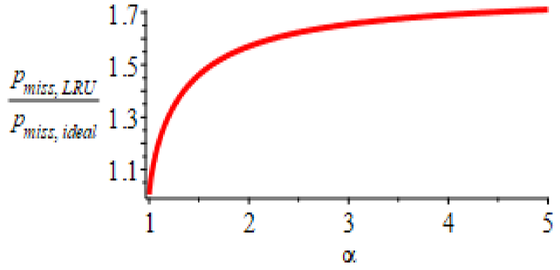
- When cache size C is large, this asymptotically turns into:

$$\boxed{p_{miss}(C,\alpha) \sim \frac{C^{1-\alpha}}{(\alpha-1)\zeta(\alpha)}\left(1 + O\left(\frac{1}{C}\right)\right)}$$

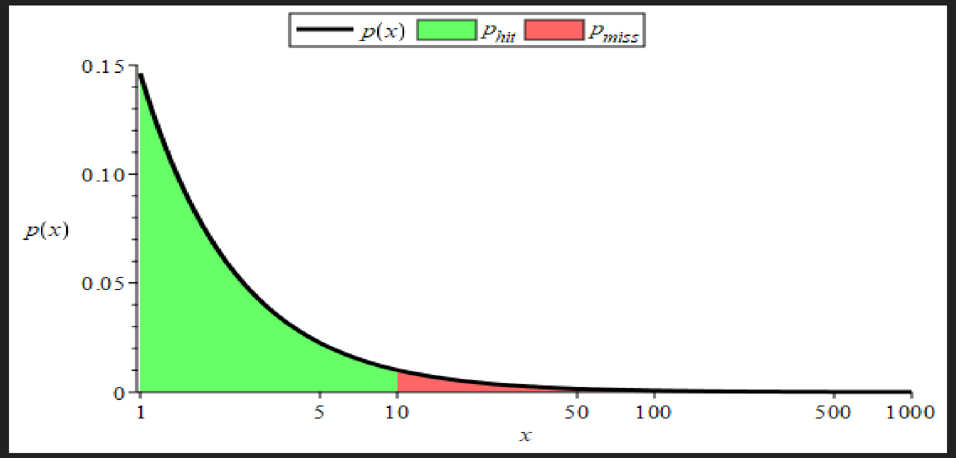
- For LRU caches it is also known that:

$$\frac{p_{miss,LRU}(C,\alpha)}{p_{miss}(C,\alpha)} \sim \left(1-\frac{1}{\alpha}\right)\left[\Gamma\left(1-\frac{1}{\alpha}\right)\right]^{\alpha}$$

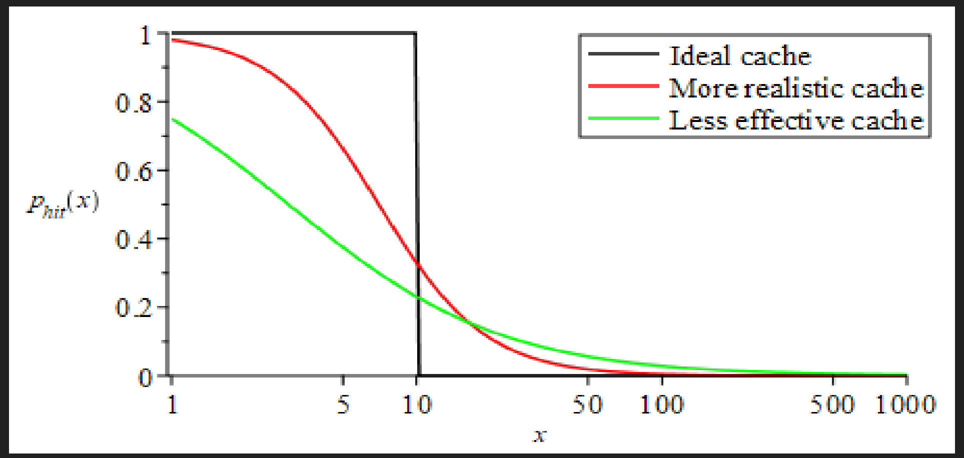
For small α this ratio is close to 1! LRU is not that far from the ideal!



Ideal cache hit/miss partition



Ideal vs realistic behavior



References

P. Franaszek and T. Wagner, "Some distribution-free aspects of paging algorithm performance," JACM, 21(1), 1974, pp.31-39.

P. R. Jelenkovic, "Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities," Ann. Appl. Probab., 9 (2), 1999, pp. 430–464.

BRIGHTCOVE

EFFECT OF 2 FORMATS

Let us next assume that

- We have 2 sets of content items (e.g. same video in HLS and DASH formats):

$$S_1 = \{s_{1,x}, x \geq 1\}, \text{ and } S_2 = \{s_{2.x}, x \geq 1\}$$

- Relative usage probabilities of these 2 sets:

$$\pi = \{\pi_1, \pi_2\}$$

- Full probabilities ($p(x)$ - content popularity distribution):

$$\boxed{p(s_{1,x}) = \pi_1 \cdot p(x) \quad \text{and} \quad p(s_{2,x}) = \pi_2 \cdot p(x)}$$

Structure of probability-ordered items in the cache ($\pi_1 > \pi_2$):

Item	Probability	Comments
$s_{1,1}$	$\pi_1 p(1)$	The 1st item in more widely used format
...	...	...
$s_{1,x}$	$\pi_1 p(x)$	$x = \left\lfloor \left(\frac{\pi_1}{\pi_2}\right)^{\frac{1}{\alpha}} \right\rfloor$, solution of $\pi_1 p(x) = \pi_2 p(1)$
$s_{2,1}$	$\pi_2 p(1)$	The 1st item in less widely used format
$s_{1,x+1}$	$\pi_1 p(x+1)$	Subsequent items in more widely used format
...	...	...
s_{1,x_2}	$\pi_1 p(x_2)$	$x_2 = \left\lfloor 2\left(\frac{\pi_1}{\pi_2}\right)^{\frac{1}{\alpha}} \right\rfloor$, solution of $\pi_1 p(x_2) = \pi_2 p(2)$
$s_{2,2}$	$\pi_2 p(2)$	The 2nd item in less widely used format
s_{1,x_2+1}	$\pi_1 p(x_2+1)$	Subsequent items in more widely used format
...	...	...

- NB: Items from less widely used set become injected with a step size of $\boldsymbol{x \sim (\pi_1/\pi_2)^{1/\alpha}}$!

BRIGHTCOVE

CACHE MISSES WITH 2 FORMATS

Some asymptotic formulae

- With large cache size C and two versions of items with π_1, π_2 usage probabilities:

$$p_{miss,2}(C,\alpha,\pi) \sim \left(\pi_1^{\frac{1}{\alpha}} + \pi_2^{\frac{1}{\alpha}}\right)^{\alpha} \frac{C^{1-\alpha}}{(\alpha-1)\zeta(\alpha)}\left(1 + O\left(\frac{1}{C}\right)\right)$$

- This looks similar to cache miss probability in case of singe set/representation:

$$p_{miss}(C,\alpha) \sim \frac{C^{1-\alpha}}{(\alpha-1)\zeta(\alpha)}\left(1 + O\left(\frac{1}{C}\right)\right)$$

Relative increase of cache miss probability

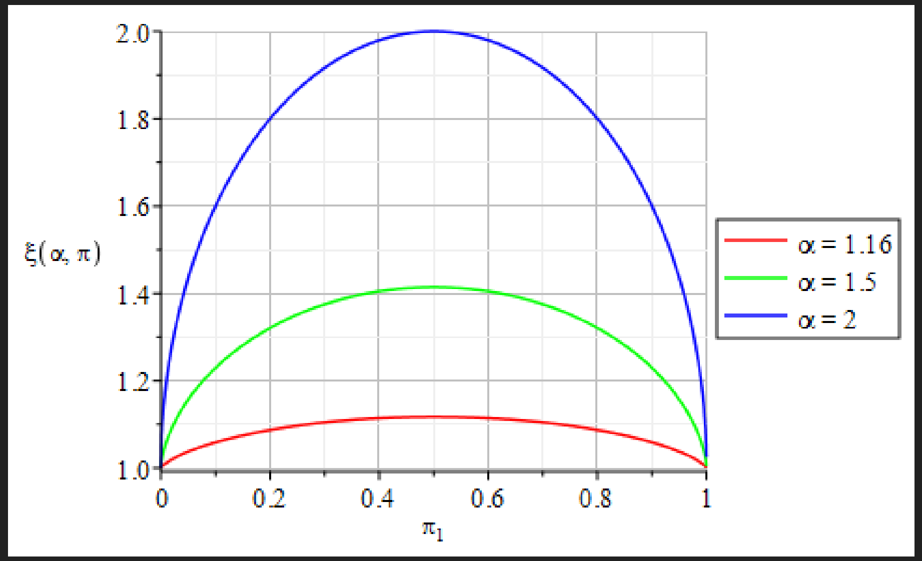
- If we next look at the ratio:

$$\boxed{\xi(\alpha,\pi) = \frac{p_{miss,2}(C,\alpha,\pi)}{p_{miss}(C,\alpha)} \sim \left(\pi_1^{\frac{1}{\alpha}} + \pi_2^{\frac{1}{\alpha}}\right)^{\alpha}}$$

  we discover that it becomes asymptotically independent on C!

- In other words, considering any CDN with reasonably large cache, we can predict that the use of 2 versions (formats) will increase its cache miss probability by $\left(\pi_1^{1/\alpha} + \pi_2^{1/\alpha}\right)^{\alpha}$

Relative increase in cache miss probability in case of using 2 formats.



BRIGHTCOVE

CACHE MISSES WITH K FORMATS

Asymptotic result for k formats

- More generally, it can be shown, that asymptotically (with large CDN cache size) the use of k versions will increase its cache miss probability by a factor of

$$\xi(\alpha,\pi) = \frac{p_{miss,k}(C,\alpha,\pi)}{p_{miss}(C,\alpha)} \sim \left(\sum_{i=1}^{k} \pi_i^{\frac{1}{\alpha}}\right)^{\alpha} = \|\pi\|_{\frac{1}{\alpha}}$$

  Where α is a parameter of content popularity model, and $\pi = \{\pi_1, \dots, \pi_k\}$ are the usage probabilities of each format

Observations

- The worst impact happens when all formats are equally probable:

$$\pi_1 = \cdots = \pi_k$$

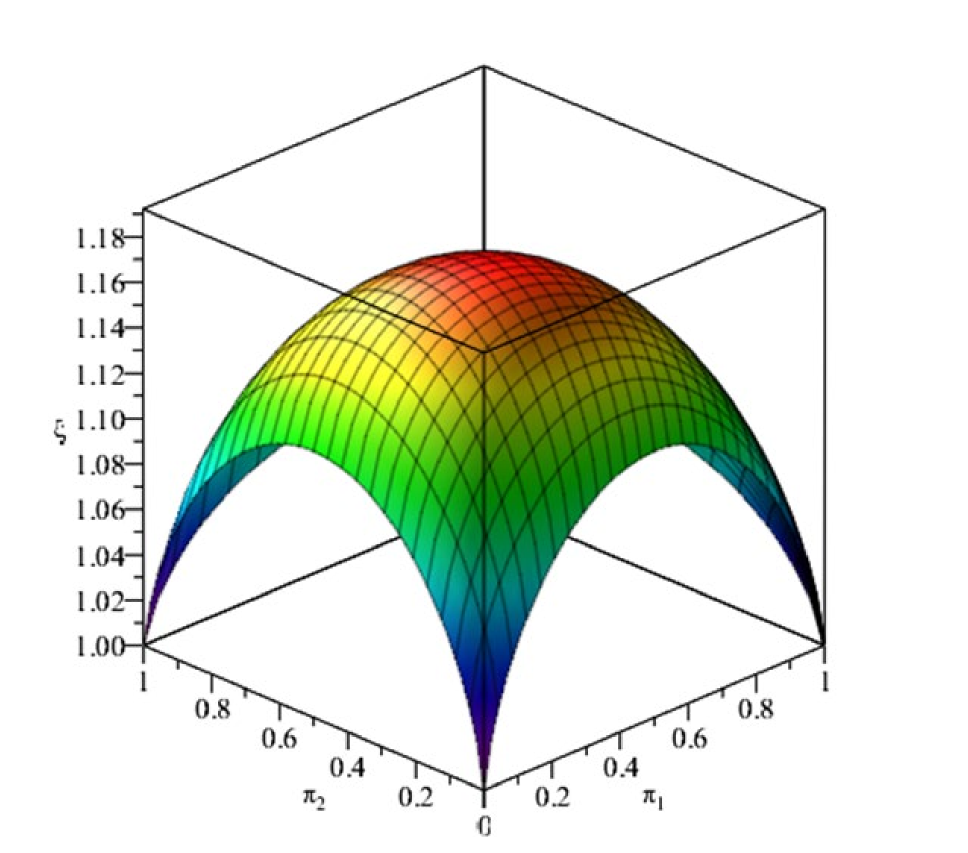
- The higher is the asymmetry in usage of different formats (or renditions), the better it is from CDN efficiency standpoint:

$$\pi_i \to 1 \Rightarrow \xi(\alpha,\pi) \to 1$$

Recipe for success

- To improve CDN performance with multiple representations/formats - pick one “preferred” representation, and direct as many possible clients/devices use it!

Relative increase in cache miss probability in case of using 3 formats.



Reference:

Y. Reznik, T. Teixeira, and R. Peck, "On Multiple Media Representations and CDN Performance," *Proc. ACM Mile-High Video Conference (MHV'22)*, March 2022 Pages 56-61.

BRIGHTCOVE

CACHE MISSES WITH ABR LADDERS

Given

- k – the number of streams
- $R = \{R_1, \ldots, R_k\}$ – bitrates of streams in encoding ladder
- $\pi = \{\pi_1, \ldots, \pi_k\}$ – load probabilities of each stream
- S – cache capacity (bits)
- α – parameter of content popularity model

It can be show that asymptotically with $S \to \infty$:

$$p_{miss,k}(S, R, \pi, \alpha) = \left(\pi_1^{\frac{1}{\alpha}} + \cdots + \pi_k^{\frac{1}{\alpha}}\right)^{\alpha} \frac{(S/R^*)^{1-\alpha}}{(1-\alpha)\zeta(\alpha)} \left(1 + O\left(\frac{1}{S}\right)\right)$$

where

$$R^* = \frac{1}{\pi_1^{\frac{1}{\alpha}} + \cdots + \pi_k^{\frac{1}{\alpha}}} \left(\pi_1^{\frac{1}{\alpha}} R_1 + \cdots + \pi_k^{\frac{1}{\alpha}} R_k\right)$$

is the average rate of the content as it stays in the cache

Recipe for success

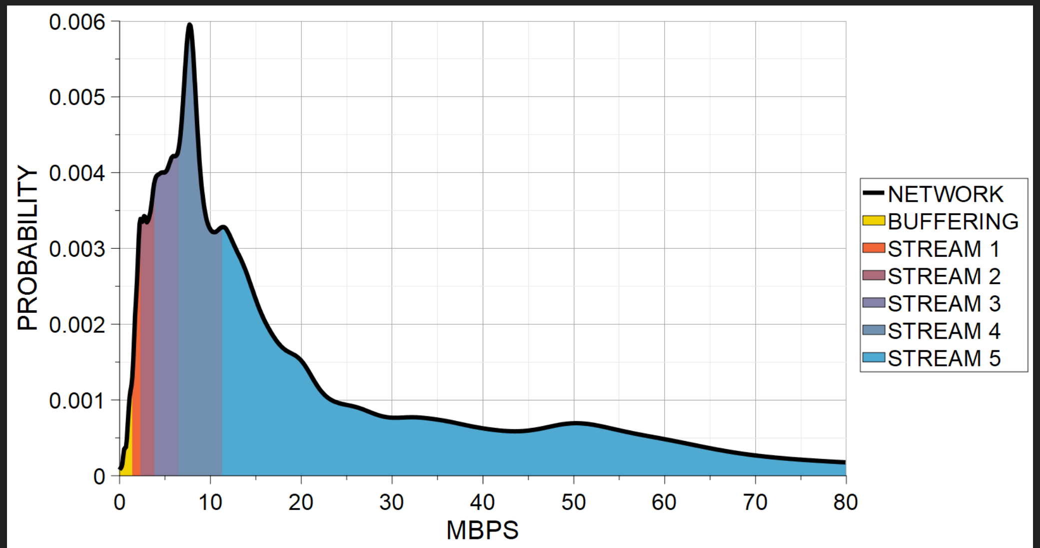
- Use network statistics as input for encoding profile generation
- Existing product: Brightcove's Content-Aware Encoding

Examples of ABR ladders

CAE-generated ladder

#	CODEC	FORMAT	RESOLUTION	FRAMERATE	BITRATE
1	HEVC	HDR10	960x540	59.94	1680
2	HEVC	HDR10	1280x720	59.94	2414
3	HEVC	HDR10	1920x1080	59.94	4050
4	HEVC	HDR10	2880x1620	59.94	6754
5	HEVC	HDR10	3840x2160	59.94	11483

Stream load probabilities



References

Y. Reznik, K. Lillevold, A. Jagannath, J. Greer, and J. Corley, "Optimal design of encoding profiles for ABR streaming," *Proc. Packet Video Workshop*, Amsterdam, The Netherlands, June 12, 2018.

Y. Reznik, T. Teixeira, and R. Peck, "On Multiple Media Representations and CDN Performance," *Proc. ACM Mile-High Video Conference (MHV'22)*, March 2022 Pages 56-61

APPLICATIONS

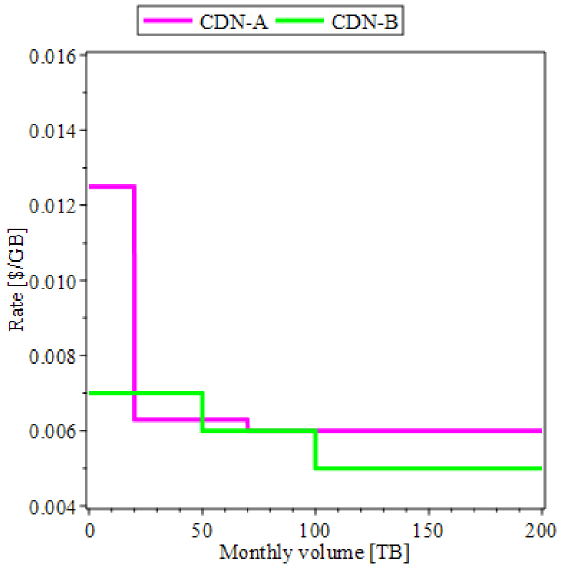
MULTI-CDN DELIVERY

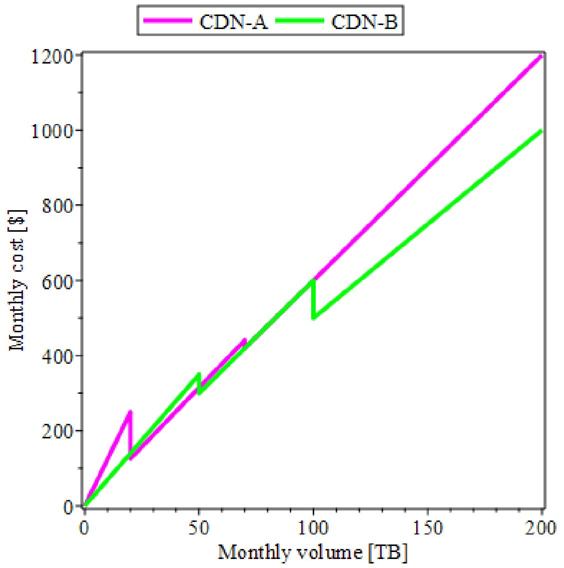
APPLICATIONS / UTILITIES

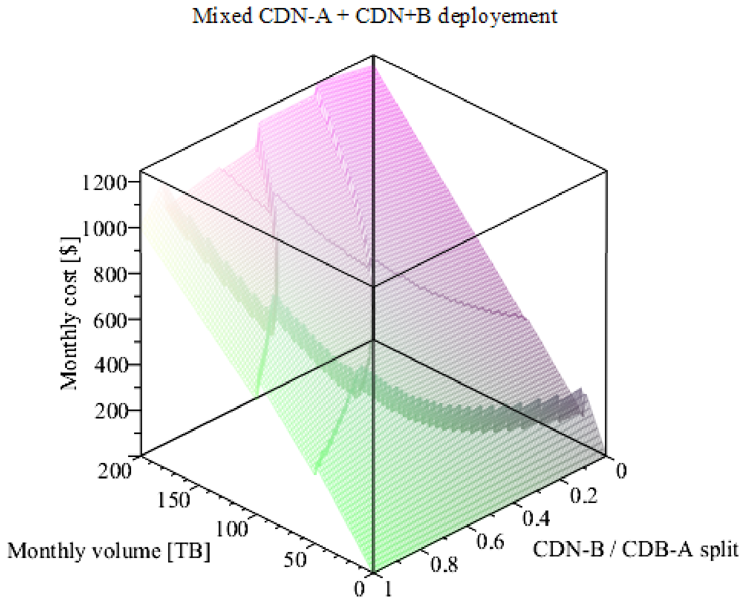
- Multi-region delivery
- Better scale (load balancing)
- Improved reliability (failover)
- Improved QOE (QOS/QOE optimizations)

BUT..CAN IT ALSO REDUCE THE COSTS?

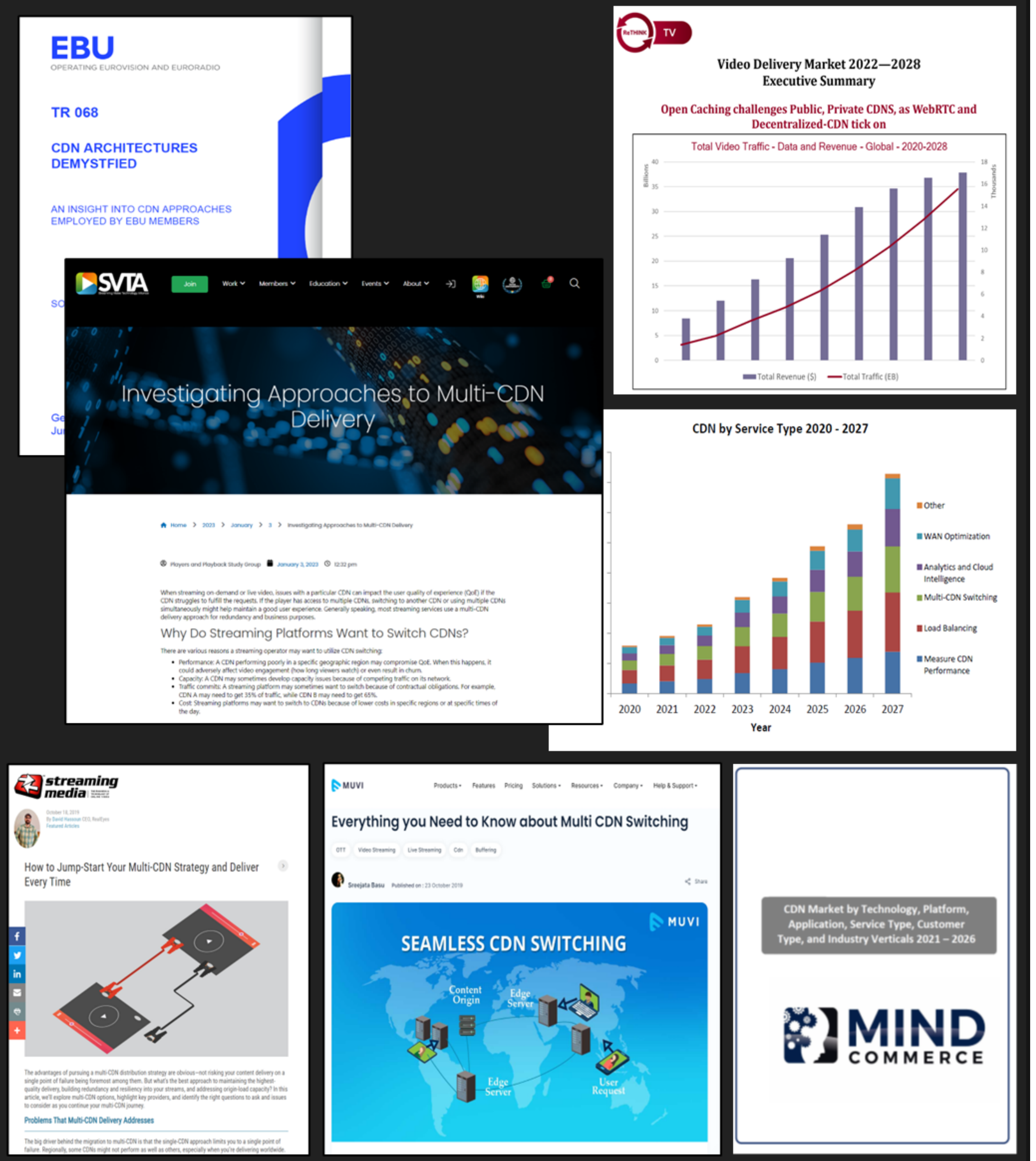
- The intuitive answer is no:
  - > Each CDN comes with a rate ladder
  - > Splitting the volume leads to higher rates





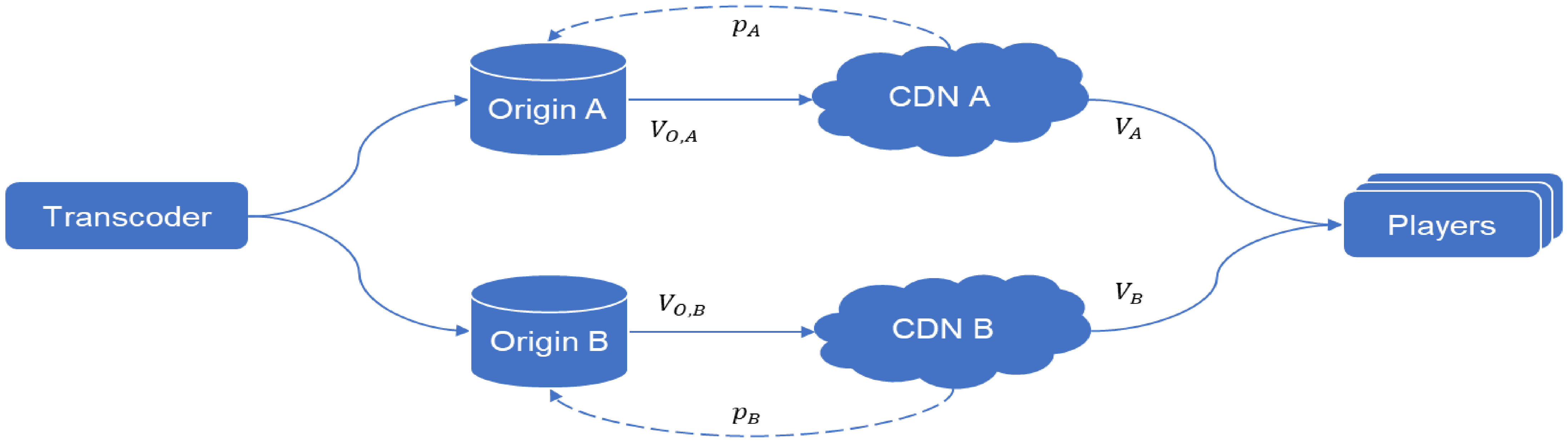


BIG TOPIC



MODEL OF MULTI-CDN SYSTEM

ARCHITECTURE



VARIABLES

- Edge volume: V_A, V_B
- Origin volume: $V_{O,A}$, $V_{O,B}$
- Cache miss probabilities: p_A, p_B

RELATIONSHIPS

- Origin/edge volume:

$$V_{O,A} = V_A \cdot p_A$$
$$V_{O,B} = V_B \cdot p_B$$

COST MODELS

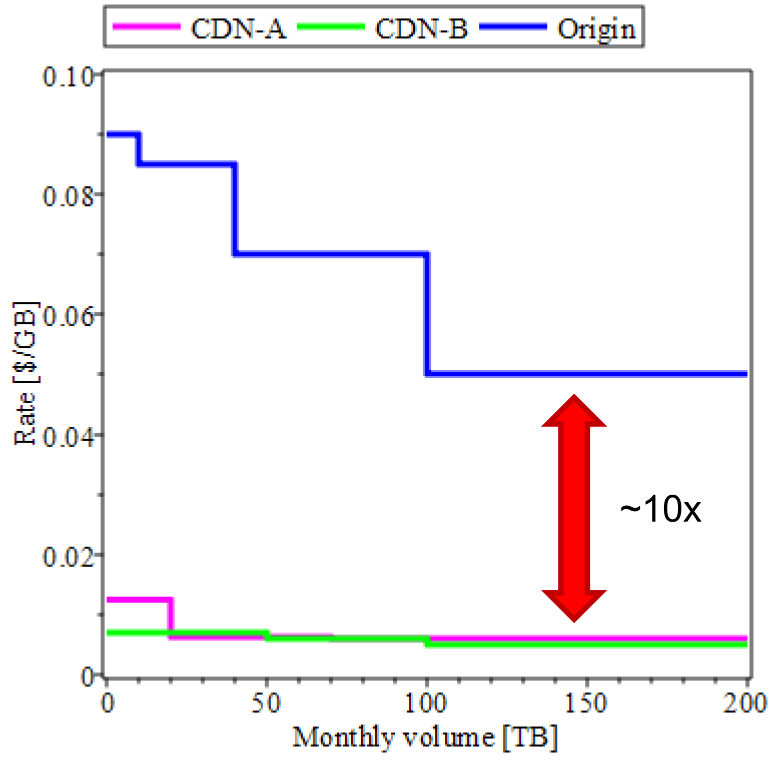
- CDN costs:

$$C_A = R_A \cdot V_A, \quad R_A = piecewise(V_A, \dots)$$
$$C_B = R_B \cdot V_B, \quad R_B = piecewise(V_B, \dots)$$

- CDN commits add extra constraints
- Origin costs:

$$C_{O,A} = R_{O,A} \cdot V_{O,A}, \quad R_{O,A} = piecewise\left(V_{O,A}, \dots\right)$$
$$C_{O,B} = R_{O,B} \cdot V_{O,B}, \quad R_{O,B} = piecewise\left(V_{O,B}, \dots\right)$$

- Generally, origins are more expensive



FULL COSTS

- Origin + CDN costs:

$$C_{\Sigma,A} = C_{O,A} + C_A = V_A\left(p_A \cdot R_{O,A} + R_A\right)$$
$$C_{\Sigma,B} = C_{O,B} + C_B = V_B\left(p_B \cdot R_{O,B} + R_B\right)$$

- Effective rates:

$$\boxed{\begin{aligned} R_{\Sigma,A} &= p_A \cdot R_{O,A} + R_A \\ R_{\Sigma,B} &= p_B \cdot R_{O,B} + R_B \end{aligned}}$$

LESS EXPENSIVE PATHWAY

COMPARING THE COSTS

- Assume that origin costs are the same:

$$R_{O,A} = R_{O,B} = R_O$$

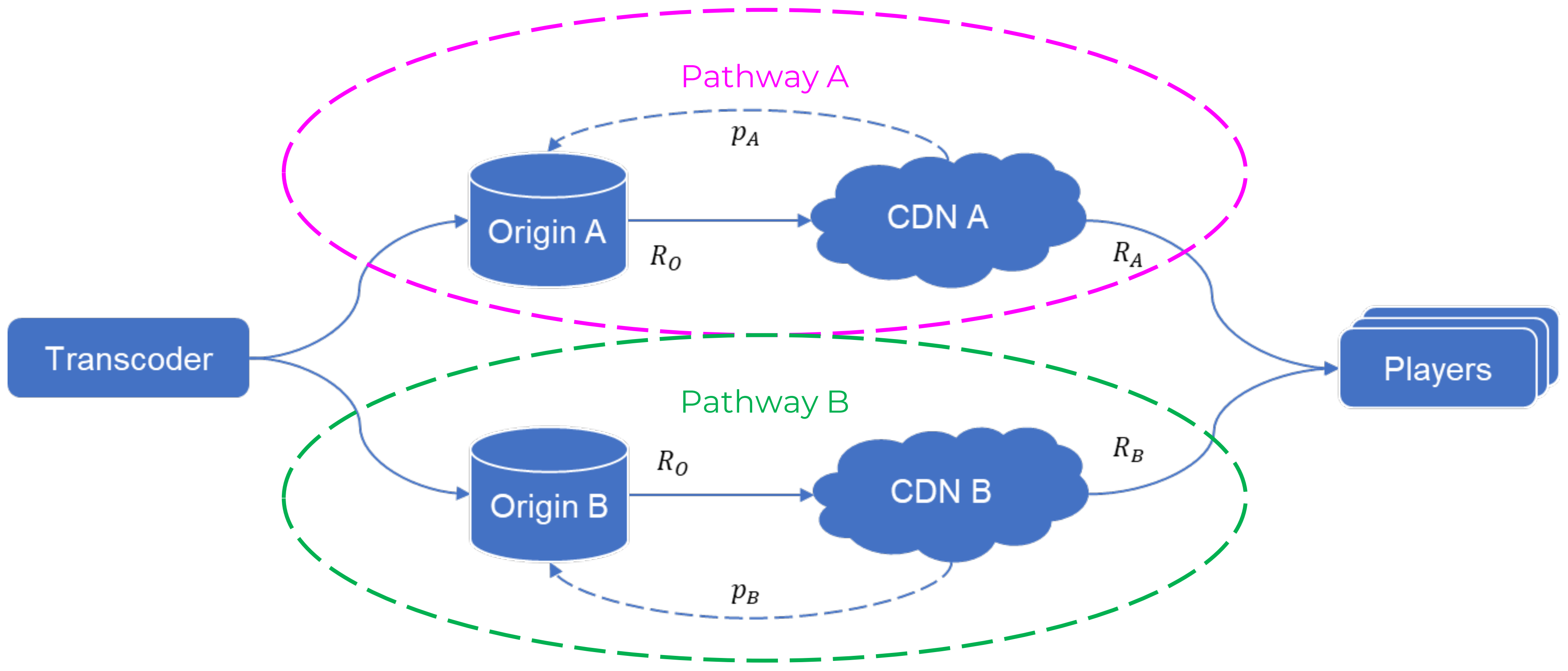
- Effective rates along each pathway:

$$R_{\Sigma,A} = p_A \cdot R_O + R_A$$
$$R_{\Sigma,B} = p_B \cdot R_O + R_B$$

- Pathway A is less expensive if:

$$R_{\Sigma,A} < R_{\Sigma,B} \quad \Rightarrow \quad p_A \cdot R_O + R_A < p_B \cdot R_O + R_B$$

$$\Rightarrow \quad \boxed{p_A - p_B < \frac{R_B - R_A}{R_O} = \xi}$$



EXAMPLES

SYNOPSIS	R_O	R_A	R_B	ξ	p_A	p_B	$p_A - p_B$	LESS EXPENSIVE PATHWAY
CDN A is cheaper & better in cache performance	0.02	0.002	0.0025	0.025	0.07	0.1	-0.03 < ξ	A
CDN A is cheaper & worse in cache performance	0.02	0.002	0.0025	0.025	0.1	0.07	0.03 > ξ	B
CDN A is more expensive & better in cache performance	0.02	0.0025	0.002	-0.025	0.07	0.1	-0.03 < ξ	A
CDN A is more expensive & worse in cache performance	0.02	0.0025	0.002	-0.025	0.1	0.07	0.03 > ξ	B

NB: Cache performance has a major impact on overall costs and choice of best pathway.

BRIGHTCOVE

CACHE MISS MODELS

FUNDAMENTAL RELATIONSHIP

- If content is popular, it gets cached with higher probability
- But…. such relationships may vary across CDNs, regions, and under different load
- To describe them we may use parametric models, e.g.:

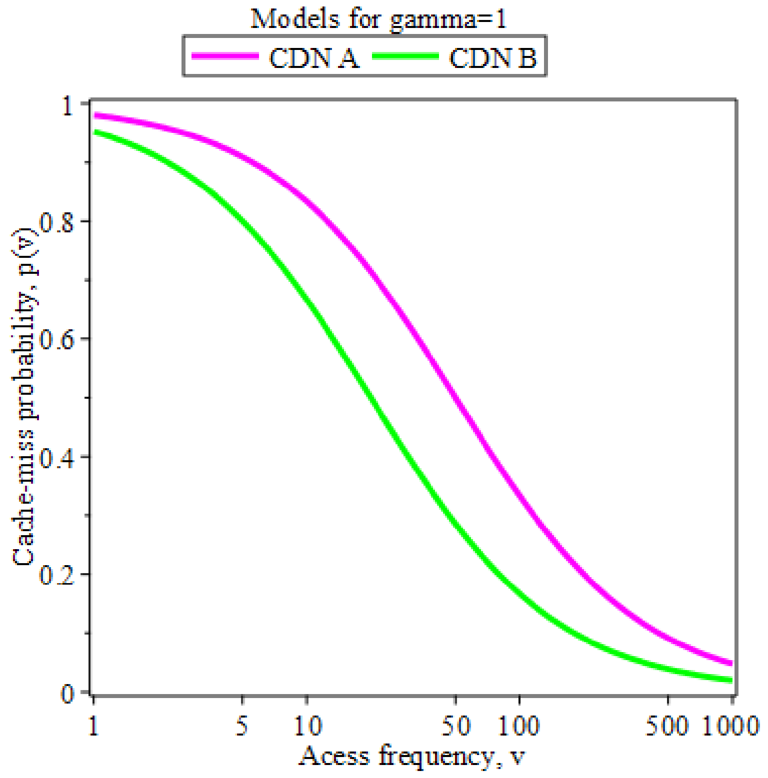
$$p_A(v) = \frac{1}{1+(v/v_A)^{\gamma}}, \qquad p_B(v) = \frac{1}{1+(v/v_B)^{\gamma}};$$

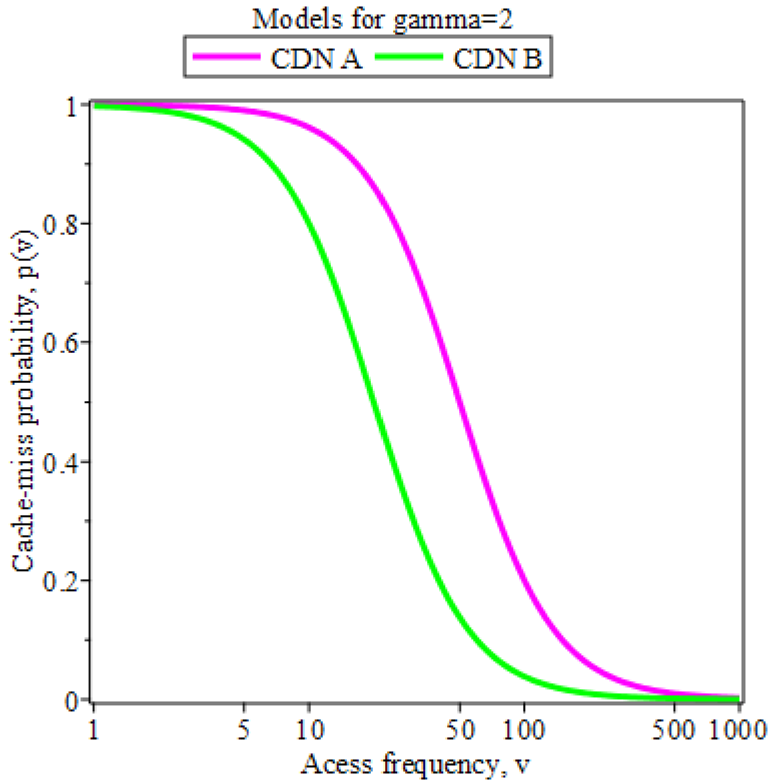
where

v	– content access frequency (e.g. requests/day)
$p_A(v),\ p_B(v)$	– cache miss probabilities of CDN A and CDN B, respectively
$v_A,\ v_B$	– CDN-specific model parameters
γ	– model shape parameter

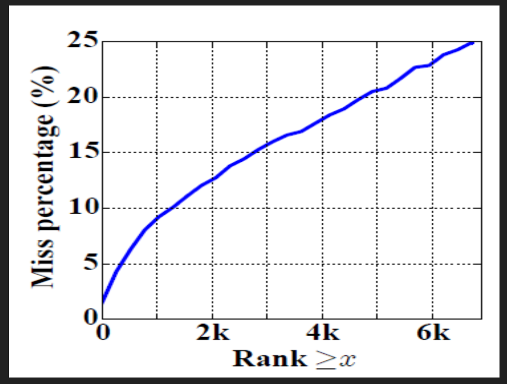
EXAMPLES

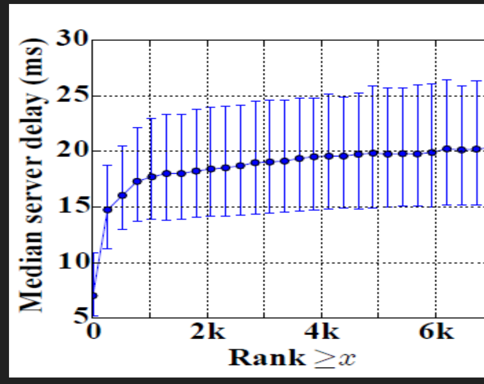
- Assume that
  $v_A = 50, v_B = 20$
- Model plots for
  $\gamma = 1, 2$:

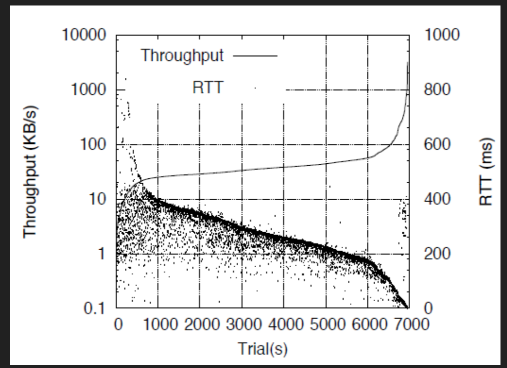


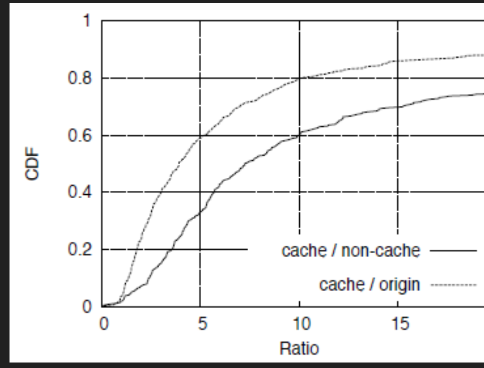


Related publications / studies









P. Franaszek and T. Wagner, "Some distribution-free aspects of paging algorithm performance," JACM, 21(1), 1974, pp.31-39.

P. R. Jelenkovic, "Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities," Ann. Appl. Probab., 9 (2), 1999, pp. 430–464.

S. Triukose, Z. Wen, and M. Rabinovich, "Measuring a Commercial Content Delivery Network," ACM WWW, 2011.

M. Ghasemi, P. Kanuparthy, A. Mansy, T. Benson, and J. Rexford, "Performance characterization of a commercial video streaming service," ACM ICM, 2016.

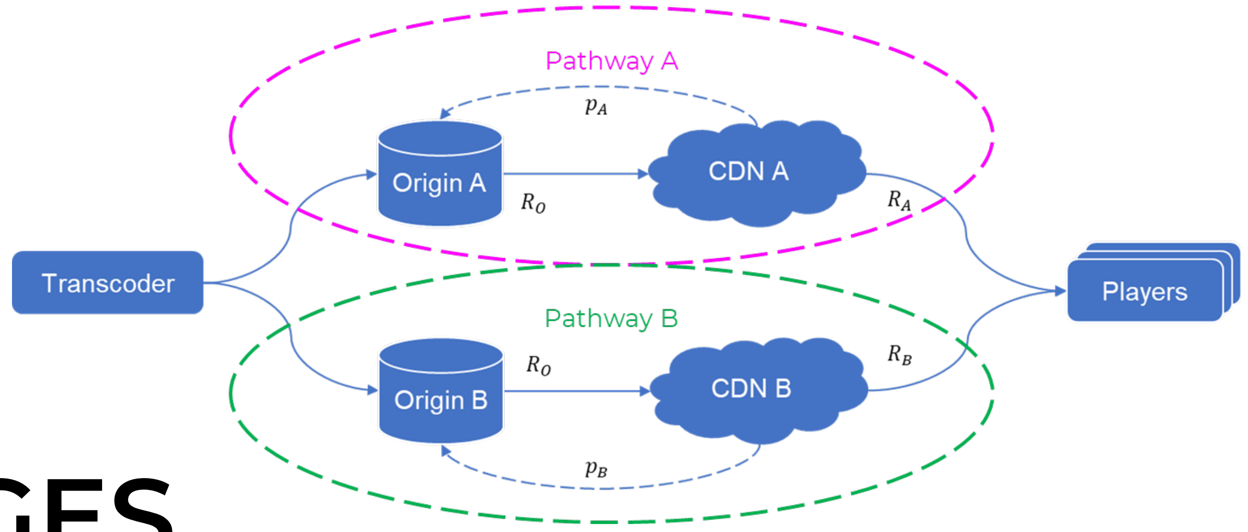
Y. Reznik, T. Teixeira, and R. Peck, "On multiple media representations and CDN performance," ACM MHV, 2022.

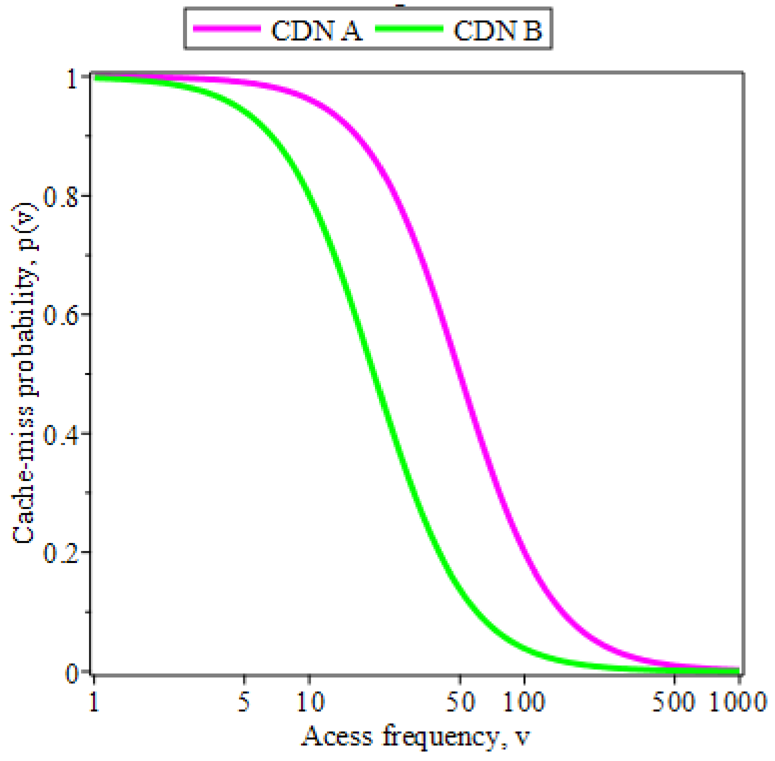
FINDING LESS EXPENSIVE PATHWAY

INITIAL SOLUTION

- For fixed p_A, p_B, we already established that:

$$R_{\Sigma,A} < R_{\Sigma,B} \;\Rightarrow\; \boxed{p_A - p_B < \frac{R_B - R_A}{R_O} = \xi}$$





ACCESS FREQUENCY RANGES

- Models for cache miss probabilities:

$$p_A(v) = \frac{1}{1 + (v/v_A)^{\gamma}}, \qquad p_B(v) = \frac{1}{1 + (v/v_B)^{\gamma}}$$

- Solution w.r.t. access frequency v:

$$R_{\Sigma,A} < R_{\Sigma,B} \;\Rightarrow\; \boxed{v \in \begin{cases} [0,\infty) & if \quad v_A < v_B \text{ and } R_A < R_B \\ (v_1^*, v_2^*) & if \quad v_A < v_B \text{ and } R_A > R_B \\ [0, v_1^*) \cup (v_2^*, \infty) & if \quad v_A > v_B \text{ and } R_A < R_B \\ \emptyset & if \quad v_A > v_B \text{ and } R_A > R_B \end{cases}}$$

- Where $v_1^* < v_2^*$ are the real positive roots of

$$\frac{1}{1 + (v/v_A)^{\gamma}} - \frac{1}{1 + (v/v_B)^{\gamma}} = \xi$$

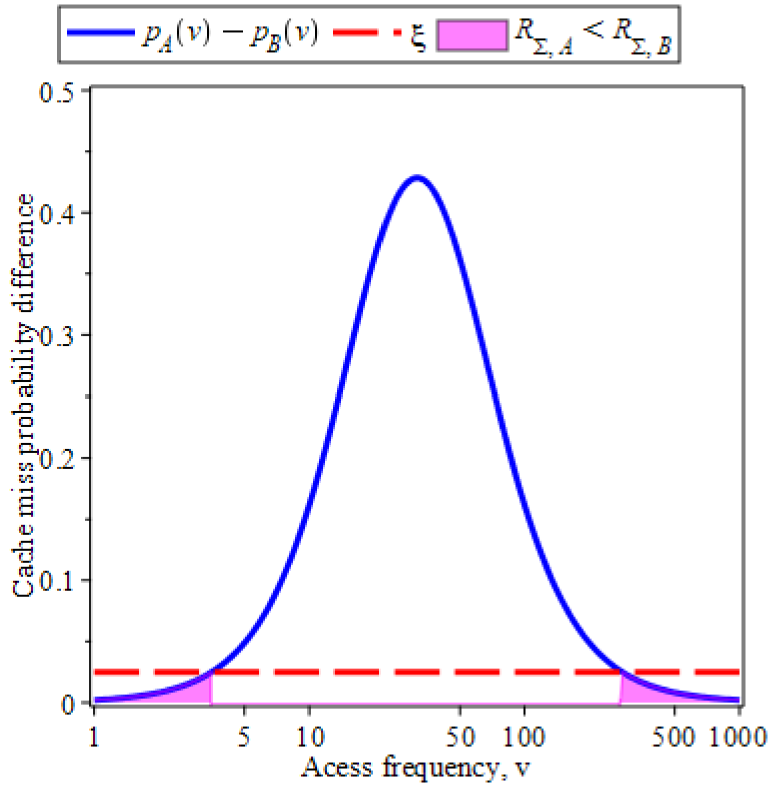
- E.g., for $\gamma = 2$:

$$v_1^* = \frac{1}{\sqrt{2\xi}}\sqrt{v_A^2 - v_B^2 - \xi(v_A^2 + v_B^2) - \sqrt{(v_A^2 - v_B^2)(v_A - v_B - \xi(v_A + v_B))(v_A + v_B - \xi(v_A - v_B))}}$$

$$v_2^* = \frac{1}{\sqrt{2\xi}}\sqrt{v_A^2 - v_B^2 - \xi(v_A^2 + v_B^2) + \sqrt{(v_A^2 - v_B^2)(v_A - v_B - \xi(v_A + v_B))(v_A + v_B - \xi(v_A - v_B))}}$$

EXAMPLE

- CDN A is cheaper: $R_A < R_B$, $\xi = 0.025$
- But worse as a cache: $v_A = 50, v_B = 20$
- Roots: $v_1^* \approx 3.51$ and $v_2^* \approx 284.76$
- The solution: $v \in [0, v_1^*) \cup (v_2^*, \infty)$
- NB: using pathway A in this case makes sense only for high access or long tail content!



BRIGHTCOVE

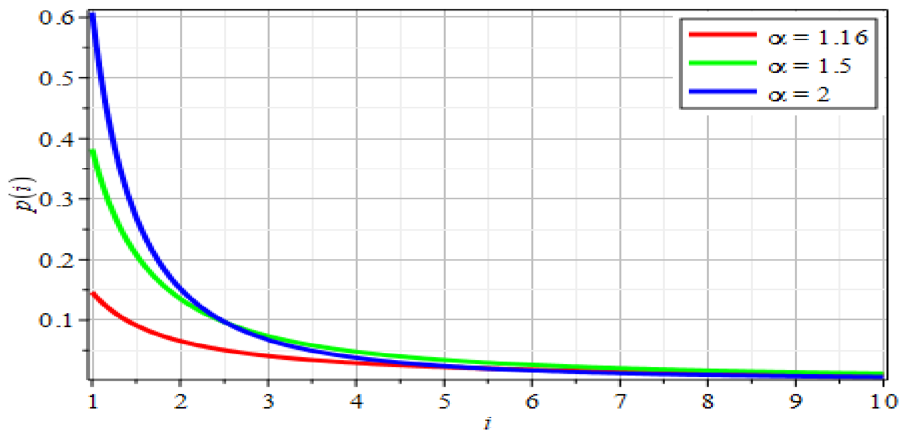
VOD: BEST PER-ASSET CDN ASSIGNMENT

CONSIDER A LARGE CATALOG

- Videos are ordered according to access frequencies
- Follow Zeta distribution:

$$u(i) = \zeta(\alpha)^{-1}\, i^{\alpha}$$

where α is a shape parameter, and $\zeta(\alpha)$ is the Riemann's Zeta function, i is an asset index.

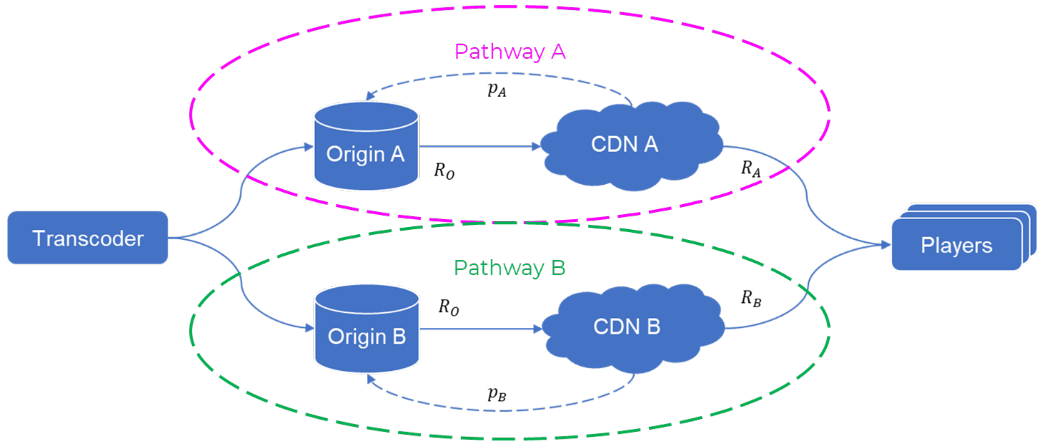


BEST PER-ASSET CDN ASSIGNMENT

- For given CDN prices R_A, R_B and cache miss models $p_A(v), p_A(v)$, we can show that:

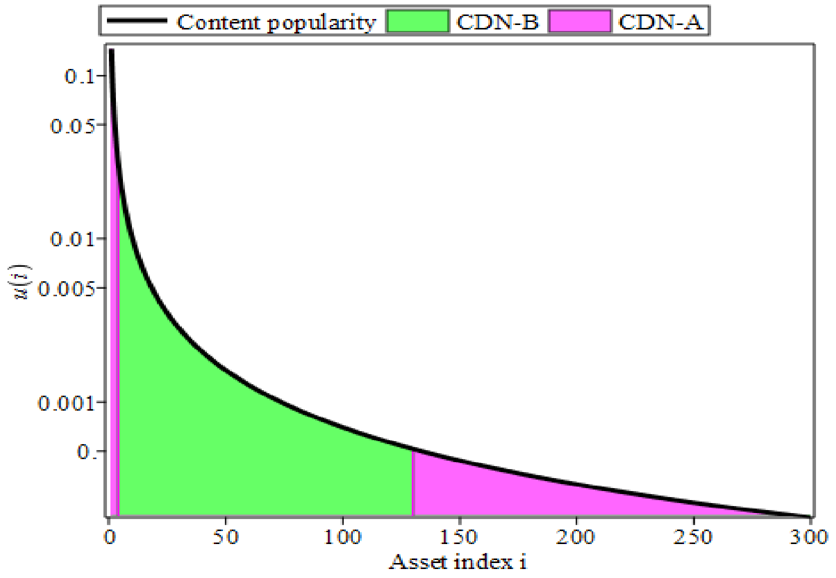
$$R_{\Sigma,A} < R_{\Sigma,B} \;\Rightarrow\; \boxed{i \in \begin{bmatrix} [1,\infty) & if & v_A < v_B \text{ and } R_A < R_B \\ (i_1^*, i_2^*) & if & v_A < v_B \text{ and } R_A > R_B \\ [1, i_1^*) \cup (i_2^*, \infty) & if & v_A > v_B \text{ and } R_A < R_B \\ \emptyset & if & v_A > v_B \text{ and } R_A > R_B \end{bmatrix}}$$

where $i_1^* = (v_2^*/C_\alpha)^{-1/\alpha}$, $i_2^* = (v_1^*/C_\alpha)^{-1/\alpha}$ are boundary points, C_α - normalization constant



EXAMPLE

- CDN A is cheaper: $R_A < R_B$, $\xi = 0.025$
- Worse as cache: $v_A = 50, v_B = 20, \gamma = 2$
- Roots: $v_1^* \approx 3.51$ and $v_2^* \approx 284.76$
- Content distribution: $\alpha = 1.16$, $C_\alpha = 1000$
- Boundary points: $i_1^* \approx 3$, $i_2^* \approx 130$
- Solution for CDN-A: $i \in [1, i_1^*) \cup (i_2^*, \infty)$



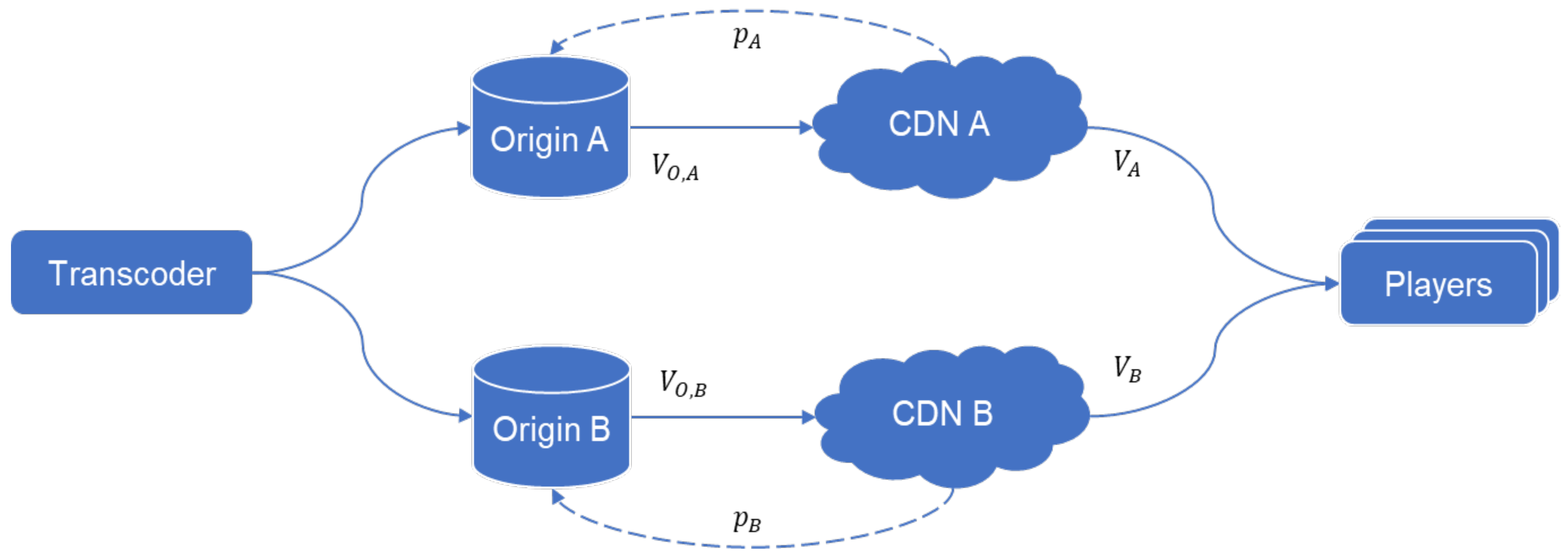
Reference: Y. Reznik, et al, "Reducing Delivery Costs by Optimal Multi-CDN Traffic Allocation", ACM Mile-High Video, Denver, CO, May 2023.

BRIGHTCOVE

VOD: COST OPTIMIZATION PROBLEM

GIVEN

- $R_A(V), R_B(V)$ – price/rate ladders for CDNs A and B
- $V_{A,\min}, V_{B,\min}$ – minimum volume commits for each CDN
- $R_{O,A}(V), R_{O,B}(V)$ – rate ladders for origins A and B
- $p_A(v,V), p_B(v,V)$ – cache miss models for CDNs A and B
- $u(i),\ i \in [1,N]$ – content popularity distribution across catalog
- $V_\Sigma = V_A + V_B$ – total volume delivered by the system



FIND

- $i_A^*, i_B^*:\ i_A^* \cup i_B^* = [1,N]$ – subsets of catalog items routed to CDN A and B, respectively

SUCH THAT

$$\boxed{C_{\Sigma,A}(i_A^*) + C_{\Sigma,B}(i_B^*) = \min_{\substack{i_A, i_B:\ i_A \cup i_B = [1,N] \\ V_A(i_A) \geq V_{A,\min} \\ V_B(i_B) \geq V_{B,\min}}} C_{\Sigma,A}(i_A) + C_{\Sigma,B}(i_B)}$$

Reference: Y. Reznik, et al, “Reducing Delivery Costs by Optimal Multi-CDN Traffic Allocation”, ACM Mile-High Video, Denver, CO, May 2023.

WHERE

- $V_A(i_A) = \sum_{i \in i_A} V_\Sigma\, u(i),\quad V_B(i_B) = \sum_{i \in i_B} V_\Sigma\, u(i)$ – edge volumes delivered by CDN A and B, respectively
- $V_{O,A}(i_A) = \sum_{i \in i_A} V_\Sigma\, u(i) \cdot p_A\left(V_\Sigma\, u(i), V_A(i_A)\right),\quad V_{O,B}(i_B) = \sum_{i \in i_B} V_\Sigma\, u(i) \cdot p_{AB}\left(V_\Sigma\, u(i), V_B(i_B)\right)$ – volumes processed by each origin server
- $C_{\Sigma,A}(i_A) = V_{O,A}(i_A) \cdot R\left(V_{O,A}(i_A)\right) + V_A(i_A) \cdot R_A\left(V_A(i_A)\right),\quad C_{\Sigma,B}(i_B) = V_{O,B}(i_B) \cdot R\left(V_{O,B}(i_B)\right) + V_B(i_B) \cdot R_B\left(V_B(i_B)\right)$ – total costs along each pathway

CONCLUSIONS

ANALYSIS OF CACHING ALGORITHMS IS AN OLD SCIENCE

- LRU/FRU schemes have been extensively studies since 1960s
- Many good results exists
- Allow simple extensions to multi-format or ABR-ladder-type cases

HIGHLY USEFUL TODAY

- Analysis and optimizations of CDN-based delivery systems
- ABR, multi-format, and multi-codec systems
- Multi-CDN systems
- Hybrid delivery systems
- Etc.

THANK YOU