Dynamic Deployment and Optimization of Virtual Content Delivery Networks

Nicolas Herbaut∗, Daniel Negru∗, David Dietrich†, Panagiotis Papadimitriou§,
∗Univ. Bordeaux, France
†Leibniz Universität Hannover, Germany
§University of Macedonia, Greece

Abstract

Today, Over-the-Top video streaming is gaining a lot of popularity. In this respect, the Virtual Content Delivery Network (CDN) is perceived as a key enabler to circumvent the technical challenges faced by Content Providers (CP) to deliver high-quality content over the Internet. In this paper, we investigate how the two main actors of the video delivery chain, i.e., the CDN operator and the Internet Service Provider (ISP), can benefit from network and server virtualization to negotiate Dynamic Service Level Agreements that reduce CDN CAPEX and OPEX, while generating more revenue for the ISP. First, we present a dataset used to simulate dynamic distributed traffic consumption. Second, we discuss the steps required to deploy and operate a virtual CDN deployed on an ISP’s network. Furthermore, we present evaluation results of the proposed solution, based on simple models. Lastly, we elaborate on operational parameters that are used to further optimize the solution.

Index Terms

Video, Network communications, Collaborative Services Delivery Platform, Modeling and prediction, Modeling of resources

I. VIDEO DELIVERY: AN EVOLVING ECOSYSTEM

Over many years, the consumption of Video On Demand streaming services Over-the-Top has been in constant progression. In particular, it has reached more that 55% of Internet Peak Traffic and is still increasing [1]. Shifts in the content distribution have hence become necessary, leading the path to Content Delivery Networks and Over-the-Top approaches.

The Content distribution ecosystem that has emerged during the past decade is structured around a variety of actors along a value chain, as presented in Figure 1. Some actors fulfill a technical role, such as ISP and CDN operators, whereas others are more oriented toward business, such as CP.

This value chain governed the democratization of online videos but the current surge of Over-the-Top IP-based streaming reshuffles the deck. Specifically, ISPs are left behind, letting the added value of content leaking through their pipes without reaping the benefits, while CDN operators must build increasingly expensive networks targeting a worldwide audience. At the same time, CP and Content Owner benefit from online videos market growth. In this this ecosystem, collaboration between actors is key to tackle those challenges, with a specific focus on finding synergies.

ISPs are solicited by CP or CDN to install streaming appliances within their network to deliver content directly to their end-users (e.g., Netflix Open Connect), reducing the number of hops as well as inter-Autonomous System (AS) traffic [2]. These solutions reduce both CDN and ISP costs and ease technical cooperation by enabling a joint control of the delivery [3]. However, collaboration is not fair, as the ISP does not benefit from the revenue generated by the CP. Finally, we witness a trend where companies concentrate the different roles of the value chain in one entity [4]. Although this may facilitate end-to-end operations, it can hinder innovation, raising the barrier for newcomers on the market. Based on these observations, we introduce the concept of CDN as a Virtual Network Function (VNF) to foster a win-win collaboration between CP/CDNs and ISPs.

The article is organized as follows. We present the vCDN concept and its main use cases. We detail the main steps to deploy and operate the the proposed solution. Finally, we conclude the whole article.

Fig. 1: Actors playing a role in the content distribution value chain and financial transaction between them
II. THE VIRTUAL CDN CONCEPT

Internet Service Providers, despite their relative decline in content delivery market still have a critical role. First of all, ISPs are the only actors capable of controlling the network end-to-end. Second, they invested tremendously in connectivity and system capacities at the edge, to support new service such as such as Telco CDN [5] or virtual Customer Premise Equipments. Finally, with the industrialization of Network Function Virtualization (NFV) and Software Defined Networking (SDN), ISPs have now the means to market “Connectivity as a service”, as promoted by several collaborative projects such at [6], making the separation of Infrastructure Provider and Network Service Provider a reality [7]. As such, ISPs can bring to the Telco world all the benefits that the Cloud brought to the IT world, i.e., reduced costs, scalability, and lower barriers to entry.

Using NFV is envisioned as an opportunity for enhancing multimedia delivery [8]. In our previous work, we went further by proposing that the ISP can rely on NFV to sell vCDN it as-a-service to CDN operators and CP [9]. In Figure 2, the ISP View shows two Network Function Virtualization Infrastructure Points of Presence (NFVI-POPs) A and B deployed inside the ISP network. NFVI-POPs are datacenters that provide computing resources (CPU, RAM, storage) for Virtual Network Functions (vNFs), in accordance with the “Network Function as a Service” (NFaaS) paradigm [10]. These NFVI-POPs are located close to the edge of the network where end-users (d1 and d2) are present. In particular, NFVI-POPs can be connected with physical links and switches possibly supporting Software Defined Networking (SDN) to facilitate network management.

On the other hand, the CDN view abstracts all the underlying ISP topology. The CDNs have only access to an overlay network that connects the end-users and their vCDNs servers. As such, the underlying network is concealed from the CDNs, preserving the confidentiality of the ISP substrate topology and resources.

Four main use cases are foreseen to strongly motivate the usage of the vCDN approach:

- **Footprint extension in low density area.** A CDN wants to expand its connectivity in an area where it is not economically sustainable to set up a permanent peering.
- **Quality improvement for niche market.** A CDN wants to deliver high-quality media (4K, 8K...) over HTTP, with the requirement of a very small delay towards the streaming server.
- **Bandwidth burst.** A CDN needs to temporarily increase its bandwidth in a geographical area to serve more end-users.
- **Virtual CDN operator.** ISPs bears entirely the content delivery, Making possible experimentation with novel content distribution paradigms through virtualization.

The variety of use cases indicates that the deployment and management of vCDNs requires flexibility and careful design of several key business and network aspects. In the following section, we elaborate on five steps envisioned for vCDN deployment and operation.

III. VCDN STEPS OF DEPLOYMENT AND OPERATION

We exclusively focus on building the virtual *infrastructure* to deliver content, based on the already deployed ISP POPs. We assume that the problem of selecting the content to be cached is solved, as it is part of the core business of the CDN. In the following, we use the term “CDN” to refer to the actor that purchases the vCDN service, irrespective whether it is a CDN operator (e.g., Akamai, Limelight) or a large CP (e.g., Netflix, Twitch).
We show the five major steps required to deploy and run a virtual CDN in Figure 3. First, the CDN needs to (1) collect historical data to (2) predict his future bandwidth consumption. The CDN use these traffic predictions to (3) generate Service Level Agreements (SLAs) that formalize its the contract with the ISP. Finally, the ISP prices each SLA and (4) embed as a Service Function Chain onto its network. As connectivity demand is very dynamic, the ISP further (5) optimizes services to cope with the fragmentation of its components to reduce costs. We discuss these five steps more details in the following.

A. Data Collection and Hypotheses

Collecting historical data is a must for the CDN operator in order to understand its traffic volume and provision servers and connectivity. In parallel, the ISP needs to take into account its vNF characteristics to determine how many servers will be used to host the software components. In this section, we present the data and hypotheses collected to support our evaluation.

**CDN traffic estimates.** Traffic data is not publicly available; however, reasonable estimates can be inferred from Internet Exchange Points (IXP). IXPs consist of infrastructures where the ISP and the CDN usually establish peering between their Autonomous Systems (AS). We extracted publicly available daily statistics of more than 50 peering points from the service monitoring webpages of 4 IXPs. The traffic of each peering point aggregates its specific mixture of Content, Enterprise Network and ISP AS. Using the data from the peeringdb.com API, we are able to precisely estimate the bandwidth allocated to each AS. Content AS represents a very significant part of the traffic, as shown on Figure 4, so we calculated each content AS traffic by dividing the total traffic by the AS bandwidth share. The dataset is updated daily and is available at http://www.labri.fr/perso/nherbaut/vcdn-dataset.html.

**VNF costs estimates.** Through the work performed at EU project T-NOVA [11], we prototyped and tested a vCDN Service on an NFV testbed. From there, we derived benchmarks to obtain the number of virtual machines needed to support a given bandwidth and combined those estimates with data from a public cloud price list to formulate the hypotheses on costs. Table I reports the computed OPEX results related to the two VNFs composing the vCDN service. The way these VNFs are combined to create the vCDN service is discussed in Section III-D. We further note that those estimates are obtained from public prices and prototype implementation and, thus, they may need to be revised according to negotiated prices and production-grade software.

**Network costs and ISP Topology estimates.** ISPs do not disclose their topology and their NFV Infrastructure deployment is still ongoing. In this respect, we considered synthetic network topologies (Erdos-Renyi, powerlaw) and performed the all the simulations over the Geant topology, since we have access to bandwidth and latencies across the 35 edges. We configured system capacities randomly over the 28 nodes.
### TABLE I: VNFs OPEX estimates

<table>
<thead>
<tr>
<th>vNF Type</th>
<th>vCPU</th>
<th>RAM</th>
<th>Storage</th>
<th>Ratio vm/Mbps</th>
<th>Equivalence</th>
<th>price per y</th>
</tr>
</thead>
<tbody>
<tr>
<td>vNF1 – Routing</td>
<td>2</td>
<td>4GB</td>
<td>20GB</td>
<td>1 VM for 1,250 Mbps</td>
<td>m4.large</td>
<td>46,252$</td>
</tr>
<tr>
<td>vNF2 – Streaming</td>
<td>36</td>
<td>244GB</td>
<td>48TB</td>
<td>1 VM for 5,000 Mbps</td>
<td>d2.8xlarge</td>
<td>51,508$</td>
</tr>
</tbody>
</table>

Fig. 5: Step 2 – Forecast, SLA and Discretization evaluations

In the next section, we use data collected from the IXPs to study the predictability of future bandwidth demand for content delivery.

### B. Traffic Prediction

For traffic prediction, we go beyond related work that relies on the prediction of resources (e.g., [12]), as we take into account changes in traffic demands. Traffic in CDNs, VOD in particular, follows a very specific diurnal pattern, as shown in Figure 5a. The patterns are also specific to geographical regions, due to time difference and broadcasting of local content (e.g., regional sport events). In addition, exogenous behavior may occur, e.g., when the first release of a popular TV program dramatically increases the data consumption. Such events cannot be incorporated into the model by relying only on historical data.

For traffic prediction, we use a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, which is well suited to the prediction of aggregated and backbone traffic [13] and can generate a forecasts easily. We also computed prediction Intervals (PI) to adjust the predicted traffic level according to model fitness.

We emphasized on measuring the accuracy of forecasts performed on a 48H window. To this end, we conducted the evaluation over 50+ peering points dataset belonging to 4 different IXPs (IXP1 to IXP4) collected over 3 weeks. We used the metrics MAPE (Mean Absolute Percentage Error, an easy to understand percentage measure of the accuracy of forecasts) and MASE (Mean Absolute Scaled Error, a more advanced measure which compares the performance of the forecast wrt a naive prediction) for the evaluation (depicted in Figure 5b). Based on MAPE, we observe that the models perform rather well with a 1.80% mean error on the whole dataset. The MASE metric shows mean scores around 1.1, with high standard deviations, meaning that depending on the considered time series, the proposed model may be very well adapted in some cases and not in others. Weaker results are due to technical maintenance occurring close to the prediction windows for IX1, as well as trend breaks on the peering point data caused by an AS being added or removed for IX2. Handling such planned events is possible with exogenous variables, but out of the scope of this paper. Once the predictability of the traffic demand has been achieved, the next step consists of the SLA generation by the ISP.

### C. SLA Generation

A vCDN SLA provides the assurance that a certain amount of bandwidth will be available to a given geographic area over a given time period. SLAs also contain two necessary specifications: (1) the maximum acceptable delay and (2) the list of existing peering points where legacy CDN traffic “exits” the ISP network. In [14], authors present a framework that exploits SLAs to create content delivery federations between ISP (for network), Cloud providers (for storage) and Content providers (purchasing the service). We follow a similar approach, albeit considering the ISP as the only actor providing the service.

CDN providers generate SLAS following three steps:

- The predictions \( \hat{t} \) are discretized to generate a step-like envelope of constant bandwidth demand chunks, as shown in Figure 6a. The algorithm filters them according to a smoothing window of length \( w \) and applies a K-Mean method, resulting in a discrete step-function with \( c \) possible values.
- The SLAs are generated by creating layers of demand chunks that have the same bandwidth value, thanks to the function \( D(\hat{t}, w, c) \) (see Figure 6b).
- The ISP applies its \( p \) pricing function to compute the right SLA price \( p_{SLAs} \) to be proposed to the CDN provider, according to the following equation:

\[
p_{SLAs} = p[D(\hat{t}, w, c)]
\]
By definition, forecasts aim at fitting the data unbiasedly, whereas the CDN needs a buffer to avoid underestimating its traffic. This buffer can be obtained by using Prediction Intervals (PIs), as a basis for SLA creation. Selecting a high prediction interval reduces SLA violations. As shown in Figure 5c, choosing 80% of 95% PI decreases the SLA violation rate to 8.2% of 4.3%, respectively. The strong advantage of using PIs is that they automatically adapt to the predictive power of the model. If data is not fitted well, the model generates large PI. The drawback of overestimating the forecast is that it leads to the increase of the resulting SLA prices, as the level of bandwidth provisioned for each SLA increases. This leads to a serious impact on the total pricing to the CDN provider, as shown in Figure 5d. Choosing the 80% or 95% PI models increases the price by respectively 10.5% or 16.1%.

From that point on, the CDN provider decides the acceptable level of risk/price ratio with respect to the expected quality. The ration is intended to be higher for sponsored contents than for premium ones.

D. SLA Embedding

After the CDN has generated the SLAs, the ISP deploys the service by embedding a vCDN onto his network. The vCDN solution (designed and prototyped in the T-NOVA project [11]) is composed of two vNFs chained together:

- The vStreamer (vSTR) constitutes an IP endpoint capable of streaming media content through HTTP. Link bandwidth toward vStreamer is reserved along the path in the ISP network.
- The Virtual Media Gateway (VMG) is a transit vNF, capable of analyzing the HTTP traffic, which can dynamically modify the routing towards content sources. For instance, if the VMG detects a request for a content available from the CDN and also from the vStreamer, it replaces the destination IP address, and IP packets are routed toward the vStreamer according to the virtual path with guaranteed QoS. The VMG is a fine-grained alternative to the traditional domain names-based routing schemes [2].

To compose the vCDN service, the Client Groups (CGs – locations of connected end-users) from one side and the Legacy CDN peering points (original locations of the content) from the other side are both taken into consideration.

One of the challenges in building the vCDN service is to adapt the service topology (CGs-VMGs-vSTRs-CDN points) to the ISP topology to lower the deployment cost. To this end, we proposed [15] an algorithm that uses heuristics to generate the service topology, then computes the embedding with a Linear Programming solver including both SLAs and ISP topology constraints. We use a cost function that minimizes ISP cost to produce the embedding of Services Function Chains. Figure 7 shows the phases of the embedding process (a demo is available on http://girafe.viotech.net).
E. Service Optimization.

We hereby present two optimization methods to cut the cost on the ISP side. We first elaborate on the possible service combinations, before studying the impact of the ISP pricing function on the SLA generated by the CDN.

Service Migration.

Over time, new SLAs may arrive or existing SLAs may expire, resulting in changes on the underlying services deployed by the ISP. If an SLA expires, the ISP releases the bandwidth reservations for the virtual links associated. On the other hand, when new SLAs arrive, the bandwidth demand increases. In this case, it may not be possible to embed the incoming SLAs using the same physical links or the same PoP, due to capacity constraints.

To alleviate this, we consider two strategies available to the ISP to minimize the embedding cost (Figure 8):

- **Strategy 1 “Separate Embedding”**: The ISP creates two services side by side: each service being in charge of implementing a subset of the SLAs required by the CDN.
- **Strategy 2 “Recombination+Migration”**: The ISP recombines all the SLAs within the same service. Even though it may appear more efficient for resource spreading, vNF migrations may be necessary, increasing the cost of the new service.

The ISP selects the strategy that minimizes the embedding cost, and subsequently deploys the new services, whenever there are new or expiring SLAs.

Figure 9a shows the evolution of the embedding cost over time. In particular, the figure highlights the variation of the optimal cost, bandwidth demand increases and decreases, and cost savings induced by Strategy 2 of recombining services (including migration costs). In the next months, we plan to release a comprehensive evaluation on the gains from the recombination mechanism that considers migration costs (Strategy 2 over Strategy 1), based on a large set of ISP topologies and extended hours of data collection, also considering the impact on NFV Orchestration softwares of such mechanism.

Adjusting the ISP pricing scheme.

The migration of vNFs is an opportunity to further optimize the service but is also a necessity to reduce service fragmentation. In this respect, we study how to reduce service fragmentation through the impact of the ISP pricing function on the SLAs.

ISPs and CDN providers have conflicting interests regarding SLAs. On one hand, the CDN provider would seek to generate SLAs “as close as possible” to its traffic forecast to limit the overhead of the discretization algorithm. The best strategy for the CDN would be to generate a new SLA for each forecast value, matching precisely the estimates. However, this strategy would lead to a large number of SLAs and, in turn, to the need for new service embeddings, resulting in substantial migration costs for the ISP. ISP thus wants to limit the number of SLAs to a bare minimum. To this end, the ISP can promote fewer and long-lasting SLAs through its pricing schemes. By applying a discount rate depending on the duration of the SLAs, the ISP can incentive the CDN to adopt its discretization strategy to reduce the number of SLAs.
(a) Example of the evolution of the cost of Service Embedding with SLAs generated from 4 IXPs, 4 legacy CDN peering points over 24h

Fig. 9: Evaluation impact of service optimization and discount policy

To understand how does the discount ratio influence the ISP revenue, we conducted the following experiment:

- we considered an exponential-decaying pricing function for SLAs up to 24h.
- we assumed that the CDN generates the SLAs by choosing the discretization parameters that minimizes its cost wrt its traffic predictions.
- we ran a simulation with the provided SLAs, and compute the overall embedding cost for the ISP.
- we computed the ISP profit from the ISP cost and the CDN cost, repeating the experiment for several discount ratios and cost hypotheses.

Figure 9b shows the results obtained. We considered 3 hypothesis (Low, Average and High) depending on the cost structure of the ISP, taking into account VM hosting, networking and VM migration costs. The main outcome is that the cost structure of the ISP influences the relevance of implementing the discount mechanism:

- In our high hypothesis, the discount mechanism must be implemented. In fact, the service is lucrative only for discounts ratios between 25% and 65%, the highest profit margin value being around 40%.
- In our average hypothesis, the discount mechanism has little impact on the ISP revenue for values below 35%. The ISP can therefore use the discount mechanism to adjust its offer to the market price for CDN service without hurting its profit.
- In our low hypothesis, implementing the discount mechanism is not effective, as it reduces the profit.

To conclude, the discount mechanism should be confronted to each ISP cost structure and can influence the hardware and software investments decisions mechanisms [18] to improve OPEX.

IV. CONCLUSIONS

The deployment and operation of vCDN, as reported in this study, would bring strong added value to a wide range of use cases and, specifically, to the four identified in Section II, as follows.

Footprint extension: For instance, one of the missions of the FEDER European development program is to fight against “Numerical Deserts” by injecting substantial amount of money towards strengthening access to ICTs on rural areas. The ISP has a role to play there, not only to bring connectivity but also to propose efficient -at all levels, including green- content availability thanks to virtualized vCDN platforms. The ease of deployment, operation and optimization, provided by the Generation, Embedding and Optimization steps of our proposal (Steps 3-4-5) should contribute to the appropriation of such use case by the ISP.

Quality improvement for niche market: regarding Ultra High Definition instant streaming services over HTTP, maximal delays needed to reach the 100 or 200Mbps marks are one order of magnitude greater than typically experienced ones today. Integrating delivery solutions within ISP networks through a vCDN deployment will greatly help Content Providers to target this UHD niche market. Indeed, thanks to the Prediction, Generation and Optimization steps (Steps 2-3-5) proposed, prediction and assurance of end-to-end latencies are possible.

Bandwidth burst: As Content Providers still face outages when releasing their most expected content to the market, vCDN could help offloading their own networks by pushing their content even deeper in the ISP Network. Such operation can as well be guaranteed in terms of results, as shown in the Prediction and Generation steps (Step 2-3), according to an interval. The strong depicted advantage towards this is the cost of such an operation, with strong gains with regards to physical deployments but also related to the desired guarantee level.
Virtual CDN operator: Virtualizing the content delivery can ease experimenting with novel architectures and algorithms such as content Centric Networks and multi-source delivery protocols. We plan in future work to quantitatively assess the gain for every actor (CP, CDN, ISP) of the vCDN.

REFERENCES


Nicolas Herbaut is a Ph.D. candidate at Bordeaux University doing research on Virtualization, Network Softwarization and Content Delivery. He received an M.Sc.Eng in Mathematics from INSA Rouen in 2004, and a M.Econ in Economical Analysis from Aix-Marseille University in 2005. He worked as a Technical leader for several companies. Contact him at nicolas.herbaut@labri.fr.

Daniel Negru Dr. Daniel NEGRU is associate professor at Bordeaux University, specializing in multimedia and networking. His current activities are focused on video streaming, content delivery, NFV/SDN. He has participated numerous collaborative research projects at the European level, published more than 60 papers in journals and conferences such as IEEE Communications Magazine, IEEE Multimedia, ICME, Globecom, ICC. Contact him at daniel.negru@labri.fr.
David Dietrich received the doctoral degree (Dr.-Ing.) from the Leibniz Universität Hannover, Germany, in 2016 and the M.Sc. from the Universität Kassel, Germany, in 2006. In 2010, he joined the Institute of Communications Technology at the Leibniz Universität Hannover where he works as postdoctoral researcher since 2016. His research interests include network virtualization and network management with emphasis on algorithms and optimization. Contact him at david.dietrich@ikt.uni-hannover.de.

Panagiotis Papadimitriou is faculty at the department of Applied Informatics in the University of Macedonia, Greece. During 2011–2016, he was an Assistant Professor at the Communications Technology Institute of Leibniz Universität Hannover, Germany. His research activities span next-generation Internet architectures, NFV, SDN, cloud networking, and mobile edge computing. Panagiotis is a Senior Member of IEEE. Contact him at papadimitriou@uom.edu.gr.