Article

Joint Location-Dependent Pricing and Request Mapping in ICN-Based Telco CDNs For 5G

Mingshuang Jin ^{1,*} , Hongbin Luo ^{2,*}, Shuai Gao ¹ and Bohao Feng ¹

¹ School of Electronic and Information Engineering, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing 100044, China; shgao@bjtu.edu.cn (S.G.); bohaofeng@bjtu.edu.cn (B.F.)

² School of Computer Science and Engineering, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China

* Correspondence: mshjin@bjtu.edu.cn (M.J.); luohb@buaa.edu.cn (H.L.)

Received: 29 April 2019; Accepted: 31 May 2019; Published: 3 June 2019

Abstract: Telco content delivery networks (CDNs) have envisioned building highly distributed and cloudified sites to provide a high-quality CDN service in the 5G era. However, there are still two open problems to be addressed. First, telco CDNs are operated upon the underlay network evolving towards information-centric networking (ICN). Different from CDNs that perform on the application layer, ICN enables information-centric forwarding to the network layer. Thus, it is challenging to take advantage of the benefits of both ICN and CDN to provide a high-quality content delivery service in the context of ICN-based telco CDNs. Second, bandwidth pricing and request mapping issues in ICN-based telco CDNs have not been thoroughly studied. In this paper, we first propose an ICN-based telco CDN framework that integrates the information-centric forwarding enabled by ICN and the powerful edge caching enabled by telco CDNs. Then, we propose a location-dependent pricing (LDP) strategy, taking into consideration the congestion level of different sites. Furthermore, on the basis of LDP, we formulate a price-aware request mapping (PARM) problem, which can be solved by existing linear programming solvers. Finally, we conduct extensive simulations to evaluate the effectiveness of our design.

Keywords: Telco CDN; ICN; pricing; request mapping; optimization

1. Introduction

Content delivery networks (CDNs) have become indispensable for data dissemination on the Internet. By employing caches that are proximal to end users, CDNs provide low-latency and better data availability, increasing the quality of experience (QoE). According to the Cisco Visual Networking Index, CDNs will carry 72% of Internet traffic by 2022, up from 56% in 2017 [1]. With the continuous growing requirements of bandwidth-hungry and latency-sensitive applications such as 8K video and automatic driving, more and more CDN companies are deploying their CDN sites closer to end users, i.e., in access networks. In this trend, telecommunication companies/mobile network operators (MNOs) have more incentives and requirements than ever to deploy more CDN sites in their hierarchical cloudified networks [2–4]. On one hand, they are re-designing the network architecture to support better flexibility and higher efficiency by introducing software-defined networking (SDN) and network function virtualization (NFV) techniques. The re-designed architecture will help provide a better content delivery service at a lower cost (e.g., by saving bandwidth). On the other hand, the representative scenarios in the 5G era require rigorous latency and bandwidth guarantees, which cannot be satisfied by existing remote CDN sites alone. Thus, sites in access networks are needed. In addition, MNOs have a cost advantage over traditional CDN companies when providing CDN services in access networks. The reason is that MNOs are the owners of the access networks, thus they

do not have to pay for transit costs or leasing of servers. China Unicom has published its network cloudification plan in the 5G era [5] and proposed the cloud-edge-fog collaborative content-oriented CDN architecture called CUBE-CDN [6].

Compared with traditional CDNs, MNOs' CDN (called telco CDNs in the rest of the paper) in the 5G era faces two main forms of challenge.

First, telco CDN should be operated in cooperation with the underlying network which evolves towards a content-oriented and information-centric architecture [7,8]. This is because telco is the owner of both the overlay and the underlay, whose performance should be taken into account simultaneously. In the context of underlay architecture evolution, Information-centric networking (ICN) has attracted much attention. In ICN, every router is cache-enabled and protocols are designed to be in content-centric abstraction in contrast to today's host-centric abstraction. Communication is based on content look-up and content caching on the path of content delivery (i.e., on-path caching), enabling content-awareness at the network layer. Once a content is cached in a router, the requests of nearby users for this content can all be satisfied by this router directly without being directed to remote servers, thus the response latency is reduced. However, the cache space in routers is limited in comparison to the powerful caching capability of CDNs. In addition, routers that execute content look-up locally can hardly support line-speed forwarding [9]. On the other hand, CDNs accelerate content delivery by directing content requests to the most suitable CDN site, which is always outside (or at least some distance) from the network that the content requests come from (i.e., off-path caching). Telco CDNs enable edge caching in the access network by deploying a large number of edge sites consisting of caching servers with a high capacity. In this context, we argue that the benefits of both ICN and telco CDN can be utilized in an integrated design fashion, which has rarely been discussed before.

Second, the distributed and cloudified infrastructure brings both flexibility and challenges in key CDN operational aspects. Existing works have focused on some of the challenges. In [10,11], they raise the concept of CDN-as-a-service (CDNaaS), which enables on-demand customization of various CDN types with different service level agreements (SLAs). On the basis of the virtualized CDNaaS architecture, important resource allocation issues are studied, e.g., QoE-aware virtual computation/storage allocation [12–14], SLA-based virtual CDN (vCDN) embedding [15,16], on-demand vCDN migration [17,18]. In addition, content caching and user request mapping should be re-designed in the hierarchical and resource-constrained cloud architecture with traffic engineering purposes taken into account [19–28]. Besides the aspects mentioned above, efficient pricing strategies and precise profitability analysis are also essential for telco CDNs to ensure sustainable income, i.e., matching revenue to cost. The profitability of introducing CDN service in an MNO is studied in [29]. The economic impact of telco CDNs' federation is studied in [30]. In addition, pricing for cache has attracted much attention [31–34].

Despite this, there is still no literature that focuses on bandwidth pricing in the context of telco CDN. Traditional CDNs adopt a unified percentile-based bandwidth pricing strategy among different CDN sites. Take Ali CDN as an example, the bandwidth pricing strategies across different CDN sites are unified on the Chinese mainland [35]. Nevertheless, the traditional bandwidth pricing strategy is not flexible and efficient enough for telco CDNs in the 5G era. The reasons are as follows. First, sites of the 5G era telco CDNs are deployed in access networks, thus a considerable amount of user requests can be satisfied within the access networks. Nevertheless, it has been shown that traffic in different access networks reveals spatial and temporal heterogeneity, which should be taken into account in the design of pricing strategies to reduce the peak-to-average ratio of the sites' bandwidth usage [36]. Second, telcos have to manage their own bandwidth cost while ensuring service quality. When the average congestion level of CDN sites reaches a certain level, telcos have to invest more to expand the sites' bandwidth capacity. Under this circumstance, a fixed pricing strategy does not provide enough economic incentives for capacity expanding.

To address the above mentioned challenges, in this paper, we first propose a framework that integrates information-centric forwarding with a powerful caching service provided by telco CDNs,

enabling more cost-efficient ICN-based telco CDNs. Then, we design a location-dependent pricing strategy (LDP) that considers the spatially heterogeneous features of traffic in geo-distributed telco CDN sites. If the variation of the sites' congestion level reaches a threshold, then the prices of the sites are set differently according to their relative congestion levels; whereas, if all sites are at a similar congestion level, then the prices are set the same according to the average congestion level. In addition, our design ensures that in both cases, the higher a site's congestion level is, the higher its bandwidth price is. This design of joint location dependent pricing and request mapping helps telco CDNs manage sites' bandwidth more efficiently and to better match revenue to cost. Specifically, the location-dependent design encourages more usage in under-utilized sites through setting lower prices. In contrast, for congested sites, it sets higher prices to let users pay more, thus its increasing bandwidth cost can be compensated.

In particular, we make four contributions.

- First, we propose an integrated framework for cost-efficient and effective deployment of ICN-based telco CDNs. Specifically, we propose to employ powerful off-path caching provided by telco CDNs to complement on-path caching in ICN. In our design, once a cache miss occurs at an ICN router, the content request is directed to the most suitable telco CDN site by the border router of the access network. Therefore, the integrated design has the features of both content-centric forwarding and powerful off-path caching, giving a cost-efficient and realistic choice for the deployment of ICN-based telco CDNs in the 5G era;
- Second, we propose a framework that consists of a location-dependent pricing (LDP) module, price-aware request mapping (PARM) module and status measurement (SM) modules. The LDP module dynamically sets prices of different CDN sites according to the sites' bandwidth usage measured by the SM modules. The PARM module calculates the request mapping rules according to the prices and the application access statistics measured by the SM modules;
- Third, we carefully design the LDP module and the PARM module. Specifically, an algorithm is presented to describe the LDP procedure, considering the spatial and temporal heterogeneity of different telco CDN sites. Then, an optimization problem that minimizes user perceived latency and the total payment is formulated to make the PARM decisions;
- Fourth, we conduct extensive simulations to evaluate the proposed design. The simulation results show that our design helps ICN-based telco CDNs flexibly price different CDN sites according to their congestion levels. When the congestion level of a CDN site increases to a certain level, LDP sets the price higher, helping the telco keep pace with its increasing bandwidth cost. Moreover, we observe the impact of some key parameter settings on the telco's revenue and cost, helping a telco to set proper parameters when adopting our design.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we illustrate the integrated design of the ICN-based telco CDN architecture. Then, we demonstrate the framework that integrates the location dependent pricing and the price-aware request mapping. In Section 4, we present the details of the LDP algorithm design followed by the problem formulation of the optimization running in the PARM module. Then, in Section 5, we conduct extensive simulations, analyze the results in different traffic patterns and different parameter settings. In Section 6, we discuss parameter setting issues in an implementation point of view. Finally, we conclude the paper in Section 7.

2. Related Work

As one of the key algorithmic functionalities in CDNs, request mapping/load balancing has always been a hot topic for its direct impact on the end users' QoE and the CDN providers' costs [37]. Specifically, different types of CDNs have different design goals in terms of request mapping.

In the context of traditional CDNs, most CDN sites are deployed in the data centers that are owned by CDN companies and the data centers are always in core networks. Therefore, a key principle

in request mapping design is to take users' location into consideration. Researchers from Bell Labs propose to use user location and content type as match fields of the routing table in the request mapping server [38]. Stability of the designed mechanism is guaranteed by the Lyapunov optimization technique while the transmission cost is minimized. Different from current request mapping mechanisms that are usually based on DNS/HTTP redirection, the proposed mechanism provides more flexibility and finer mapping granularity, which help reduce average delivery time and promote cache hit ratio. Similarly, researchers from Akamai also propose to take user location into account and realize their design in commercial deployment [39]. They make it possible to specify the client's IP prefix by using an EDNS0 extension of the existing DNS protocol. Specifically, the IP prefix of a user is included when a DNS look-up request is forwarded from a local DNS (LDNS) to an up-level authoritative DNS server. The EDNS0-based request mapping design significantly reduces round-trip-time (RTT), the time to first byte, and content downloading time.

In the context of cloud CDNs [40], CDN sites are deployed in cloud data centers which are highly distributed and always have to transit a number of Internet Service Providers (ISPs). Therefore, a key principle of request mapping is to optimize the cost efficiency. For example, the authors of [41] argue that geographical diversity of bandwidth cost brings a highly challenging problem of how to minimize total cost. In this context, bandwidth cost means the bandwidth prices that are paid by a cloud CDN provider to different transit ISPs. They formulate a problem that minimizes the summation of bandwidth cost and energy cost. In addition, they propose a hybrid particle swarm algorithm (HPSO) to solve the problem. Simulations show that the proposed design can minimize the total cost and promote throughput of the system. The authors of [42] also take into account the bandwidth prices charged by transit ISPs when solving request mapping and routing problems. In order to reduce complexity and promote robustness in the implementation, they design a distributed tide algorithm and evaluate the effectiveness. Simulations show that the proposed design can guarantee minimized average latency and average cost.

In the context of telco CDN (ISP-operated CDN), request mapping is often designed in collaboration with congestion management of the underlying network infrastructure. The reason is that a telco CDN has to optimize both the service performance that its overlay (CDN) provides and the efficiency of its underlying network infrastructure. The authors of [25] argue that in Mobile CDNs that enhance Base Stations with storage capability, blind redirection of user requests upon content placement can cause traffic congestion. They investigate a joint optimization problem of content placement and request mapping to achieve both congestion avoidance and load balancing. By using the stochastic optimization model and Lyapunov optimization technique, they design an algorithm that efficiently achieves a low transmission cost. An evaluation confirms that the design goals were achieved. The authors of [43] propose a flexible content-based and network-status aware request mapping mechanism, taking the content prefix and traffic statistics as input parameters. Simulations show that average traffic volume and average access latency are both reduced.

3. System Model

In this section, we propose the overall design of the ICN-based telco CDN architecture. Then, we present the framework that consists of an LDP module, PARM module, and SM modules. Finally, we illustrate the functions and relationship between the listed modules.

Figure 1 shows the ICN-based telco CDN service scenario we consider in this paper in comparison with the traditional CDN service scenario. Traditional CDNs, e.g., Alibaba CDN and Tencent CDN, always deploy their sites in core networks or connect their sites with telcos on Point-of-Presence (PoP) [22] and serve traditional IP access networks. In comparison, telcos in the 5G era can deploy amounts of sites in their cloudified hierarchical access networks. We propose to use these sites to provide a caching service for both ICN-based and IP-based underlay network regions. In IP network regions, content requests are directed to CDN sites after DNS look-up and data flows are forwarded based on IP address. In ICN network regions, content requests are looked up based on content name at

every router; if the requested content is cached locally, then the content chunks are directly sent back. Once a cache miss occurs, the request is forwarded to the border router, which is employed to direct the unsatisfied requests to the CDN sites. Take ICN-based region 1 as an example, if a cache miss for content A occurs at R1, border router R2 will direct the request to CDN site 1 or site 2 according to the forwarding information table (FIB) rules (as shown by R2's FIB in Figure 1). In this paper, we assume there are M telco CDN sites, serving L user access regions. The regions may be partitioned in base station level or aggregation-level. Let C_m denote the bandwidth capacity of site m , and d_m denote the historical traffic usage of site m ($m \in M$). Let t_{lm} denote the latency between region l ($l \in L$) and site m ($m \in M$). Assume that N applications adopt the telco CDN service, paying their bandwidth and cache expenses to the telco. Let λ_{nl} denote the request arrival rate of application n ($n \in N$) in region l ($l \in L$). Let t_n denote the expected response time of application n ($n \in N$).

Figure 1. Telco content delivery network (CDN) service scenario in comparison with traditional CDN scenario.

Given the input parameters above, the telco should set the bandwidth price and calculate the rules of user request mapping. Let p_m denote the unit bandwidth price of site m ($m \in M$). Let α_{nlm} denote the portion of application n requests from region l directed to site m . For clarity, we summarize the notations mentioned above in Table 1.

In our design, the bandwidth prices and the request mapping rules are periodically calculated by the LDP module and PARM module, respectively. As shown in Figure 2, the SM module measures the bandwidth usage, the application access statistics at different sites, and updates the latest state to the LDP module and the PARM module. The LDP module calculates bandwidth prices according to the congestion level and its variation. Then, on the basis of the prices and other system parameters as listed in Table 1, the PARM module runs an optimization to obtain the latest request mapping rules and install the rules on the relevant sites. Given T_{LDP} , T_{PARM} , T_{SM} as the update period of LDP, PARM, SM, respectively, then they should be set according to the following inequality:

$$T_{LDP} > T_{PARM} \geq T_{SM}. \quad (1)$$

Figure 2. The relationship between the modules in our design.

Table 1. Summary of the parameters and decision variables.

Parameters or Variables	Meaning
M	Number of telco CDN sites
L	Number of user access regions
N	Set of applications that adopt the telco CDN service
C_m	Bandwidth capacity of site m
d_m	Historical traffic usage of site m
t_{lm}	Latency between region l and site m
t_n	Expected response time of application n
λ_{nl}	Request arrival rate of application n in region l
θ_m	Congestion level of site m
Θ_m	Congestion level set of all CDN sites
$\hat{\theta}$	Reference congestion level to evaluate relative congestion level of a certain site
σ_1, σ_2	Congestion level thresholds in Algorithm 1
P_{min}	The lowest price of a site in LDP design
P_{max}	The highest price of a site in LDP design
P_{const}	Fixed price of a site in traditional pricing strategy
p_m	Unit bandwidth price of site m
α_{nlm}	Portion of application n requests from region l directed to site m

4. Joint Pricing and Request Mapping

On the basis of the illustration of the system model above, in this section, we present the design details of the LDP module and PARM module, respectively.

4.1. The Design of LDP Module

The key design principle of the LDP module is to set the price of a CDN site according to the site's congestion level. Specifically, the principle includes two features: (a) the higher the congestion level is, the higher the price is; (b) the pricing strategy should have a location-dependent feature. Here, we define θ_m as the historical congestion level of site m , as Equation (3) shows. We use Θ to represent the congestion level set of all CDN sites.

$$\theta_m = \frac{d_m}{C_m}, m \in M \quad (2)$$

$$\Theta = [\theta_1, \theta_2, \dots, \theta_M]. \quad (3)$$

According to principle (a), we first design a congestion-aware bandwidth pricing strategy (CABP). The bandwidth cost function for network provider is always modeled as no-decrease and a convex function of congestion level, for simplicity, researchers always use the piecewise linear form [44,45]. Considering the objective of pricing is to match the bandwidth cost, we believe that it is reasonable to design a piecewise linear pricing function in CABP. Specifically, by defining two thresholds, i.e., a lower-bound threshold and an upper-bound threshold. If the congestion level of a site is lower than the lower-bound threshold, the price is set as the lowest price. If the congestion level of a site is higher than the upper-bound threshold, the price is set as the highest price. If the congestion level of a site is between two thresholds, the price is set according to a linear function of the congestion level. Given

θ_m , we calculate p_m by comparing θ_m with a reference value $\hat{\theta}$. Define two congestion level thresholds as σ_1 and σ_2 , respectively. Define two price thresholds as P_{min} and P_{max} , respectively. Let P_{const} denote the fixed bandwidth price in a traditional CDN, then $P_{min} \leq P_{const} \leq P_{max}$ holds. The impact of setting P_{min} and P_{max} will be discussed in Section 5. Then, p_m is calculated as described in Algorithm 1. If θ_m is smaller than $\sigma_1 \hat{\theta}$, p_m is set as P_{min} (line 2). If θ_m is between $\sigma_1 \hat{\theta}$ and $\sigma_2 \hat{\theta}$, p_m is set according to a linear function with a positive slope of $\frac{P_{max}-P_{min}}{\sigma_2 \hat{\theta}-\sigma_1 \hat{\theta}}$, i.e., p_m increases with θ_m (line 4). If θ_m is bigger than $\sigma_2 \hat{\theta}$, p_m is set as P_{max} . Figure 3 shows the relationship between congestion level and bandwidth price.

Figure 3. The relationship between congestion level and bandwidth price.

Algorithm 1 The congestion-aware bandwidth pricing strategy (CABP) algorithm

Input: $\theta_m, \hat{\theta}, \sigma_1, \sigma_2, P_{min}, P_{max}$

Output: p_m

- 1: **if** $\theta_m < \sigma_1 \hat{\theta}$ **then**
 - 2: $p_m = P_{min}$; // Set the price of site m as the lowest price
 - 3: **else if** $\sigma_1 \hat{\theta} \leq \theta_m < \sigma_2 \hat{\theta}$ **then**
 - 4: $p_m = P_{min} + \frac{P_{max}-P_{min}}{\sigma_2 \hat{\theta}-\sigma_1 \hat{\theta}} (\theta_m - \sigma_1 \hat{\theta})$; // Set the price of site m according to its congestion level based on linear part of the piecewise function in Figure 3
 - 5: **else if** $\theta_m \geq \sigma_2 \hat{\theta}$ **then**
 - 6: $p_m = P_{max}$; // Set the price of site m as the highest price
 - 7: **end if**
 - 8: **return** p_m ;
-

On the basis of the CABP strategy, the procedure of LDP is described in Algorithm 2 based on principle (b). Note that the temporal and spatial heterogeneity of traffic patterns of different sites should be taken into account to reduce network congestion [36]. Therefore, the LDP algorithm takes the variation of the sites' congestion level as a key factor that influences the pricing procedure. At the beginning of every price update period (T_{LDP}), the LDP module first checks the variation of Θ to determine whether or not to set the prices to be location-dependent. If the variation is smaller than δ (a pre-defined parameter, in this work we set as $\delta = 0.03$), p_m ($\forall m \in M$) are set to the same value, calculated by running the CABP function with mean value of Θ as congestion level and 1 as reference congestion level (line 5). If the variation is larger than δ , the prices are set to be location-dependent.

We fuse Θ based on the population weighted average method to obtain the reference congestion level $\hat{\theta}$ (line 7). Note that given U_m as the user number of site m , the population weight of site m is defined as γ . Then, by running the function CABP with θ_{avg} and $\hat{\theta}$ as the input parameters, p_m is obtained (line 8). σ_{11}, σ_{12} and σ_{21}, σ_{22} are set as σ_1, σ_2 , respectively, in the above two cases.

Algorithm 2 The LDP Algorithm

Input: $\theta_m, \forall m \in M$

Output: P

```

1: for Every beginning of  $T_{LDP}$  do
2:   if  $var(\Theta) < \delta$  then
3:      $\theta_{avg} = mean(\Theta)$ ; // Obtain the average value of the sites' congestion level
4:      $\hat{\theta} = 1$ ; // Set 1 as the reference congestion level
5:      $p_m = CABP(\theta_{avg}, \hat{\theta}, \sigma_{11}, \sigma_{12}), \forall m \in M$ ; // Calculate price for every site with  $\theta_{avg}$  as congestion
        level in CABP
6:   else
7:      $\hat{\theta} = \sum_{m=1}^M \gamma_m \theta_m (\gamma_m = \frac{U_m}{\sum_{i=1}^M U_i})$ ; // Set the data fusion result obtained from population-weighted
        average method as the reference congestion level
8:      $p_m = CABP(\theta_m, \hat{\theta}, \sigma_{21}, \sigma_{22}), \forall m \in M$ ; // Calculate price for site  $m$  with  $\theta_m$  as congestion level in
        CABP
9:   end if
10: end for
11: return  $P = [p_1, p_2, \dots, p_M]$ ;

```

4.2. The Design of PARM Module

Taking the sites' bandwidth prices P calculated by LDP and the other parameters in Table 1 as input, in this subsection, we formulate the optimization running in the PARM module.

To guarantee the service quality and provide economic incentives to application providers, the ICN-based telco CDN should minimize both the latency and the application providers' bandwidth payment. Therefore, the objective is set as minimizing the weighted sum of latency and total payment of all the application providers, as Equation (4a) shows. Note that Ψ is the weight factor. In practice, Ψ is set according to a CDN's actual operational objectives. For example, if the CDN focus more on guaranteeing the latency performance, Ψ should be set as a relatively small value (or even as zero). Or if the CDN wants to give users more economic incentives, Ψ should be set as a big value. In summary, the determination of the weight parameter is an engineering issue, which is out of the scope of this work. Constraint (4b) ensures that the aggregate traffic served by site m is within the site's capacity. Furthermore, we use $\bar{\zeta}$ to ensure that a site's target congestion level is within a certain range of the mean value of Θ , denoted by $\bar{\theta}$. Thus, the case that too many requests are directed to possible site with the smallest latency/price will be prevented. Constraint (4c) ensures that average latency perceived by application n 's requests is within its expected response time. Constraint (4d) ensures that the proportions that the application n 's requests generated in region l directed to different sites sum up to one. Constraint (4e) enforces the ranges of the variables, i.e., α_{nlm} .

$$\min \sum_n \sum_l \sum_m \lambda_{nl} \alpha_{nlm} t_{lm} + \Psi \cdot \sum_n \sum_l \sum_m \lambda_{nl} \alpha_{nlm} p_m \quad (4a)$$

$$s.t. \left| \frac{\sum_n \sum_l \lambda_{nl} \alpha_{nlm}}{C_m} - \bar{\theta} \right| \leq \xi, \forall m \in M \quad (4b)$$

$$\frac{\sum_m \sum_l \alpha_{nlm} t_{lm}}{L} \leq t_n, \forall n \in N \quad (4c)$$

$$\sum_m \alpha_{nlm} = 1, \forall n \in N, \forall l \in L \quad (4d)$$

$$0 \leq \alpha_{nlm} \leq 1, \forall n \in N, \forall l \in L, \forall m \in M \quad (4e)$$

$$var : \alpha_{nlm}, \forall n \in N, \forall l \in L, \forall m \in M. \quad (4f)$$

Note that Constraint (4b) is equivalent to the following two inequalities:

$$\frac{\sum_n \sum_l \lambda_{nl} \alpha_{nlm}}{C_m} - \bar{\theta} \geq -\xi, \forall m \in M \quad (5a)$$

$$\frac{\sum_n \sum_l \lambda_{nl} \alpha_{nlm}}{C_m} - \bar{\theta} \leq \xi, \forall m \in M. \quad (5b)$$

Therefore, the optimization is a linear programming (LP) problem, which can be solved by existing LP solvers (e.g., linprog function in Matlab).

5. Evaluation

In this section, we first introduce the experiment settings, then we analyze the experiment results to evaluate the effectiveness of our design.

5.1. Experiment Setup

5.1.1. Traffic Model

In simulation, we consider there to be 10 user regions and 10 CDN sites. End users from different regions consume traffic for three types of applications. To evaluate the proposed joint design under traffic that matches the trend of the practical traffic, we adopt the traffic generation approach in [45] to generate the traffic demand. Specifically, we use a Poisson process to simulate the process of end users' Internet access for different applications. User requests for application n ($1 \leq n \leq 3$) arrive according to a Poisson process with λ_{nu} as intensity, the value of λ_{nu} is set according to the parameters in [45]. To evaluate the proposed mechanism in different traffic patterns, user requests from different access regions are generated according to Table 2, which gives the user number in different access regions. λ_{nl} is obtained according to Equation (6). According to the measurement of average latency of Internet access in China in [42], we set the latency parameters as follows. For site m , t_{lm} ($m \in M$) is generated randomly according to a normal distribution with t_m as mean and 1 ms as standard deviation. t_m is generated randomly according to a continuous uniform distributions with 40 ms as the lower bound and 100 ms as the upper bound. C_m is set as 1.1 times the max bandwidth usage in Traffic Pattern 1. For the price-related parameters, we set P_{const} according to the bandwidth unit price in Ali CDN [35], set P_{min} as 0.2, P_{max} as 0.5 as default. According to the trend of traffic fluctuation throughout a day, 1 h is an appropriate length to distinguish the varied congestion level of a site [36,46]. Therefore, the update period of LDP is set as 1 h. SM, PARM are both set as 10 min. Here, the relationship between periods of SM, LDP, and PARM conforms with Equation (1). The simulation time is set as 1 h. In summary, all parameters are shown in Table 3.

$$\lambda_{nl} = \sum_{u \in l} \lambda_{nu}, \forall n \in N, \forall l \in L. \quad (6)$$

Table 2. Traffic patterns in the experiment.

Index	User Numbers in Every Region
Traffic Pattern 1	[110, 150, 90, 230, 320, 190, 430, 380, 450, 500]
Traffic Pattern 2	[90, 255, 375, 160, 80, 150, 310, 75, 150, 175]
Traffic Pattern 3	[230, 210, 190, 240, 220, 200, 205, 225, 245, 215]
Traffic Pattern 4	[300, 320, 305, 310, 325, 330, 315, 290, 305, 325]

Table 3. Parameters' setting in the experiment.

Parameter	Value
M	10
L	10
N	3
t_n	[90 ms, 70 ms, 60 ms]
λ_{nu}	[61.92, 39.84, 22.08]
P_{const}	0.3
P_{min}	0.2
P_{max}	0.5
Ψ	0.5
σ_1, σ_2	$\sigma_{11} = 0.3, \sigma_{12} = 0.7, \sigma_{21} = 0.4, \sigma_{22} = 1.2$

5.1.2. Performance Metrics and the Compared Cases

We first observe the relationship between the CDN sites' congestion level and their prices under the LDP design. Then, we evaluate the performance of PARM by observing the following metrics:

- The variation of the sites' congestion level;
- Bandwidth cost faced by the telco CDN. The bandwidth cost of a site m is calculated according to Equation (7) [47], where D_m is the actual traffic demand and C_m is the bandwidth capacity of the site;
- Total bandwidth payment of the application providers;
- Average latency perceived by end users.

$$g(D_m) = \begin{cases} D_m & 0 \leq D_m/C_m < 1/3 \\ 3D_m - 2/3C_m & 1/3 \leq D_m/C_m < 2/3 \\ 10D_m - 16/3C_m & 2/3 \leq D_m/C_m < 9/10 \\ 70D_m - 178/3C_m & 9/10 \leq D_m/C_m < 1 \\ 500D_m - 1468/3C_m & 1 \leq D_m/C_m < 11/10 \\ 5000D_m - 16318/3C_m & 11/10 \leq D_m/C_m < \infty. \end{cases} \quad (7)$$

The metrics above are observed in the three cases, which are compared in Table 4. In the first case, bandwidth usage is charged with a fixed price, and user requests are directed to the nearest CDN site. In the second case, bandwidth usage is charged with a fixed price, and request mapping rules are calculated based on optimization that only minimize the latency perceived by all user requests (i.e., the optimization is price-unaware, set $Psi = 0$ in the PARM formulation). In the third case, bandwidth usage is charged with our LDP design, and request mapping rules are calculated by the PARM module. The above three are called **BASE**, **PURM**, and **PARM**, respectively. In Section 5.2.1, the congestion level in BASE is set as the historical congestion level. In Sections 5.2.2 and 5.2.3, the congestion level in BASE is set as the initial historical congestion level and the congestion level of the previous PARM period is always set as the historical congestion level of the upcoming PARM period.

Table 4. Details of the compared cases.

Case Name	The Bandwidth Pricing Strategy	The Request Mapping Approach
BASE	Fixed pricing as P_{const}	Direct Requests to the Nearest site
PURM	Fixed pricing as P_{const}	PARM with $\Psi = 0$
PARM	Location dependent pricing as p_m	PARM with $\Psi > 0$

In addition, in Section 5.2.3, we observe the impact of ξ and P_{max} on performance metrics of PARM. (In Section 5.2.2, ξ is set as $0.15\hat{\theta}$ and P_{max} is set as 0.5)

5.2. Experiment Results

5.2.1. Results of LDP

Using the sites' congestion level under the BASE case as historical data, we first evaluate the performance of LDP. The relationship of the sites' congestion level and bandwidth price of four traffic patterns is shown in Figures 4–7, respectively. In general, variation of the sites' congestion level of Traffic Pattern 1 and 2 is obviously bigger than that of Traffic Pattern 3 and 4. Prices in the former two traffic patterns are set differently across all sites. In Figure 4, sites 1–6 is not as congested as sites 7–10, thus have a lower price. The prices of sites 7–10 are higher than those of sites 1–6. Note that among all the sites, the site with a higher congestion level has a higher price, conforming with the pricing function shown in Figure 3. In Figure 5, the prices are also location-dependent. Remember that in Algorithm 2, when prices are set differently in different locations, a site's congestion level is compared with the weighted sum of congestion level as reference value. Site 3 and site 7, the most congested sites in Traffic Pattern 2, contribute much to the weighted sum. Therefore, they are priced the same as the most congested sites in Figure 4. In Traffic Pattern 3 as shown in Figure 6, the traffic is distributed smoothly across different sites, the congestion level of every site is around 0.4. Thus, the prices are set as the same value based on the mean value of the congestion level. In Traffic Pattern 4, as shown in Figure 7, the sites' congestion level fluctuates around 0.6 within a very small range. Thus, the prices are set as the same value that is higher than that in Figure 6. The reason is that Traffic Pattern 4 has a higher mean value of the sites' congestion level.

Figure 4. LDP Prices in Traffic Pattern 1.

Figure 5. LDP prices in Traffic Pattern 2.

Figure 6. LDP prices in Traffic Pattern 3.

Figure 7. LDP prices in Traffic Pattern 4.

5.2.2. Results of PARM

On the basis of the prices calculated in the previous part, we further evaluate the performance of the PARM design, compared with another two cases, i.e., BASE and PURM.

Figure 8 shows the comparison of variation of CDN sites in four traffic patterns of BASE, PURM, PARM, respectively. We observe that the variation of the former two patterns in BASE is obviously higher than in the latter two patterns. The observation here conforms with the distributions of the sites' congestion level observed in Figures 4–7, respectively. In PURM and PARM, the variations of all traffic patterns are nearly the same. This is because Constraint (4b) (which exists in both PURM and PARM) bounds the target congestion level. For details, variations in traffic pattern 1 and 2 under PURM and PARM are much lower than the relative values under BASE. Variations in the latter two traffic patterns are higher than those in BASE.

Figure 8. Comparison of variation of CDN sites.

To quantify the impact of three cases on the telco CDN's cost, we compare total bandwidth cost of all CDN sites in Figure 9. First we observe that bandwidth costs under PURM and PARM are nearly the same in all traffic patterns and are not higher than those under BASE. This observation proves the effectiveness of Constraint (4b) in reducing the total bandwidth cost by making sites' target congestion level fluctuate in certain range. Specifically, PURM and PARM obviously reduce bandwidth cost compared with BASE in former two traffic patterns. In comparison, three cases bring nearly the same bandwidth cost in traffic pattern 3 and 4. That is to say, the variation increase brought by PURM and PARM in the latter two traffic patterns (as shown in Figure 8) is acceptable.

Figure 10 shows the total payment for bandwidth in three cases. First, we observe that payments under BASE and PURM are the same. This is because the bandwidth prices are the same in these two cases, i.e., P_{const} , and the total bandwidth usage is unchanged after request direction by PURM or PARM (as constrained by Constraint (4d)). In comparison, payment under PARM is different from that under BASE/PURM due to the location-dependent and congestion-aware features in the LDP strategy. We observe that in Traffic Pattern 2 and 3, which are less congested, payment under PARM is lower than that in BASE/PURM with 7.94% and 10.4% as the reduction ratios, respectively. This is because LDP sets the price lower than P_{const} when the congestion level is relatively low. Thus, the application providers will save more money if they use more. In contrast, in Traffic Pattern 1 and 4, which are more congested, payment under PARM is higher than that in BASE/PURM with 3.25% and 31.7% as the increasing ratios, respectively. This is because LDP sets the price higher than P_{const} when the congestion level is relatively high. The higher the congestion level, the higher the price. That is to say,

when the overall congestion level of all the CDN sites is high, a telco CDN could be paid more under PARM than under BASE/PURM, thus it will have more economic incentives and be positive in terms of expanding its bandwidth capacity.

Figure 9. Comparison of total Bandwidth Cost.

Figure 10. Comparison of total bandwidth payment.

Figure 11 shows the average latency of all applications in the three cases. First, we observe that PURM and PARM both outperform BASE. This is because they both minimize latency in the optimization and BASE conducts request mapping without considering the latency performance. Then, we compare PURM and PARM and find that in Traffic Pattern 1 and 3, PARM outperforms PURM, and in the other two traffic patterns, PURM outperforms PARM. Yet, it is confirmed that the latency constraint of every application is satisfied. Thus, the result is acceptable.

Figure 11. Comparison of application's average latency.

From the above analysis, three conclusions could be obtained as follows:

- For the traffic patterns with high variations, our design is efficient to reduce the variation and thus reduce the telco CDN's total bandwidth cost;
- For any traffic pattern, our design helps a telco CDN better match its revenue to cost. Specifically, if the overall congestion level is relatively low, a discount is provided (compared with P_{const}) to encourage more usage. Otherwise, prices of the congested sites are set higher than P_{const} , making the total revenue match with the increasing bandwidth cost;
- The latency performance provided by PARM is similar to that provided by PURM, which minimizes latency only.

5.2.3. Impact of ξ and P_{max} on Performance of PARM

From the analysis in Section 5.2.2, we find that in our design, PARM is effective in reducing the variation of the sites' congestion level, LDP is effective in helping telco CDN match revenue to cost by setting prices according to congestion level. In this part, we further explore the impact of key parameters (i.e., ξ , P_{max}) on the effectiveness of PARM.

Impact of ξ . From Figure 8 we have already observed that ξ is effective in bounding the variation. Here we observe the trend of different performance metrics when ξ varies. Define ξ as τ percent of $\bar{\theta}$, as shown by Equation (8). Figures 12–15 show the metric when τ vary from 0.1 to 0.3.

$$\xi = \tau \cdot \bar{\theta} \quad (8)$$

Figure 12 shows the trend of variation of the sites' congestion level. We observe that as τ increases, the variation of four traffic patterns all increase. This is because a larger τ enables a larger range for congestion level fluctuation among all sites when conducting request mapping. When $\tau = 0.1$, the variations are nearly the same. When τ increases, the variation of the four values gets bigger and bigger. In addition, for any τ , the relationship between the variations in Traffic Pattern 4, Traffic Pattern 1, Traffic Pattern 3, Traffic Pattern 2 is consistent, i.e., from the largest to the smallest. This order conforms with the order of the overall congestion level of four traffic patterns. (Traffic Pattern 4 is the most congested and with the largest total bandwidth cost, Traffic Pattern 2 is the least congested and with the least total bandwidth cost, as shown in Figure 9) The larger the overall congestion level is, the larger the possible range for request mapping is, thus the numerical relationship holds.

Figure 12. Impact of τ on sites' variation in PARM.

Figure 13 shows the trend of total bandwidth cost. First we observe that as τ increases, the relationship between the total bandwidth cost in Traffic Pattern 4, Traffic Pattern 1, Traffic Pattern 3, Traffic Pattern 2 is consistent, i.e., from the largest to the smallest, conforming with the analysis in the previous paragraph. Specifically, when $\tau \geq 0.15$, total bandwidth cost of Traffic Pattern 2 and Traffic Pattern 3 gradually increase. When $\tau = 0.3$, these two values reach 1.139 times and 1.123 times of the values when $\tau = 0.15$, respectively. For Traffic Pattern 1, the value nearly stay the same as τ varies. For Traffic Pattern 4, the value stay nearly the same when τ varies from 0.1 to 0.175, then it increase much at every step. When $\tau = 0.3$, the value reaches 1.202 times of the value when $\tau = 0.175$.

Figure 13. Impact of τ on total bandwidth cost in PARM.

Figure 14 shows the trend of total payment. First we observe that the value in Traffic Pattern 3 and Traffic Pattern 4 does not change as τ increases. This is because in these two patterns, different CDN sites are with the same price, as shown in Figures 6 and 7, respectively. In comparison, the values in Traffic Pattern 1 and Traffic Pattern 2 both decrease as τ increases. This is because in these two patterns, different CDN sites are with different prices. A larger τ provides much more space for PARM to map requests to sites that are with lower prices. Note that the overall congestion level of Traffic Pattern 1 is higher than that of Traffic Pattern 2, thus the range of different prices in the former pattern

is larger than that in the latter pattern. Therefore, the absolute value of gradient of the former pattern is bigger than that of the latter pattern.

Figure 15 shows the trend of average latency. We observe that this metric is independent of τ . The average latency curves of Traffic Pattern 3 and Traffic Pattern 4 are more flat than those of the other two traffic patterns.

Figure 14. Impact of τ on total payment in PARM.

Figure 15. Impact of τ on average latency in PARM.

Impact of P_{max} . In our design, the differentiated prices imply the difference in the sites' congestion level and affect the procedure of request mapping. Thus, it is essential to evaluate the impact of price range on the performance of PARM. For simplicity, we make P_{min} fixed as 0.2 and make P_{max} vary from 0.35 to 0.8. The trend of four observed metrics are shown in Figures 16–19, respectively.

In Figure 16, we observe that four curves are stable around certain values, whose numerical relationship conforms with that in Figure 12. This is because given a fixed ζ , the various price only influence the optimized value of Equation (4a), not the variables. On the basis of the stability in Figure 16, it is not surprising that the trend of total bandwidth cost is also stable, as shown in Figure 17.

Figure 16. Impact of P_{max} on sites' variation in PARM.

In Figure 18, we observe that the rank of four traffic patterns at every P_{max} conforms with the respective rank in Figures 16 and 17, i.e., the higher the overall congestion level of a traffic pattern, the higher the total payment. In addition, four curves all increase but with different gradients, specifically, the higher the overall congestion level of a traffic pattern, the bigger the gradient. Note that the curves of Traffic Pattern 3 and Traffic Pattern 4 are both linear. Thus, for simplicity, we take these two patterns as examples to analyze the observation above. First, the sites' prices are location-independent (all sites have the same price) in these two patterns, as shown in Figures 6 and 7. Therefore, the total payment could be formulated as Equation (9), where P_{fixed} is the bandwidth price in Traffic Pattern 3 or 4. d'_m is the traffic demand of site m after request mapping under PARM. Note that as P_{max} increases, the slope of CABP also increases. Under the same pricing function, P_{fixed} in Traffic Pattern 4 increases faster because it is calculated with a bigger variable (the average congestion level) in the linear function. Therefore, the curve of Traffic Pattern 4 has a bigger gradient.

Figure 17. Impact of P_{max} on total bandwidth cost in PARM.

Finally, it is easy to understand the curves in Figure 19, considering that price variables do not have direct relationship with latency optimization.

$$BandwidthPayment_{total} = P_{fixed} \sum_{m=1}^M d'_m. \quad (9)$$

From the above analysis, three conclusions could be obtained as follows:

- ζ is effective in affecting the sites' variation and the total bandwidth cost. Note that if ζ is too big, the total bandwidth cost will grow rapidly when sites become too congested (as Traffic Pattern 4 in our simulation). Therefore, the overall congestion level should be taken into account when setting ζ ;
- P_{max} is effective in affecting the application providers' total payment. If the overall congestion level is high, a bigger P_{max} will give more economic incentives to help a telco CDN match its increasing bandwidth cost. Certainly, the tolerance of CDN users (i.e., application providers) should be taken into account when a telco CDN is making the decision of raising P_{max} ;
- Average latency does not have a direct relationship with ζ or P_{max} .

Figure 18. Impact of P_{max} on total payment in PARM.

Figure 19. Impact of P_{max} on average latency in PARM.

6. Discussions of Parameter Setting in Implementation Aspect

Parameter setting in practical use should be determined according to a CDN's actual operational purpose and its historical traffic patterns. In practical implementation, the adopter of our design could set the parameters according to our analysis. For example, if the CDN is over-congested, then τ should be set lower than 0.2. Figure 13 shows that $\tau \geq 0.2$ will lead to a significant increase in bandwidth cost in Traffic Pattern 4. In addition, if congestion continuously exists for a period of time, the CDN may raise the max price but keep the max price lower than 0.5. Figure 18 shows that $P_{max} \geq 0.5$ will lead to a significant increase in user payment that may lead to critical user complaint.

7. Conclusions

In this paper, we first propose a ICN-based telco CDN framework that integrates the benefits of both ICN (i.e., information-centric forwarding) and telco CDNs (i.e., powerful distributed edge caching) to provide high-quality content delivery service in the 5G era. Then, we design a location-dependent pricing (LDP) strategy for ICN-based telco CDNs by taking into account the spatial heterogeneity of the traffic usage of geo-distributed CDN sites. Thereafter, on the basis of the LDP strategy, we formulate a price-aware request mapping optimization which minimizes both user perceived latency and application providers' bandwidth payment. Simulations under four representative traffic patterns show that our design helps ICN-based telco CDNs efficiently manage the sites' congestion and thus better match their revenue to cost. In addition, the impacts of key parameters in LDP and PARM on the performance metrics are analyzed and setting suggestions are provided for adopters.

Author Contributions: Formal analysis, M.J.; Funding acquisition, S.G.; Methodology, M.J. and B.F.; Project administration, S.G.; Supervision, H.L.; Validation, M.J.; Writing—original draft, M.J., H.L. and B.F.

Funding: This research is funded by the project of Beijing New-Star Plan of Science and Technology under Grant No. Z151100000315052.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cisco Visual Networking Index: Forecast and Methodology 2017–2022; Cisco: San Jose, CA, USA, 2019.
2. Sousa, B.; Cordeiro, L.; Simoes, P.; Edmonds, A.; Ruiz, S.; Carella, G.A.; Corici, M.; Nikaein, N.; Gomes, A.S.; Schiller, E.; et al. Toward a fully cloudified mobile network infrastructure. *IEEE Trans. Netw. Serv. Manag.* **2016**, *13*, 547–563. [\[CrossRef\]](#)
3. Alnoman, A.; Carvalho, G.H.; Anpalagan, A.; Woungang, I. Energy efficiency on fully cloudified mobile networks: Survey, challenges, and open issues. *IEEE Commun. Surv. Tutor.* **2017**, *20*, 1271–1291. [\[CrossRef\]](#)
4. Ibrahim, A.I.A.; Khalifa, F.E.I. Cloudified NGN infrastructure based on SDN and NFV. In Proceedings of the 2015 International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE), Khartoum, Sudan, 7–9 September 2015; pp. 270–274.
5. White Paper for China Unicom's Edge-Cloud; Technical Report; China Unicom: Beijing, China, 2018.
6. White Paper for China Unicom's CUBE CDN; Technical Report; China Unicom: Beijing, China, 2018.
7. Ibrahim, G.; Chadli, Y.; Kofman, D.; Ansiaux, A. Toward a new Telco role in future content distribution services. In Proceedings of the 2012 16th International Conference on Intelligence in Next Generation Networks (ICIN), Berlin, Germany, 8–11 October 2012; pp. 22–29.
8. Spagna, S.; Liebsch, M.; Baldessari, R.; Niccolini, S.; Schmid, S.; Garroppo, R.; Ozawa, K.; Awano, J. Design principles of an operator-owned highly distributed content delivery network. *IEEE Commun. Mag.* **2013**, *51*, 132–140. [\[CrossRef\]](#)
9. Fayazbakhsh, S.K.; Lin, Y.; Tootoonchian, A.; Ghodsi, A.; Koponen, T.; Maggs, B.; Ng, K.; Sekar, V.; Shenker, S. Less pain, most of the gain: Incrementally deployable icn. *ACM SIGCOMM Comput. Commun. Rev.* **2013**, *43*, 147–158. [\[CrossRef\]](#)
10. Frangoudis, P.A.; Yala, L.; Ksentini, A. CDN-as-a-Service Provision over a Telecom Operator's Cloud. *IEEE Trans. Netw. Serv. Manag.* **2017**, *14*, 702–716. [\[CrossRef\]](#)

11. Herbaut, N.; Negru, D.; Dietrich, D.; Papadimitriou, P. Dynamic Deployment and Optimization of Virtual Content Delivery Networks. *IEEE Multimed.* **2017**, *24*, 28–37. [[CrossRef](#)]
12. Yala, L.; Frangoudis, P.A.; Ksentini, A. QoE-aware computing resource allocation for CDN-as-a-service provision. In Proceedings of the IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–6.
13. Yala, L.; Frangoudis, P.A.; Lucarelli, G.; Ksentini, A. Balancing between Cost and Availability for CDNaas Resource Placement. In Proceedings of the GLOBECOM 2017—2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–7. [[CrossRef](#)]
14. Llorca, J.; Sterle, C.; Tulino, A.M.; Choi, N.; Sforza, A.; Amideo, A.E. Joint content-resource allocation in software defined virtual CDNs. In Proceedings of the 2015 IEEE International Conference on Communication Workshop (ICCW), London, UK, 8–12 June 2015; pp. 1839–1844.
15. Herbaut, N.; Negru, D.; Magoni, D.; Frangoudis, P.A. Deploying a content delivery service function chain on an SDN-NFV operator infrastructure. In Proceedings of the 2016 International Conference on Telecommunications and Multimedia (TEMU), Heraklion, Greece, 25–27 July 2016; pp. 1–7.
16. Herbaut, N.; Negru, D.; Dietrich, D.; Papadimitriou, P. Service chain modeling and embedding for NFV-based content delivery. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–7.
17. Ibn-Khedher, H.; Hadji, M.; Abd-Elrahman, E.; Afifi, H.; Kamal, A.E. Scalable and cost efficient algorithms for virtual CDN migration. In Proceedings of the 2016 IEEE 41st Conference on Local Computer Networks (LCN), Dubai, UAE, 7–10 November 2016; pp. 112–120.
18. Ibn-Khedher, H.; Abd-Elrahman, E. Cdnas framework: Topsis as multi-criteria decision making for vcdn migration. *Procedia Comput. Sci.* **2017**, *110*, 274–281. [[CrossRef](#)]
19. De Vleeschauwer, D.; Robinson, D.C. Optimum caching strategies for a telco CDN. *Bell Labs Tech. J.* **2011**, *16*, 115–132. [[CrossRef](#)]
20. Liu, D.; Chen, B.; Yang, C.; Molisch, A.F. Caching at the wireless edge: Design aspects, challenges, and future directions. *IEEE Commun. Mag.* **2016**, *54*, 22–28. [[CrossRef](#)]
21. Sharma, A.; Venkataramani, A.; Sitaraman, R.K. Distributing content simplifies ISP traffic engineering. *ACM SIGMETRICS Perform. Eval. Rev.* **2013**, *41*, 229–242. [[CrossRef](#)]
22. Li, Z.; Simon, G. In a Telco-CDN, pushing content makes sense. *IEEE Trans. Netw. Serv. Manag.* **2013**, *10*, 300–311.
23. Claeys, M.; Tuncer, D.; Famaey, J.; Charalambides, M.; Latré, S.; Pavlou, G.; Turck, F.D. Hybrid multi-tenant cache management for virtualized ISP networks. *J. Netw. Comput. Appl.* **2016**, *68*, 28–41. [[CrossRef](#)]
24. Diab, K.; Hefeeda, M. Cooperative Active Distribution of Videos in Telco-CDNs. In Proceedings of the SIGCOMM Posters and Demos, Los Angeles, CA, USA, 22–24 August 2017; pp. 19–21.
25. Liu, J.; Yang, Q.; Simon, G. Congestion Avoidance and Load Balancing in Content Placement and Request Redirection for Mobile CDN. *IEEE/ACM Trans. Netw.* **2018**, *26*, 851–863. [[CrossRef](#)]
26. Li, X.; Wang, X.; Wan, P.J.; Han, Z.; Leung, V.C. Hierarchical Edge Caching in Device-to-Device Aided Mobile Networks: Modeling, Optimization, and Design. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 1768–1785. [[CrossRef](#)]
27. Tran, T.X.; Le, D.V.; Yue, G.; Pompili, D. Cooperative Hierarchical Caching and Request Scheduling in a Cloud Radio Access Network. *IEEE Trans. Mob. Comput.* **2018**. [[CrossRef](#)]
28. Yang, P.; Zhang, N.; Zhang, S.; Yu, L.; Zhang, J.; Shen, X. Dynamic Mobile Edge Caching with Location Differentiation. In Proceedings of the GLOBECOM 2017—2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–6.
29. Lee, D.; Mo, J.; Park, J. ISP vs. ISP+ CDN: Can ISPs in Duopoly profit by introducing CDN services? *ACM SIGMETRICS Perform. Eval. Rev.* **2012**, *40*, 46–48. [[CrossRef](#)]
30. Lee, H.; Duan, L.; Yi, Y. On the competition of CDN companies: Impact of new telco-CDNs' federation. In Proceedings of the 2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Tempe, AZ, USA, 9–13 May 2016; pp. 1–8.
31. Zhong, Y.; Xu, K.; Li, X.Y.; Su, H.; Xiao, Q. Estra: Incentivizing storage trading for edge caching in mobile content delivery. In Proceedings of the 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, 6–10 December 2015; pp. 1–6.

32. Shen, F.; Hamidouche, K.; Bastug, E.; Debbah, M. A Stackelberg game for incentive proactive caching mechanisms in wireless networks. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–6.
33. Liu, T.; Li, J.; Shu, F.; Guan, H.; Yan, S.; Jayakody, D.N.K. On the incentive mechanisms for commercial edge caching in 5G Wireless networks. *IEEE Wirel. Commun.* **2018**, *25*, 72–78. [[CrossRef](#)]
34. Krolikowski, J.; Giovanidis, A.; Di Renzo, M. A decomposition framework for optimal edge-cache leasing. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 1345–1359. [[CrossRef](#)]
35. Cloud, A. Pricing for CDN. Available online: <https://www.aliyun.com/price/?spm=5176.7933777.1280361.25.113d56f5wjp666> (accessed on 12 April 2019).
36. Li, Y.; Xu, F. Trace-driven analysis for location-dependent pricing in mobile cellular networks. *IEEE Netw.* **2016**, *30*, 40–45. [[CrossRef](#)]
37. Maggs, B.M.; Sitaraman, R.K. Algorithmic nuggets in content delivery. *ACM SIGCOMM Comput. Commun. Rev.* **2015**, *45*, 52–66. [[CrossRef](#)]
38. Benchaita, W.; Ghamri-Doudane, S.; Tixeuil, S. On the Optimization of Request Routing for Content Delivery. *ACM SIGCOMM Comput. Commun. Rev.* **2015**, *45*, 347–348. [[CrossRef](#)]
39. Chen, F.; Sitaraman, R.K.; Torres, M. End-user mapping: Next generation request routing for content delivery. *ACM SIGCOMM Comput. Commun. Rev.* **2015**, *45*, 167–181. [[CrossRef](#)]
40. Wang, M.; Jayaraman, P.P.; Ranjan, R.; Mitra, K.; Zhang, M.; Li, E.; Khan, S.; Pathan, M.; Georgeakopoulos, D. An overview of cloud based content delivery networks: research dimensions and state-of-the-art. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XX*; Springer: Berlin, Germany, 2015; pp. 131–158.
41. Yuan, H.; Bi, J.; Li, B.H.; Tan, W. Cost-aware request routing in multi-geography cloud data centres using software-defined networking. *Enterp. Inf. Syst.* **2017**, *11*, 359–388. [[CrossRef](#)]
42. Fan, Q.; Yin, H.; Jiao, L.; Lv, Y.; Huang, H.; Zhang, X. Towards optimal request mapping and response routing for content delivery networks. *IEEE Trans. Serv. Comput.* **2018**. [[CrossRef](#)]
43. Lin, T.; Xu, Y.; Zhang, G.; Xin, Y.; Li, Y.; Ci, S. R-iCDN: An approach supporting flexible content routing for ISP-operated CDN. In Proceedings of the 9th ACM workshop on Mobility in the evolving internet architecture, Maui, HI, USA, 11 September 2014; pp. 61–66.
44. Jiang, W.; Zhang-Shen, R.; Rexford, J.; Chiang, M. Cooperative content distribution and traffic engineering in an ISP network. *ACM SIGMETRICS Perform. Eval. Rev.* **2009**, *37*, 239–250.
45. Chang, C.H.; Lin, P.; Zhang, J.; Jeng, J.Y. Time dependent adaptive pricing for mobile internet access. In Proceedings of the 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Hong Kong, China, 26 April–1 May 2015; pp. 540–545.
46. Ma, Q.; Liu, Y.F.; Huang, J. Time and location aware mobile data pricing. *IEEE Trans. Mob. Comput.* **2016**, *15*, 2599–2613. [[CrossRef](#)]
47. Jin, M.; Gao, S.; Luo, H.; Li, J.; Zhang, Y.; Das, S.K. An Approach to Pre-Schedule Traffic in Time-Dependent Pricing Systems. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 334–347. [[CrossRef](#)]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).