Analyzing the Potential Benefits of CDN Augmentation Strategies for Internet Video Workloads

Athula Balachandran
Carnegie Mellon University
abalacha@cs.cmu.edu

Vyas Sekar
Stony Brook University
vyas@cs.stonybrook.edu

Aditya Akella
University of Wisconsin Madison
akella@cs.wisc.edu

Srinivasan Seshan
Carnegie Mellon University
srini@cs.cmu.edu

ABSTRACT

Video viewership over the Internet is rising rapidly, and market predictions suggest that video will comprise over 90% of Internet traffic in the next few years. At the same time, there have been signs that the Content Delivery Network (CDN) infrastructure is being stressed by ever-increasing amounts of video traffic. To meet these growing demands, the CDN infrastructure must be designed, provisioned and managed appropriately. Federated telco-CDNs and hybrid P2P-CDNs are two content delivery infrastructure designs that have gained significant industry attention recently. We observed several user access patterns that have important implications to these two designs in our unique dataset consisting of 30 million video sessions spanning around two months of video viewership from two large Internet video providers. These include partial interest in content, regional interests, temporal shift in peak load and patterns in evolution of interest. We analyze the impact of our findings on these two designs by performing a large scale measurement study. Surprisingly, we find significant amount of synchronous viewing behavior for Video On Demand (VOD) content, which makes hybrid P2P-CDN approach feasible for VOD and suggest new strategies for CDNs to reduce their infrastructure costs. We also find that federation can significantly reduce telco-CDN provisioning costs by as much as 95%.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: performance attributes; C.2.4 [**Computer-Communication Networks**]: Distributed Systems—Client/server

General Terms

Experimentation, Measurement, Performance

Keywords

Internet video, Measurement, User behavior

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IMC'13, October 23–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-1953-9/13/10 ...$15.00.
http://dx.doi.org/10.1145/2504730.2504743.

1. INTRODUCTION

Internet video has been consistently growing for the last few years, with some reports showing that around 51% of Internet traffic in 2011 was video [3]. Market predictions suggest that video will comprise over 90% of the traffic on the Internet in the next few years. The increasing video workload is placing the onus on content providers for efficient distribution of the content.

Today, content providers rely on CDNs to leverage their presence across different geographical locations to serve video content. With the ever-increasing amounts of video traffic, however, there are signs that the CDN infrastructure is being stressed [31, 32]. In this context, hybrid P2P-CDN and telco-CDN federation are two emerging strategies to augment the existing infrastructure that have received significant industry attention recently:

- Telco-CDN federation is based on the recent development amongst various CDNs operated by telecommunication companies to federate by interconnecting their networks and compete directly with the traditional CDNs [35, 10, 19, 4]. This would enable users to reach CDN caches that are closer. Interconnecting resources across telco-CDNs would also ensure better availability and will benefit the participating ISPs in terms of provisioning costs [4].
- A hybrid strategy of serving content from dedicated CDN servers using P2P technology (e.g., [16, 17]) has been around for a while in the research literature but has only recently seen traction in the industry [1, 37]. A hybrid P2P-CDN approach would provide the scalability advantage of P2P along with the reliability and manageability of CDNs.

Given that several industry efforts and working groups are underway on both fronts [35, 10, 19, 1, 37], it is crucial to analyze the potential benefits that these CDN augmentation strategies can offer for Internet video workloads. Our main contribution in this paper is in identifying video access patterns that have significant implications to these two strategies and analyzing the potential benefits of these two strategies. To the best of our knowledge, there has not been any previous large-scale study on the benefits of federated telco-CDN infrastructures. While there is prior work on analyzing the benefits of P2P augmentation, these were done long before Internet video became mainstream [16, 17], and hence were ahead of their times. Thus, it is also timely to revisit the benefits of P2P augmentation.

Using a dataset of around 30 million VOD and live sessions collected over two months from viewers across the United States, we characterize several video viewing patterns that have implications to these two designs including:

- **Regional interest:** Typically, we observe significant population induced difference in load across different regions (e.g., US East coast, US West coast, Mid-West). But, for live events with regional biases like a local team playing a match, we observe significantly skewed access rates from regions that exhibit low load in the typical case.
- **Temporal shift in peak load:** We observe strong diurnal effects in access patterns and also confirm temporal shifts between regions in the demand for VOD objects using cross-correlation analysis. The temporal shift in access pattern is caused by time zone differences. The video access load peaks at around 8pm local time for each region.
- **Evolution of interest:** We observe that peak demand for VOD objects occur on the day of release and the decay in demand in the subsequent days can be modeled using an exponential decay process. Interestingly, overall user viewing patterns are very different across genres (e.g., TV series, reality shows, news shows). For example, decay rates of news shows are much higher than TV series episodes. Also, TV series episodes have highly predictable and stable demand from week to week (i.e., across successive episodes).
- **Synchronized viewing patterns:** While we expect synchronous viewing behavior for live video, we unexpectedly observe synchrony in the viewership of VOD objects. This is especially true for popular shows during the peak demand period (e.g., evening of the day of release of the show).
- **Partial interest in content:** We reconfirm prior observations that users watch only part of the video during a session [11, 24]. For instance, in the case of VOD, a significant fraction of the viewers typically watch only the first 10 minutes of the video before quitting. We observe that around 4.5% of the users are "serial" early-quitters (analogous to channel surfing) while 16.6% of the users consistently watch videos to completion.

We develop simple models to capture the deployment of federated telco-CDNs and analyze the potential benefit of federation to increase availability and reduce provisioning required to serve video workloads. We also revisit the potential benefits that P2P-assisted architectures provide in the light of these video access patterns. Our key findings are:

- Telco-CDN federation can reduce the provisioning cost by as much as 95%. VOD workloads benefit from federation by offloading daily peak loads and live workloads benefit by offloading unexpected high traffic triggered by regional events.
- Using P2P can lead up to 87% bandwidth savings for the CDNs during peak access hours. Employing a strategy to filter out users who quit early by serving them using P2P can alone lead to 30% bandwidth savings for VOD traffic and 60% savings for live traffic.

In the rest of the paper, we discuss related work in Section 2 and provide an overview of our dataset in Section 3. We analyze the implications and potential benefits for federation across telco-CDNs and for hybrid P2P-CDNs in Section 4 and Section 5 respectively before concluding in Section 6.

2. RELATED WORK

In this section, we discuss the key similarities and differences with respect to past work in measuring different aspects of Internet video.

Video performance: Previous work confirms that video quality impacts user engagement across different content genres [20, 28]. Past work also identifies that many of the quality problems observed today are a result of spatial and temporal differences in CDN performance and suggest potential workarounds via cross-CDN optimization [31, 32]. The quality problems these studies uncover suggest that CDNs are stressed to deliver high-quality video and this motivates the need to explore strategies for augmenting CDNs.

Content popularity: There have been studies to understand content popularity in user-generated content systems (e.g., [18, 25]), IPTV systems (e.g., [13, 34, 12]), and other VOD systems (e.g., [27, 30, 21]). The focus of these studies was on understanding content popularity to enable efficient content caching and prefetching. Other studies analyze the impact of recommendation systems on program popularity (e.g., [38]) or the impact of flash-crowd like events (e.g. [22]). In contrast, our work focuses on analyzing the benefit of CDN augmentation techniques and extends these studies along two key dimensions. First, we model the longitudinal evolution in interest for different genres of video content and analyze its implications for designing a hybrid P2P-CDN infrastructure. Second, we analyze regional variations and biases in content popularity and its implications for provisioning a federated telco-CDN infrastructure.

P2P: Several pure P2P VOD systems aim to provide performance comparable to a server-side infrastructure at significantly lower cost (e.g., [14, 26, 27, 36]). There are already recent commercial efforts by CDNs to augment their infrastructures with P2P based solutions [1, 37]. Early work in the P2P space presented measurement-driven analysis on the feasibility and cost savings that hybrid-P2P technologies can bring [16, 17]. In some sense, these studies were ahead of their time—given that Internet video has really taken off only in the last 3-4 years, we believe it is critical to revisit these findings in light of new video viewing patterns. Specifically, our observations on synchronized viewing behavior for VOD and user join-leave patterns lead us to question the conventional wisdom in this space and we explore and evaluate new strategies for designing hybrid-P2P CDNs.

User behavior: Previous studies show that many users leave after a very short duration possibly due to low interest in the content (e.g., [11, 24]). While we reconfirm these observations, we also provide a systematic model for the fraction of video viewed by users using mixture model and gamma distributions, and highlight key differences between live and VOD viewing behavior. Furthermore, we analyze the implications of such partial user interest in the context of hybrid-P2P CDN deployments and explore new strategies for CDNs to reduce their bandwidth costs.

3. DATASET

The data used for this analysis was collected by `conviva.com` in real time using a client-side instrumentation library in the video player that collects information pertaining to a session. This library gets loaded when the user watches video on `conviva.com`'s affiliate content providers' websites. The library also listens to events from the player (e.g., seek, pause). The data is then aggregated and processed using Hadoop [5].

We focus on two of the most popular content providers (based in the US). These two providers appear consistently in the Top 500 sites in overall popularity ranking. Our analysis is based on data queried over two months—January 2012 and March 2012—and consists of over 30 million video viewing sessions during this period. We classify the video content into two categories:

- *VOD:* The first provider serves VOD objects that are between 35 minutes and 60 minutes long. These comprise TV series episodes, news shows, and reality show episodes.

- *Live:* The second provider serves sports events that are broadcast while the event is happening, and hence the viewing behavior is synchronized.

The VOD dataset consists of approximately 4 million users and 14 million viewing sessions and covers 1,000 video shows. The live dataset consists of around 4.5 million users and 16 million video viewing sessions covering around 10,000 different events. As in several prior studies on content popularity [30, 12], we also observe a heavy tailed Zipf distribution for overall popularity of objects for both VOD and live. Whereas most objects have few accesses over the two months, some extremely popular objects had significant viewership. On average, users viewed 4 VOD objects and 2 live events during the course of a month, which amounts to 85 minutes of VOD objects and 65 minutes of live events per month. We also observed a few heavy-hitters who watched upwards of 500 videos per month on these websites.

Session characteristics: In order to understand user behavior, we look at several characteristics of individual video sessions. Specifically, for each session we collected the following information:

- *ClientID:* The first time a client watches a video on the player, a unique identifier is assigned to the player and stored in a Flash cookie to be used by subsequent views.
- *Geographical location:* Country, state and city of the user.
- *Provider:* Information on the AS/ISP from which the request originated.
- *Session events:* Start time and duration of the session along with details on other user interaction events like pausing and stopping.
- *Session Performance:* Average bitrate, estimated bandwidth etc. during the playback.
- *Content:* Information on the content being watched, in particular, the name of the video (which we use for classifying videos into genres) and the actual duration of the content (e.g., 45 minute show).

Region	States
1	MA, NH, TV, ME, RI, CT
2	NY, PA, NJ
3	WI, MI, IL, IN, OH
4	MO, ND, SD, NE, KS, MN, IA
5	DE, MD, DC, VA, WV, NC, SC, GA, FL
6	KY, TN, MS, AL
7	OK, TX, AR, LA
8	ID, MT, WY, NV, UT, CO, AZ, NM
9	AK, WA, OR, CA, HI

Table 1: *List of Regions*

Geographical regions: We limit our study to clients within the United States. We classified the country into 9 regions using the censor bureau designated areas [2] as shown in Table 1. Not surprisingly, we observe that the average load in terms of number of accesses is significantly different across different regions, and they are largely correlated with the total population of the region. We observe that this pattern holds for both live and VOD traffic except in the case of some events that have regional bias. We explore this aspect further in Section 4.

4. ANALYZING TELCO-CDN FEDERATION

The tremendous increase in video traffic on the Internet over the past few years has caused great challenges for ISPs. The increasing traffic has strained the ISP networks leading to higher costs and maintenance issues. However, this trend has not significantly contributed to much increase in revenue for ISPs since most of the video content is served by content providers using CDNs. As a result, several ISPs have started deploying proprietary CDNs inside their own network, providing services to content providers to serve content from caches closer to customers. This could result in increased revenue for the ISPs along with traffic reduction caused by content caching [4].

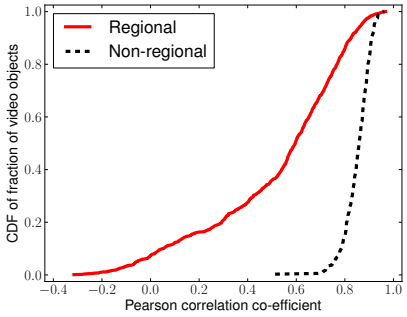


Figure 1: *The result shows the CDF of the correlation coefficient between the #views and the population of the region for the live dataset. Non-regional content is strongly correlated with population whereas regional content is uncorrelated or negatively correlated.*

There has also been recent developments that point to interest among ISPs to deploy *telco CDN federations* by consolidating their CDN capacity and offering services to users in other ISPs [35, 10, 19]. By interconnecting telco-CDNs, consumers can reach CDN caches that are closer and are also ensured of better availability and service in case of local network congestion. Pooling resources across ISPs could potentially benefit the participating ISPs in terms of provisioning costs. It also enables ISPs to provide a global "virtual CDN" service to the content providers [4].

Although there have been pilot deployment efforts and initiatives for standardization of a federated-CDN architecture in the industry [4, 10], we are not aware of any study quantifying the benefits of telco-CDN federation, specifically in the context of Internet video. We first present video access patterns that we observed in our dataset that have significant implications for CDN federation in Section 4.1. We further quantify the potential benefits that telco-CDN federation can provide and put them in context with our findings on the user access patterns. To this end, we develop simple models to capture the deployment of telco-CDN federations that help us determine the potential benefits that such federation offers in Sections 4.2 and 4.3. We use this to evaluate the benefits of telco-CDN federation using our dataset for live and VOD content separately in Section 4.4. To the best of our knowledge this is the first large scale study to quantify the benefits of telco-CDN federation.

4.1 User Access Patterns

We observed video access patterns for live and VOD content that have implications to telco-CDN federation. For instance, in our live dataset, we observed unexpected surges in demand for certain objects from regions which can potentially be served using spare capacity in servers in other regions if CDNs federate. Similarly, we observed strong temporal shifts in when specific regions hit peak load in the VOD dataset opening up new possibilities for handling peak loads using federation. We finally also present statistics on ISP coverage and their relative performance which also have important implications when ISPs decide to federate.

4.1.1 Regional Interests

Typically, the number of accesses to a particular content from a geographical region is strongly correlated with the total population of the region. However, in our live dataset, we observed anomalies in the case of content with region-specific interest (e.g., when a local team is playing a game). Such unexpected surges in demands triggered by regional interests can potentially be served from servers in other regions if CDNs federate.

Our data consists of only clients within the United States and it does not contain tags with event region details. Hence, we manually classified the content as regional or non-regional based on whether it appeals to a particular region within the US. Sports matches between local teams within the US (e.g., NCAA) were classified as regional as opposed to events that are non-regional to the US viewers (e.g., Eurocup soccer).

We computed the Pearson correlation coefficient [33] between the number of accesses from each region to the population of the region (obtained from census data [2]). Figure 1 shows the CDF of the correlation coefficient across video objects for all the live objects. We observe that access rates of non-regional content show strong correlation to the population, whereas for regional matches it is uncorrelated or negatively correlated. This is because of skewed access rates from normally not so active regions because of a sporting event that has a local team. However, some regional matches show high correlation. These are highly popular events (e.g., final rounds of NCAA are of interest to everyone in the US).

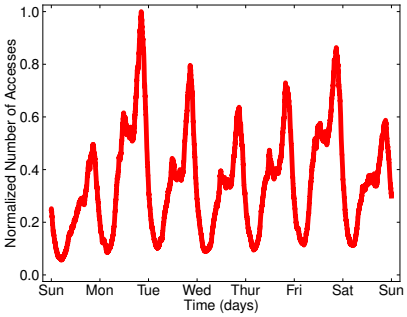
Implications: The skeweness in access rates caused by regional interest is an important factor to be considered while provisioning the delivery infrastructure to handle unexpected high loads. Federation can potentially help offload such unexpected surges triggered by regional interests by using spare capacity in CDNs in other regions.

4.1.2 Temporal shift in peak loads

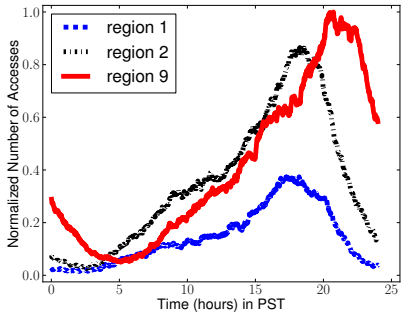
Figure 2a provides an overview of the VOD dataset by plotting the time series of the normalized number of videos accessed across all regions at per minute granularity for a week. As expected, we clearly observe strong time-of-day effects. To identify regional variations in peak load, we zoom into a day and plot the time series of the normalized number of accesses separately for each region in Figure 2b. Due to lack of space, we only show the results for the top 3 regions. The number of accesses peaks around 8 pm local time with a lull in the night. We observe that there is a difference between the time when the load peaks at different regions (caused by time zone difference). Also, we see that the peak loads are different across regions—they are largely correlated with the total population of the region.

We perform cross-correlation analysis to confirm the temporal shift in access patterns over the entire two months of data. Cross-correlation measures the degree of similarity between two time series as a function of a time lag applied to one of them. Let $X =< X_0, X_1, ..., X_i, ... >$ denote the time series of the number of accesses as a vector where X_i is the number of accesses at time index i and $E(X_i)$ and σ_{X_i} represents the expected value and standard deviation respectively. For a given time lag k, the cross-correlation co-efficient between two time series vectors $X =< X_0, X_1, ..., X_i, ... >$ and $Y =< Y_0, Y_1, ..., Y_j, ... >$ is defined as:

$$\tau(k) = \frac{E(X_i Y_{i+k}) - E(X_i)E(Y_{i+k})}{\sigma_{X_i}\sigma_{Y_{i+k}}} \quad (1)$$



(a) Aggregate access pattern



(b) Region-wise access pattern

Figure 2: *Diurnal characteristics of access pattern*

The cross-correlation coefficient lies in the range of [-1,1] where $\tau(k) = 1$ implies perfect correlation at lag k and $\tau(k) = 0$ implies no correlation at lag k. We use cross-correlation to analyze the time shift in the access pattern across regions. We performed analysis across all region pairs at lags of one hour each. Due to space constraint, we present the co-coefficients plotted at different lags for the top 3 region pairs in Figure 3. Regions 1 and 2 fall in the same time zone and hence the $\tau(k)$ is highest at $k = 0$. Region 9 is 3 hours behind regions 1 and 2 and hence $\tau(k)$ is highest at $k = 3$. We observe this pattern holds for all the region pairs.

Implications:
The temporal shift in peak access times across different regions opens up new opportunities to handle peak loads—e.g., spare capacity at servers in regions 1 and 2 can be used to serve content in region 9 when access rates peak at region 9.

4.1.3 ISP performance

We study the relative performance of the ISPs over the month in terms of video quality using two key metrics identified in [20]: (1) buffering ratio defined as the percentage of session time spent in buffering, and (2) the average bitrate for each session. We summarize the relative performance of top ISPs using box-and-whiskers plots (Figure 4) showing the minimum, 25%ile, median, 75%ile, and 90%ile values observed across sessions. Our results corroborate a similar report released by Netflix in May 2011 [6]. The mean performance of the ISPs are very similar, with cable ISPs like Comcast and Cox providing marginally better bitrates in the median case. We also see that wireless providers like Clearwire and Verizon Wireless provide lower bitrates compared to their wired counterparts. As observed in [20], majority of the sessions have very low buffering ratio. The median buffering ratio is zero for all

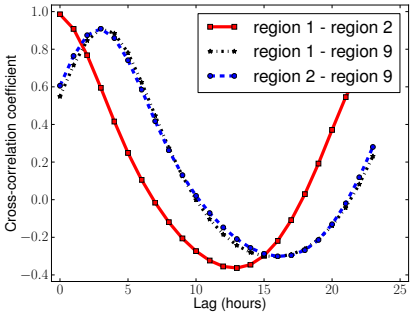


Figure 3: *Cross correlation analysis confirms the temporal shift in access pattern over two months of data*

the ISPs. Verizon Wireless and Windstream have marginally higher buffering ratio in the 75%ile and 90%ile case.[1]

Implications: Since the overall performance of most ISPs are very similar, they can potentially collaboratively use their resources without worrying that their customers may see poor performance from their federating "peers" due to network effects.

4.1.4 ISP Regional presence

ISP	NY (%)	LA (%)
Comcast	1.4	1.7
AT&T	6.1	24.7
Verizon	41.7	56.3
RoadRunner	34.1	2.0
Cox	-	-
Charter	-	1.2
Qwest	-	-
Cablevision	2.9	-
Frontier	-	-
Windstream	-	-
Others	13.8	14.1

Table 2: *Fraction of clients observed from individual ISPs for top-2 cities*

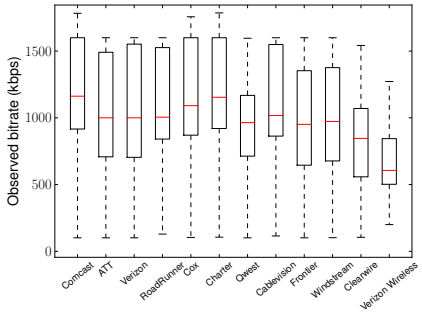
Table 2 shows a different split across ISPs for two large cities. We observe that ISPs have significant regional biases in their coverage. For instance, while Verizon and RoadRunner have a large fraction of clients in New York city, AT&T and Verizon have a more dominant presence in LA. We also observe that some ISPs have a small fraction of their clients in cities where they are not dominant. For example, RoadRunner appears to contribute 2% of the total users in LA and AT&T has 6% in NY.

Implications: An ISP may not want to roll out new video delivery infrastructure in regions where it does not already have a significant coverage and in this case might want to direct its customers to servers located in cooperating ISPs.

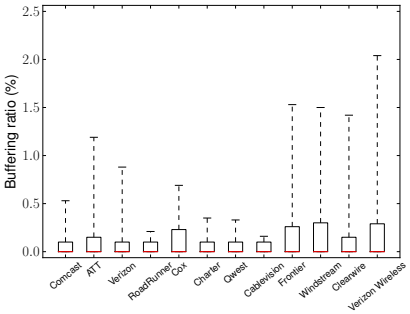
4.2 System Model

In order to analyze the potential benefits of federation, we use a simplified system model for federated telco-CDNs. Figure 5 provides a high-level overview of our system model. As we have discussed earlier, there are several geographical regions, represented

[1]Even though these wireless providers have worse performance, we did not observe any significant difference in user behavior. This is somewhat surprising given recent studies on how quality affects behavior [28]. It probably points to how users get "trained" for lower expectations, thereby increasing their tolerance.



(a) Observed bitrate



(b) Buffering ratio

Figure 4: *Performance of top ISPs*

by $Region_r$. We currently use the regions described in Table 1, but this could also be more fine-grained (e.g., city or metro-area). Each region may have several ISPs, each denoted by ISP_i and each such ISP has some provisioned CDN capacity (number of users that can be served) in each region denoted by $Cap_{r,i}$.

Similar to today's ISP peering for connectivity, we envision pre-established "peering" relationships between ISPs across different regions to share their spare capacity. Let $P(r,i)$ be the set of all region-ISP tuples with whom ISP_i in $Region_r$ has peering relationships. We use $r'i' \in P(r,i)$ to specify that $ISP_{i'}$ in $Region_{r'}$ has such a peering relationship with ISP_i in $Region_r$. This means that $ISP_{i'}$ in $Region_{r'}$ can offer its services or spare capacity to serve users from ISP_i in $Region_r$. This allows us to flexibly capture different kinds of telco CDN peering relationships.[2] For example, in the trivial case, without any cross-ISP federation or cross-region resource sharing $P(r,i)$ relationship only contains the current ISP-region combination. In the most general case, all ISPs can share capacity with each other.

Let $Demand_{ri}(t)$ be the number of video users in region $Region_r$ from ISP_i at a given epoch t. As a first step, we only focus on the number of users and not on the specific bitrates they choose. Let $N_{r',i' \to r,i}(t)$ denote the number of users served using servers located in $ISP_{i'}$ in $Region_{r'}$ to clients from ISP_i in $Region_r$. at epoch t. We use $L_{r'i';r,i}$ to denote the latency cost incurred in this process. For clarity of discussion, we use a simple latency function at the level of "region-hops" between neighboring regions; we can extend this to incorporate more fine-grained inter-ISP latency within and across regions.

[2]ISPs can also employ other relationships and policies. For example, ISPs with higher server capacity can potentially employ "provider-customer" relationships. Our current model does not capture such fine-grained policies and cost models.

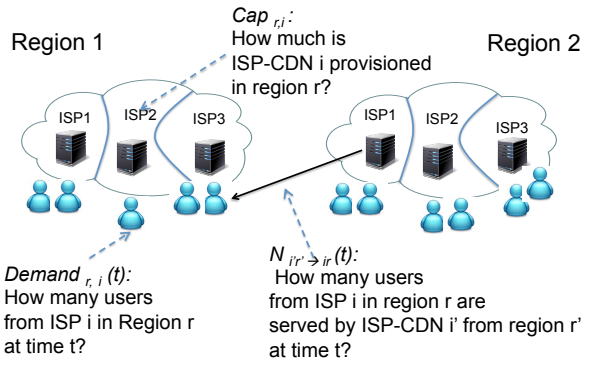


Figure 5: *System model for telco CDN federation*

$$\text{Minimize: } Latency(t) + \alpha \times Dropped(t)$$

$$\forall r, i : Dropped_{r,i}(t) = Demand_{r,i}(t) - \sum_{r',i':r',i' \in P(r,i)} N_{r',i' \to r,i}(t) \quad (2)$$

$$\forall r, i : Dropped_{r,i}(t) \geq 0 \quad (3)$$

$$Dropped(t) = \sum_{r,i} Dropped_{r,i}(t) \quad (4)$$

$$Latency(t) = \sum_{r,i:r'i' \in P(r,i)} L_{r'i';r,i} \times N_{r',i' \to ri}(t) \quad (5)$$

$$\forall r', i' : \sum_{ri:r'i' \in P(r,i)} N_{r'i' \to ri}(t) \leq Cap_{r'i'} \quad (6)$$

Figure 6: *Linear program for finding the optimal allocation in each logical epoch*

4.3 Global provisioning problem

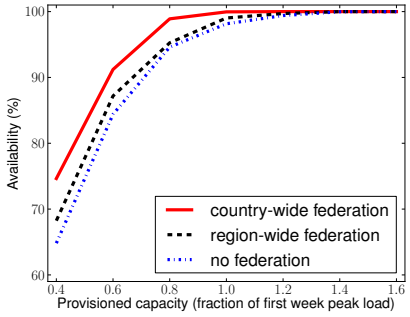
Given this setup, we can now formulate the telco CDN federation problem as a *resource allocation* problem with the resources being the servers in different ISP/region combinations and the demands being the users in ISP/region combination. The linear program in Figure 6 formally describes the high-level optimization problem.

There are two high-level goals here. First, we want to accommodate as many users as possible given the current capacity provisioned at the different ISPs in various regions. Second, we want to minimize the network footprint of these assignments and ensure that requests are served as locally as possible. However, given a specific provisioning regime, it may not always be possible to fully meet the demands and some requests have to be invariably dropped. We trade off the relative importance of these objectives (i.e., latency vs. coverage) using the cost factor α in the objective function that captures the penalty for dropping users. By setting α to be very high, we can ensure that the demand is maximally met even if it requires fetching content from remote servers.

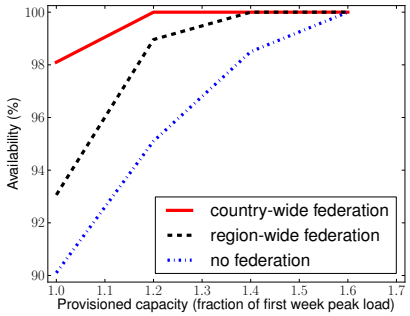
We capture the number of requests that are *dropped* in each ISP-region tuple Eq (2) and the total number of drops in Eq (4). (Of course, the number of requests dropped cannot be negative so we have the sanity check in Eq (3).) We model the overall latency footprint in Eq (5) using a simple weighted sum. Finally, we have a natural capacity constraint that in each region no ISP exceeds its provisioned capacity and this is captured in Eq (6).

4.4 Evaluation

We use the above formal model to evaluate the potential benefits of telco CDN federation for live and VOD content using our dataset.



(a) Over multiple weeks



(b) 5 hour peak access time

Figure 7: *Benefits from federation for VOD*

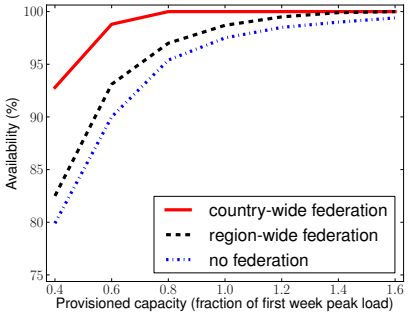
Methodology: We use the user access behavior during the first week and find the peak load at each ISP_i in each $Region_r$ to determine a baseline for provisioning $Cap_{r,i}$ at each ISP-region combination. Specifically, we consider a provisioning exercise where each ISP-region combination is provisioned to handle a fraction of this peak load. Then, we use the remaining three weeks of user arrival patterns to analyze the effectiveness of such provisioning with and without federation. We set the value of α to be extremely high to minimize drops. The particular measure of interest in this exercise is the *availability* which we define as the fraction of requests that are served by the overall distribution infrastructure. Formally, this can be expressed as:

$$Availability = \frac{\sum_{r,i,t} \sum_{ri:r'i' \in P(r,i)} N_{r'i' \to ri}(t)}{\sum_{r,i,t} Demand_{ri}(t)} \quad (7)$$

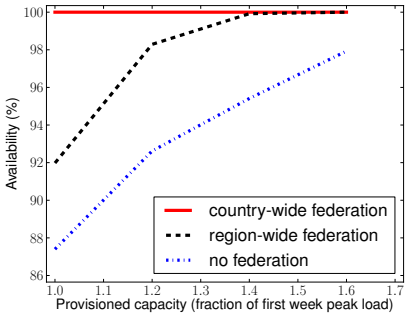
In the following evaluation, we consider three scenarios:

- *No federation:* Here, $P(r,i)$ consists of just itself.
- *Region-wide federation:* $P(r,i)$ consists of all ISPs within the same region
- *Country-wide federation:* $P(r,i)$ consists of all ISPs in all regions.

Benefits for VOD content: Figure 7a shows the overall benefits of federation using the VOD dataset. As mentioned before, each telco-CDN provisions for a fraction of the observed peak load from the first week. For instance, as shown in Figure 7a, when each telco-CDN provisions for 40% of the observed peak load in the first week (this roughly corresponds to the average observed load), we see that there is almost a 5% increase in availability with just region-wide federation when evaluated over the workload from the next 3 weeks. Country-wide federation results in about 10% increase in availability of the system.



(a) Over multiple weeks



(b) During a popular regional event

Figure 8: *Benefits from federation for live*

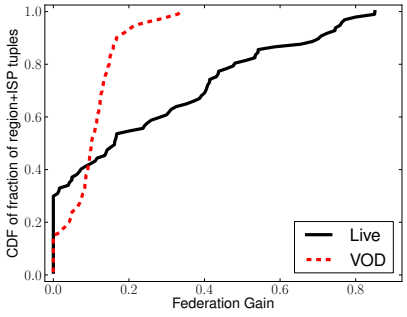


Figure 9: *CDF of federation gain*

Although peak loads are roughly predictable for VOD content, in order to achieve 100% availability without federation, each ISP-region needs to over-provision with 1.6 times the observed first week peak load. Whereas, provisioning with 1.4 times the peak load would be enough with region-wide cooperation and provisioning with 1.2 times the observed first week peak load is sufficient to sustain the workload over the next 3 weeks with country-wide federation. This points to the fact that despite the synchrony in viewing behavior, peak loads are slightly offset across different ISPs within a region enabling using spare resources from other ISPs within the same region to improve availability. Similarly, the temporal shift in peak loads across regions due to time zone effect enables even more sharing of resources, reducing the provisioning cost to meet unexpected demands.

This result focuses on the average availability across the entire three week period. The benefits of federation are the most pronounced during peak access times. In order to highlight this further, we evaluate the availability of the system during a five-hour peak access period in Figure 7b. This result shows that without federation, roughly 10% of users will need to be dropped if each ISP-region was simply provisioned for the peak load observed in the first week, whereas we get only 2% dropped users with country-wide federation.

Benefits for live content: Live events have more unpredictable workloads due to interest-induced regional effects leading to unexpected higher load from typically low-load regions (e.g., when the local team plays a match). Consequently, we expect that pooling in resources from other ISPs and regions via federation will be even more beneficial.

We use the live dataset and show the overall benefits from federation in Figure 8a. For instance, as seen in Figure 8a, when provisioned for 40% of the peak load from first week, region-wide federation would increase the availability by around 3% (lower than the VOD case) while country-wide federation would increase the availability by 13% (higher than VOD) when evaluated on the next 3 week workload.

We zoom into a peak access time of 3 hours when a regional match was being broadcasted and repeat the study to show the benefits of federation in Figure 8b. We observe that employing country-wide federation, the system can achieve 100% availability by just provisioning for the observed peak load from the first week. Region-wide federation would require provisioning the system with 1.4 times the peak load. Without any federation, we observed that provisioning for 20 times the peak load is required to meet 100% availability—i.e., federation decreases the required provisioning by around 95%. This clearly shows that live events can benefit a lot from federation because unpredictable local peaks in access rates are much more common.

Which ISPs benefit the most: The immediate question that arises when we consider peering and federation is fairness. We analyze if specific categories of ISPs and/or regions are more likely to gain from federation compared to others. To this end, we define a *federation gain* metric for each ISP-region combination as the ratio between the total volume of requests served by other ISP/regions to the total capacity of this ISP-region $\frac{TotalServedbyOthers}{Capacity}$. Figure 9 shows the CDF of federation gain over all ISP-region combinations using country-wide federation. We observe that federation gains are lower and more uniform for VOD (highest gain is 0.4) while they are more skewed and higher in value in the case of live (highest gain is 0.8). Looking at the ISP-region combinations that benefit the most, we observe that ISPs in typically low-load regions have higher benefits in the case of live. This is because of unpredictable local peaks caused by events of regional interest. In the case of VOD, the ISPs in high-load regions have larger benefits. The benefits were mostly from offloading unexpected daily peak loads.

Performance Costs: Employing telco-CDN federation might lead to the selection of CDN servers far from a user, which would increase latency. Our approach to limit these performance issues is to use a very simple hop-based latency model, but a more systematic scheme would take into consideration the impact of CDN server selection on users' quality-of-experience [32, 15]. Design and analysis of a system taking these into consideration is outside the scope of this paper.

4.5 Main observations

To summarize, the key observations are:

- Federation increases the overall availability of the system with lower provisioning overhead (as much as 95% reduction in the case of live). The benefits are higher with higher level of co-

operation (the upper bound being pooling in all resources within the country).
- VOD workload benefits from federation by offloading daily peak loads. We notice that ISPs from typically high load regions benefit the most.
- Live workload benefits from federation by offloading unexpected high traffic triggered by regional events. Here, the benefits are higher for ISPs in typically low-load regions.

5. ANALYZING HYBRID P2P-CDN

The two predominant technologies for delivering videos to end users are CDNs based on the client-server model of delivery, and server-less P2P mechanisms. Whereas CDNs provide reliable delivery using a geographically distributed delivery infrastructure, P2P enables scalability by leveraging on the bandwidth resources of individual users. There has been renewed interest in the CDN industry to augment traditional CDN based video delivery with P2P technologies. This trend is driven by the need for higher quality (e.g., [7]), and is further enabled by new technologies that allow P2P modules to run within browsers and video players without requiring separate applications [9, 1, 8].

Conventional wisdom in the use of P2P-assisted hybrid CDNs suggests that:

- P2P is only likely to be useful for live content because VOD may have low synchrony with very few users viewing the same part of the video at the same time.
- It is better to use the CDN for the early bootstrap process as clients arrive and use P2P only for the steady-state once the "swarm" dynamics stabilize.

However, we observed several user access patterns and behaviors in our dataset that give us reason to revisit and question these traditional assumptions in hybrid CDN-P2P designs. We present these observations in Section 5.1. Based on these observations, we propose new strategies for CDNs to reduce their infrastructure costs by using a hybrid-P2P approach and evaluate these proposals in Section 5.2.

5.1 User Access Patterns

We observed several user access patterns that have very important implications to the design of hybrid P2P-CDN architecture. For example, we observed that several users watch only the first few minutes of a video in the case of both VOD and live content. This could imply that some parts of the video objects are more amenable to P2P than the rest. We also explore the evolution of interest for both VOD and live content to understand when it would be more beneficial to employ P2P strategies.

5.1.1 Partial Interest in content

We observed that several users had partial interest in the content that they are viewing and they quit the session without watching the content fully in the case of both VOD and live. If most users watch only the first few minutes of the video before quitting, P2P might be more amenable for the first few chunks since there will be more copies of them compared to the rest of the video. Hence, we further investigated the temporal characteristics of user behavior within a given video session and analyzed what fraction of a video object users typically view before quitting.

For VOD content, Figure 10a shows that based on the fraction of video that a user viewed within a session, users can be classified into three categories:

- **Early-quitter:** A large fraction of the users watch less than 10% of the video before quitting the sessions. These users might be "sampling" the video.
- **Drop-out:** We observe that further on, users steadily drop out of the video session possibly due to quality issues or lack of interest in the content.
- **Steady viewer:** A significant fraction of the users watch the video to completion.

We can model this using a mixture model with three separate components [33]. As shown in Figures 10b and 10c, we try to find the best fitting probability distribution for the early-quitter and steady viewer components. Inspecting visually and using mean squared error test, we choose the gamma distribution to represent both the early-quitter and the steady viewer components. We model the drop-out component using a uniform distribution. We then use expectation maximization [33] to estimate the mixture model parameters and obtain the model as shown in Figure 10a. These models can be used for simulating video viewing behaviors in the future.

The previous result considers the behavior of users in aggregate. A natural question then is whether specific users behave in a consistent way across multiple video sessions. To this end, we profile users' viewing history across multiple sessions by grouping sessions by the user as identified using their unique *ClientID*. We find that 4.5% of the users quit the session early for more than 75% of the sessions; i.e., these users are "serial" early quitters. Similarly, 16.6% of the users consistently viewed the video to completion; i.e., these are consistently steady viewers.

Similar to the analysis that we did for VOD content, we also analyze what fraction of the live content users typically view before quitting and plot the distribution in Figure 11a. We observe that based on the fraction of video viewed within a session, users watching live content can be classified into two categories:

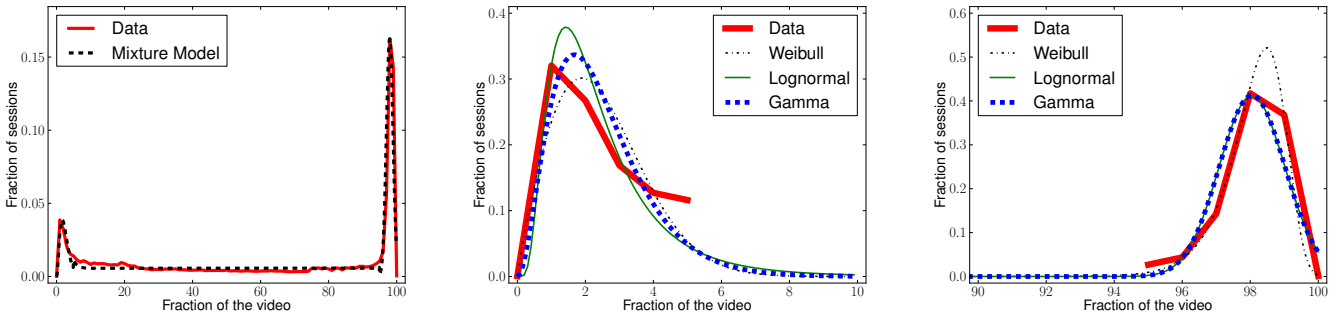
- **Early-quitter:** A very large fraction of users watch less than 20% of the video before quitting the session.
- **Drop-out:** The remaining fraction of users steadily drop out of the video session.

Figure 11b zooms into the early-quitter part of the plot and shows how well different distributions fit the data. Inspecting visually and using mean squared error test, we find that the gamma distribution is the best fit and model it in Figure 11a. A large fraction of users quitting the session early for live content might imply that the first part of the event is the most popular part. However, as we see in Figure 11c users arrive randomly within the event and stay for short periods of time before quitting. Hence the first part of the event is not necessarily the most popular part.

We also profile users' viewing history (based on the unique *Client ID*) and notice that around 20.7% of the clients are "serial" early quitters—i.e., they quit the session early for more than 75% of the sessions for live content. We also observe several users joining and quitting multiple times during the same event. Since our dataset consists of sporting events, one possibility is that they might be checking for the current score of the match.

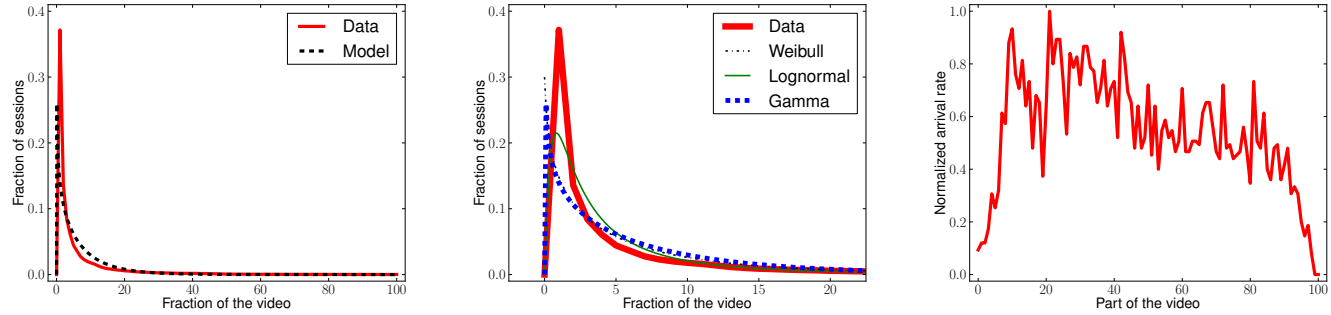
Contrasting the observations of live and VOD, we observe the following key differences:

- The early-quitters watch higher fractions of video in the case of live (up to 20% of the video) when compared to VOD (up to 10% of the video). Drop-out percentage is less pronounced in the case of live and we also do not observe a significant fraction of users viewing the entire event.
- In the case of VOD, users typically view the video from the start as opposed to live where people join at random times.



(a) PDF of fraction of video viewed before quitting (b) Component 1: Early quitters (c) Component 3: Steady viewers

Figure 10: *Distribution of the fraction of video viewed for VOD*



(a) PDF of fraction of video viewed (b) Modeling fraction of video viewed (c) Arrival rate

Figure 11: *Fraction of video viewed and arrival rate for live objects*

- We observe a higher fraction of “serial" early-quitters in the case of live.

Implications:
(1) This analysis is particularly relevant in the context of augmenting CDNs with P2P based delivery. For example, if most users are likely to only watch a small fraction of video, then P2P will be less effective at offloading some of the server load as there may not be sufficient number of cached copies of the content.
(2) Content providers and content delivery infrastructures can identify the early quitters and steady viewers and customize the allocation of resources (e.g., use P2P to serve the content to early quitters who are “sampling" the video).
(3) Although user behavior like early-quitting are similar for live and VOD, we need to consider the differences in access patterns. For example, since early-quitters watch the video for longer in the case of live, employing P2P to serve early-quitters might imply serving more content using P2P in the case of live than VOD.
(4) Similarly, the fact that users typically view VOD objects from the start and quit early might imply higher availability for the first few chunks of the video. For live, even though users quit quickly, they arrive randomly in between the event and hence the first part of the event may not necessarily be the most popular part.
(5) Beyond hybrid P2P designs, this analysis is very interesting because understanding such patterns is especially useful for content providers and content delivery infrastructures in order to maximize some higher-level objective (e.g., where to place ad impressions to maximize revenue).

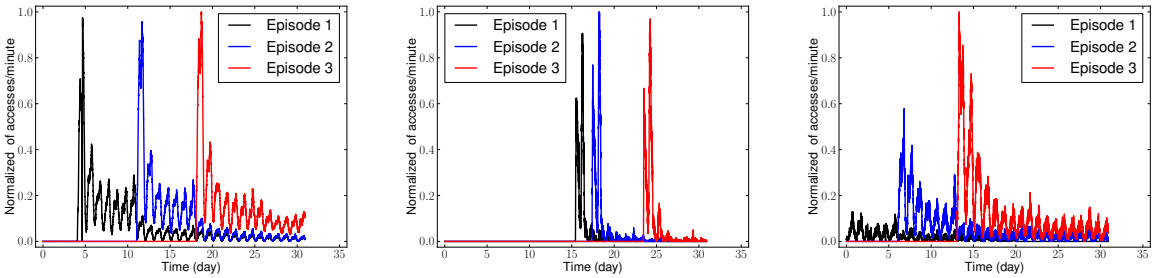
5.1.2 Evolution of interest

It is crucial to investigate how popularity of content evolves over time since it could point to certain times when P2P strategies might be more beneficial. For example, if more users watch VOD videos on the day of release, there would be higher synchrony in viewership that could lead to higher benefits from employing P2P.

We classify VOD objects into three categories: TV series, news show or reality show and model the evolution in interest along two key dimensions: (1) temporal decay in popularity for a given object (i.e., a fixed episode for a fixed show) over days, and (2) demand predictability across multiple episodes for a given show. We develop models for these parameters that can be used for simulating video workloads in the future. Live objects are viewed while the event is happening and are not available afterwards. Hence, we explore how the interest in the content evolves during the event by analyzing hotspot points in events.

Figure 12 shows the temporal variation in popularity and how demand for the content decays for sample objects from the three categories of VOD objects. First, for TV series episodes, the demand for episodes appears relatively stable and predictable week to week, and it decays gradually over time. Second, for news shows, we see the demand hits a peak on the release date and decreases quite dramatically. Finally, for reality shows, while we see a decay in demand from the time of release, there is less predictable viewership across different episodes. We further characterize the temporal decay and demand predictability for VOD objects.

Temporal decay in popularity for VOD objects: We observe that the highest number of accesses occurs on the day of release for



(a) TV series (b) News shows (c) Reality shows

Figure 12: *Temporal change in popularity of VOD objects*

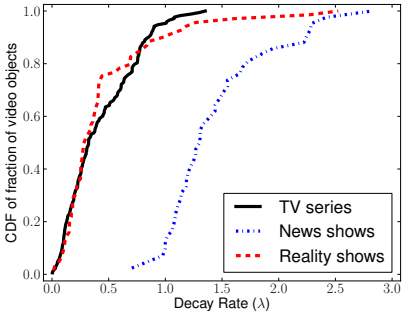
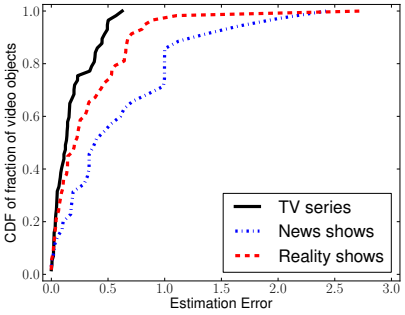
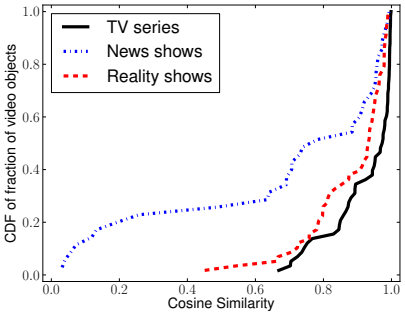


Figure 13: *CDF of decay rate for different genres*



(a) Estimation Error



(b) Cosine Similarity

Figure 14: *Characterizing demand predictability*

all the VOD objects, and the daily peak number of access for each object decreases with time. Exponential decay appears to be the best fit for modeling the decay (compared to linear decay process) based on aggregate mean-squared error test across multiple objects. The decay in peak number of accesses can hence be characterized using an exponential decay function as follows:

$$P(t) = P_0 e^{-\lambda t} \tag{8}$$

where P_0 is the peak number of access on the day of release, $P(t)$ is the peak number of access on the day t since release and λ is the decay constant. Figure 13 shows the CDF of the estimated decay rate (λ) for all the VOD objects categorized by their genres. News shows have high decay rates which implies that these objects turn stale quicker and their demand decreases dramatically within a day of release. In contrast, TV shows have lower decay rates. The decay rate of reality shows have more variability.

Demand predictability for VOD objects: We analyze how predictable the demand for shows are based on their viewership history. For this, we use the viewership pattern of the latest episode as an estimate for the next episode.[3] We characterize (1) how close were the peak number of accesses on the day of release? (2) how similar were the decay patterns?

(1) *Estimation Error:* Using the most recent episode as a predictor for the peak demand for the next episode, we calculate:

$$Estimation\ error = \frac{|P_{actual} - P_{estimated}|}{P_{\texttt{actual}}} \tag{9}$$

where P_{actual} is the peak number of accesses on the day of release of the show and $P_{estimated}$ is the estimated peak number of accesses (i.e., the peak number of accesses observed for the previous show in the series). Figure 14a shows the CDF of relative error for different genres. We observe that TV series have lower relative error values, implying that their peak access rates across episodes are more steady and predictable. News shows and reality shows tend to have more variable peak accesses.

(2) *Cosine similarity:* Apart from categorizing the predictability of the peak number of accesses, we also want to estimate how similar the decay patterns are across episodes within a series. If $X =< x_0, x_1, ..., x_i, ... >$ denotes the vector of the number of accesses for the object starting from the hour of release and $Y =< y_0, y_1, ..., y_j, ... >$ denote the vector of number of accesses for the previous episode of the series, we compute the similarity between the episodes as:

$$Cosine\ similarity = \frac{\sum_{i=0}^{n} x_i \times y_i}{\sqrt{\sum_{i=0}^{n}(x_i)^2} \times \sqrt{\sum_{i=0}^{n}(y_i)^2}} \tag{10}$$

[3]Our dataset is limited to 2 to 4 episodes per show. Modeling viewership history over a larger span is an interesting direction for future work.

Cosine similarity takes values in the range [0,1] where 1 implies high similarity and 0 indicates independence.[4] Figure 14b shows the CDF of cosine similarity for different VOD objects. We observe that TV series have the highest similarity. The access patterns of news shows tend to be very different from the previous episodes. The cosine similarity of reality shows falls in between the TV series and news shows.

Hotspots in live events: From a provisioning perspective, it is important to understand how the interest in the content evolves during the live event. Figure 15 gives two extreme examples of how overall interest in the content changes within a session. Figure 15a shows an example of an event where the number of viewers watching the event was steady throughout the event whereas Figure 15b is an example of an event where there was a particular point in the event where interest peaked and then it died down. We refer to the location with the peak number of simultaneous viewers as the *hotspot point* within the event.

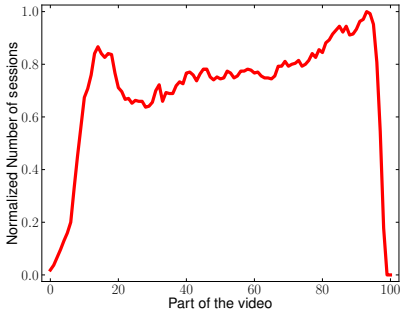
Given these extremes, a natural question is what does a typical live event look like? To this end, we systematically analyze the live events on two dimensions: (1) where do hotspots occur in a video? (2) how pronounced is the hotspot? Figure 16a shows the CDF of the hotspot point location for all the live events. We see that there is no particular location where hotspots typically occur. To capture how pronounced a hotspot is, we compute the *peak-to-average* ratio of the number of simultaneous viewers at a given point of time during the session. Looking at the distribution of the peak-to-average ratio (Figure 16b), we observe that majority of the events have flat access rates (similar to Figure 15a). However, events with pronounced hotspots tend to have the hotspot point towards the beginning of the event.

Implications:
(1) The strong diurnal patterns observed from the time series plots again point to high synchrony of viewing even at a per-object basis. This bodes favorably in using P2P augmentation strategies for delivering VOD content.
(2) The decay rates indicate higher synchronous viewing behavior on the day of release of the show. This is also when we see higher demand in objects and when the CDNs might benefit more from using P2P strategies.
(3) Comparing genres, news shows have very high decay rates and are least predictable. This could potentially lead to sudden unexpected surges in demands and hence CDNs may need to invoke P2P-based strategies dynamically to handle these loads. However, TV series have more stable demands that are predictable and with lower decay. This means that the delivery infrastructure can be provisioned accordingly. Reality shows have much more variability in terms of decay and predictability.
(4) Since we do not observe any typical pattern for hotspot locations across live objects, CDNs may need to dynamically invoke strategies to handle the peak loads by using P2P depending on how interest evolves for the particular content.

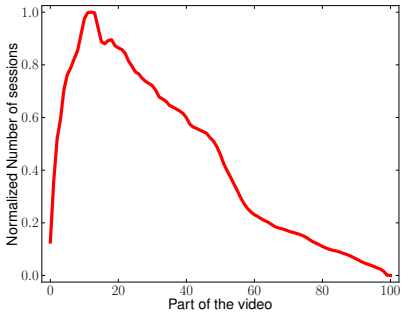
5.2 Revisiting P2P-CDN benefits

Contrary to the conventional wisdom in this space, first, we posit that P2P might be more useful for VOD than previously assumed and that these benefits can be achieved even without assuming that each peer is caching the whole content as in [27]. Second, the presence of early quitters suggest CDNs may want to rethink how they allocate constrained server resources. Specifically, we leverage the higher interest in the early chunks coupled with the tendency of

[4]Because X and Y are both positive vectors, the cosine similarity can't be negative.



(a) Flat



(b) Non-flat

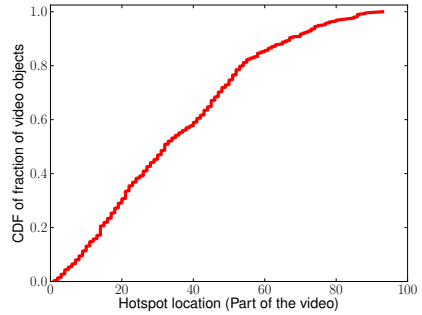
Figure 15: *Two extreme examples of temporal change in interest in live content during the duration of the event*

users to sample videos to consider a (possibly counter-intuitive) option where we can use P2P to bootstrap serving the content and later use the CDN servers. This allows the CDN to invest resources more usefully for viewers who are more likely to yield increased revenue from ad-impressions.

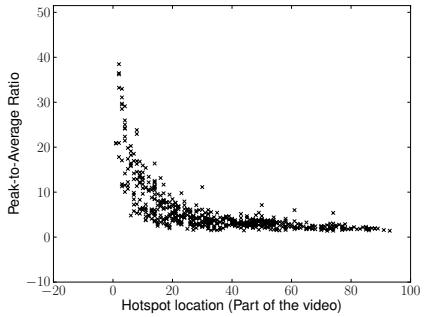
Methodology: The metric of interest here is the reduced number of accesses at the CDN server as a result of using P2P. Our goal in this exercise is to evaluate the *potential benefit* of P2P-assisted CDNs. To this end, we consider a very simplistic P2P-assisted model where peers are organized in a swarm with a specific *scope* and *size*. For each swarm, we assume that only one request needs to go to CDN server and the remaining nodes receive the content through P2P. The scope represents a trade off between network-friendliness and the availability of peers who are in synchronized viewing:

- Nodes form peers only with nodes within the *same city+ISP*.
- Based on the region classification in Table 1, nodes form peers only with other nodes within the *same region+ISP*.
- Nodes form peers within the *same region*

Because our goal is to estimate the potential benefit, we do not model swarm dynamics or evaluate the choice of chunk selection policies [23]. We consider a simple sequential chunk selection policy where peers are organized in swarms corresponding to the location in the video. For live content we do not consider the impact of cache size since all viewers are in sync. However, for VOD we cache a limited number of previous chunks (e.g., several services like Netflix do not allow caching more than a few minutes of the content) and nodes typically peer with other nodes that have the required content cached. We set the chunk size to 5 minutes of the video consistent with what is predominant in the industry. We limit the maximum swarm size to 100.



(a) Where do hotspots occur?



(b) How pronounced are the hotspots?

Figure 16: *Investigating hotspots in live content*

Scope	Live (%)	VOD (%)
Same region	98.94	87.09
Same region+ISP	96.91	40.90
Same city+ISP	92.65	13.79

Table 3: *Overall benefit for using P2P for different scopes*

Impact of varying scope: Table 3 summarizes the overall benefit from using a P2P-augmented CDN system for live and VOD content. Not surprisingly, live content has higher savings than VOD. For live, the potential savings are as high as 92% even with same city+ISP scope. While the benefits are higher as we increase the scope, the resulting increase in savings shows diminishing returns. This suggests that realizing simple and network-friendly P2P solutions to augment today's CDN infrastructure is a promising path. In the case of VOD, we limit the cache size to 1 chunk. We observe that savings can be as high as 87% when nodes are allowed to peer with other nodes within the same region. Unlike live content, however, the savings are not as large when the scope of peering is limited (e.g. same city + ISP).

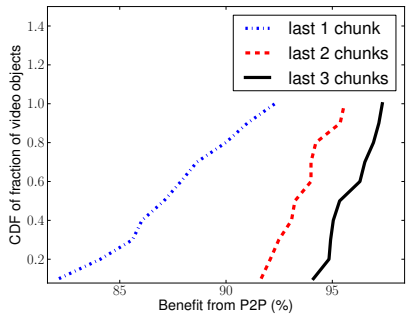


Figure 17: *Impact of cache size on the benefit of P2P for VOD*

Effect of varying cache size for VOD: In the case of VOD, it is interesting to investigate how increasing cache size affects the performance of the system. Figure 17 shows the CDF of the benefits from P2P for different VOD objects with same region scope. Although increasing cache size leads to greater savings, we observe diminishing returns for increased cache size.

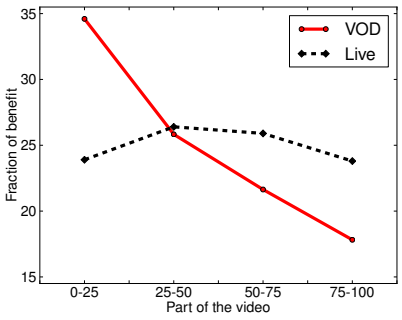


Figure 18: *Chunks that are more likely to benefit from P2P; for VOD we see that the early chunks are the ones that benefit the most*

Which part of the video gives most benefit? In Figure 18, we look at the percentage of savings that chunks in different parts of the video provide. We observe that most of the benefits for VOD is due to the earlier chunks. This is because users typically watch VOD videos from the start and the large number of early-quitters cause the earlier chunks to be more available than the later ones. However, in the case of live, the benefits appear to be more uniform. This is because of the pattern that we observed earlier—although users quit early, they also join the event at random times.

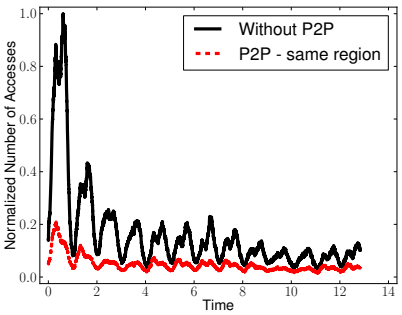


Figure 19: *Evolution of the benefit of P2P assistance over time*

How does the benefit of P2P vary over time: Unlike live events, VOD objects have demands that last several weeks. In this context, it is interesting to observe how the savings vary with time. Figure 19 shows the temporal variation in access demands at the server with and without P2P. We observe that the savings are as high as 80% during the peak access hour on the day of release because of larger number of users synchronously viewing content. This is also the time when the CDN would benefit the most from savings.

Using P2P earlier: Last, we explore the benefits via an alternative strategy of using P2P for the early chunks and later serving the content directly from the CDN. This can be a viewed as a mechanism to filter out the early quitters and serve them without wasting precious server resources. We analyze the benefits of serving only the first few chunks using P2P for both live and VOD in Figure 20. We observe that with about 2 chunks (which covers most of the early-quitter scenarios for VOD), we can get savings of around 30%. In

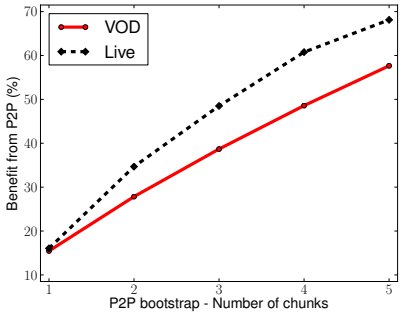


Figure 20: *Using P2P in the early stages of user arrival*

the case of VOD, this is almost equivalent to the savings obtained from the first 2 chunks of the video (as in Figure 18) since users typically watch the video from start. In the case of live, serving the first four chunks since the user starts the session (which covers all of the early-quitters) results in around 60% savings. However, note that this is not the same as the savings from the first 4 chunks of the video since users join at random times during the event and start viewing from random parts of the event.

Performance Costs: Using P2P to bootstrap video delivery might have an impact on the time it takes for the video to start playback (*start up delay*) [29]. Also, higher node churns in the P2P swarm can potentially result in disruptions in the video playback. System designs to circumvent such performance issues have been studied in previous work [37] and is not the focus of this paper.

5.3 Main observations

To summarize, the key observations are:

- VOD has more synchrony in viewership than expected, especially during the peak access hours on the day of release. This is also when we observe the highest demand for the object. Hence, P2P can be used to offload some of the load from the server during peak hours. We observed around 87% savings for P2P-assisted VOD.
- We explore the option of bootstrapping using P2P as a means of "filtering out" early-quitters for both VOD and live and see that this alone could lead to 30% savings in the case of VOD and 60% savings in the case of live.

6. CONCLUSIONS

As Internet-based video consumption becomes mainstream, the video delivery infrastructure needs to be designed and provisioned to deliver high-quality content to larger user populations. But current trends indicate that the CDN infrastructure is being stressed by the increasing video traffic. Telco-CDN federation and hybrid P2P-CDNs are two oft-discussed strategies to augment existing infrastructure, but there are no recent studies on the benefits of these two strategies. Given the ongoing industry efforts and discussions to deploy federated and P2P-based solutions, we believe our work is timely: we provide a quantitative basis to justify, motivate, and inform these initiatives.

Our analysis of over 30 million live and VOD sessions reveals several interesting access patterns that have important implications to these two strategies including regional and time of day effects, synchronous viewing behavior, demand predictability, and partial interest in content. Building on these observations, we analyzed the potential benefits of hybrid P2P-CDN approaches and telco-CDN federation. We found that federation can signiïňĄcantly reduce telco-CDN provisioning costs and equivalently increase the effective capacity of the system by exploiting regional and cross-ISP skews in access popularity. Surprisingly, we found that P2P approaches can work for VOD content as well, especially at peak loads when we see highly synchronous viewing patterns, and proposed and evaluated new strategies for hybrid-P2P systems based on prevalent user behavior.

Acknowledgments

We thank our shepherd Kuai Xu and the anonymous reviewers for their feedback that helped improve the paper. We thank Conviva Inc. for making the video viewing data available for our study. We also thank the staff at Conviva for answering several questions about the dataset and the data collection infrastructure. This work is partially supported by the National Science Foundation under grants CNS-1050170, CNS-1017545, CNS-0905134 and CNS-0746531

7. REFERENCES

[1] Akamai NetSession. http://www.akamai.com/client/.

[2] Census Bureau Divisioning. http://www.census.gov/geo/www/us_regdiv.pdf.

[3] Cisco forecast. http://blogs.cisco.com/sp/comments/cisco_visual_networking_index_forecast_annual_update/.

[4] Cisco Report on CDN Federation - Solutions for SPs and Content Providers To Scale a Great Customer Experience.

[5] Hadoop. http://hadoop.apache.org/.

[6] Mail service costs Netflix 20 times more than streaming. http://www.techspot.com/news/42036-mail-service-costs-netflix-20-times-more-than-streaming.html.

[7] NFL What are HQ videos? http://www.nfl.com/help/faq.

[8] Use RTMFP for developing real-time collaboration applications. http://labs.adobe.com/technologies/cirrus/.

[9] WebRTC 1.0: Real-time Communication Between Browsers. http://www.w3.org/TR/webrtc/.

[10] B. Niven-Jenkins, F. L. Faucheur, and N. Bitar. Content distribution network interconnection (CDNI) problem statement. http://datatracker.ietf.org/doc/draft-ietf-cdni-problem-statement/, Jan. 2012.

[11] L. Plissonneau and E. Biersack. A Longitudinal View of HTTP Video Streaming Performance. In *Proc. MMSys*, 2012.

[12] H. Abrahamsson and M. Nordmark. Program popularity and viewer behavior in a large TV-on-Demand system. In *IMC*, 2012.

[13] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. K. Ramakrishnan. Optimal Content Placement for a Large-Scale VoD System. In *Proc. CoNext*, 2010.

[14] B. Cheng, L. Stein, H. Jin, and Z. Zheng. Towards Cinematic Internet Video-On-Demand. In *Proc. Eurosys*, 2008.

[15] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a Predictive Model of Quality of Experience for Internet Video. In *SIGCOMM*, 2013.

[16] C. Huang, A. Wang, J. Li, and K. W. Ross. Understanding Hybrid CDN-P2P: Why Limelight Needs its Own Red Swoosh. In *Proc. NOSSDAV*, 2008.

[17] C. Huang, J. Li, and K. W. Ross. Can Internet Video-on-Demand be Profitable? In *Proc. SIGCOMM*, 2007.

[18] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. IMC*, 2007.

[19] D. Rayburn. Telcos and carriers forming new federated cdn group called ocx (operator carrier exchange). http://goo.gl/wUhXr.

[20] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. A. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. In *Proc. SIGCOMM*, 2011.

[21] J. Erman, A. Gerber, K. Ramakrishnan, S. Sen, and O. Spatscheck. Over the top video: The gorilla in cellular networks. In *IMC*, 2011.

[22] H. Y. et al. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics. In *Proc. IMC*, 2009.

[23] B. Fan, D. Andersen, M. Kaminsky, and K. Papagiannaki. Balancing Throughput, Robustness, and In-Order Delivery in P2P VoD. In *Proc. ACM CoNEXT*, 2010.

[24] A. Finamore, M. Mellia, M. Munafo, R. Torres, and S. G. Rao. Youtube everywhere: Impact of device and infrastructure synergies on user experience. In *Proc. IMC*, 2011.

[25] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *Proc. Eurosys*, 2006.

[26] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross. A measurement study of a large-scale P2P IPTV system. *IEEE Transactions on Multimedia*, 2007.

[27] Y. Huang, D.-M. C. Tom Z. J. Fu, J. C. S. Lui, and C. Huang. Challenges, Design and Analysis of a Large-scale P2P-VoD System. In *Proc. SIGCOMM*, 2008.

[28] S. Krishnan and R. Sitaraman. Video Stream Quality Impacts Viewer Behavior: Infering Causality Using Quasi-Experimental Designs. In *IMC*, 2012.

[29] B. Li, S. Xie, Y. Qu, and K. G.Y. Inside the New Coolstreaming: Principles, Measurement and Performance Implications. In *INFOCOMM*, 2008.

[30] Z. Li, J. Lin, M.-I. Akodjenou-Jeannin, G. Xie, M. A. Kaafar, Y. Jin, and G. Peng. Watching video from everywhere: a study of the pptv mobile vod system. In *IMC*, 2012.

[31] H. Liu, Y. Wang, Y. R. Yang, A. Tian, and H. Wang. Optimizing Cost and Performance for Content Multihoming. In *SIGCOMM*, 2012.

[32] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang. A Case for a Coordinated Internet-Scale Video Control Plane. In *Proc. SIGCOMM*, 2012.

[33] T. Mitchell. *Machine Learning*. McGraw-Hill.

[34] T. Qiu, Z. Ge, S. Lee, J. Wang, Q. Zhao, and J. Xu. Modeling channel popularity dynamics in a large IPTV system. In *Proc. SIGMETRICS*, 2009.

[35] R. Powell. The federated cdn cometh. May 2011. TelecomRamblings.com.

[36] S. Guha, S. Annapureddy, C. Gkantsidis, D. Gunawardena, and P. Rodriguez. Is High-Quality VoD Feasible using P2P Swarming? In *Proc. WWW*, 2007.

[37] H. Yin, X. Liu, T. Zhan, V. Sekar, F. Qiu, C. Lin, H. Zhang, and B. Li. Design and Deployment of a Hybrid CDN-P2P System for Live Video Streaming: Experiences with LiveSky. In *Proc. ACM Multimedia*, 2008.

[38] R. Zhou, S. Khemmarat, and L. Gao. The impact of YouTube recommendation system on video views. In *Proc. IMC*, 2010.