Temporal Scalable Live Video Streaming over Hybrid CDN-P2P Architecture

Anahita Fellah Jahromi
Department of Computer Science and Eng.
Sharif University of Technology, International Campus
Kish Island, Iran

ABSTRACT
Most of currently deployed live video streaming systems don’t consider heterogeneous bandwidth of peers. To address this issue, many researches combine P2P solution with scalable video streaming, however none of the existing hybrid CDN-P2P consider it. This paper presents the combination of Hybrid CDN-P2P and temporal scalable streaming. The main design goal of our approach is changing requesting and transmission policy of peers commensurate with bandwidth to optimize video distribution and frame losses. Heterogeneity is addressed by changing upload policy of peers to send different layers for neighbors. The experimental result shows that, it achieves significant improvement to decrees video distortion and hop count.

Keywords
Hybrid CDN-P2P, Video layered, Temporal scalability, Live video streaming, Bandwidth heterogeneity

1. INTRODUCTION
In recent years, live video streaming approaches have gained popularity in video distribution over the Internet. With widespread adaptation of video application by users; scalable and cost effective approaches for delivering video with high quality have remained interesting and challenging problem. Since IP multicasting has not been deployed on the internet, two main architectures for video streaming at application layer exist; Content Distribution Network (CDN) and Peer to Peer (P2P).

CDN network is an overlay of geographically distributed servers (replica server) over the Internet. In traditional client-server model, clients suffer from delays by physical distance between client and source streaming server. In CDN model instead of one server, servers are placed at the edges of Internet. Clients connect to a central server and redirect to nearest server to get video directly with shortest start up delay. However, scalability is the main problem of this approach. By increasing number of users more resources (capacity and bandwidth) are required to provide video with acceptable quality, more costs is imposed for CDN service provider.

P2P is a new paradigm which solves scalability problem of CDN approach. Peers’ resources use to reduce resource utilization of servers. In this decentralize method, each user works as both client and server. In a mesh system, there is no specific topology for peers, peers contact to a tracker and send their information to it, and the tracker returns a list of active peers. A peer tries to make neighbor ship with peers mentioned in the list. When a peer sets up a connection with neighbors, it exchanges data availability by a buffermap and requests video chunk based on it. However, this solution has gained scalability in the cost of losing shorter start up delay and stable video quality of CDN architecture.

Hybrid CDN-P2P architecture has been introduced to combine scalability advantages of P2P as well as reliable and shortest start up delay transmission of CDN systems. While CDN-P2P architecture has been proved more efficient than other architectures for live video streaming it is still a question, how it is possible to be more efficient in providing higher quality of service (QoS).

A simple definition of temporal scalability is, providing different frame rate for delivering of video. Some video frames discard from a GOP to provide basic frame rates in base layers, and this layer can encode independently from enhancement layers. Enhancement layers consist of remaining video frames from the GOP but these frames depend on video frames in the base layer and are encoded with reference temporal prediction in the base layer. Heterogeneous bandwidth has an important role in CDN-P2P systems. Lack of bandwidth is the main reason of video transmission losses and video quality degradation in this architecture. The idea behind temporal scalable video coding is, allows for adaptation to different bandwidth of peers and helps to provide better quality for peers due to heterogeneity of bandwidth. Although scalable video streaming has been proposed for many systems individually or combination with P2P streaming, but integration with Hybrid CDN-P2P architecture to improve quality is a challenging task.

Main design goals of our proposed method are; provide important layer (basic layer) for all peers and change requesting and transmission policies of peers in regard to peer’s upload bandwidth to provide better quality for system. To achieve these, three version of video produced in our method, each peer send one version of video to their neighbors. When a peer hasn’t enough bandwidth to provide all layers, it should change its transmission policy and just transmit missing chunks from the lower layer. On the other hand when a peer has enough bandwidth to serve all layers, it should transmit all missing chunks.

Following this introduction, a review of related works on hybrid CDN-P2P together with Scalable Video Coding in P2P system presents. In section 3, we propose some methods that enhance the hybrid CDN-P2P protocol with an efficient temporal scalability method. Section 4 contains comparative studying on connected mesh with layered method and without it based on evaluation criteria which are important for live video streaming.
2. RELATED WORK
To meet the requirements of live media streaming, hybrid CDN-P2P architectures have gained popularity in research community.

D. Xu et al. proposed a hybrid CDN-P2P architecture for distribution of on demand video streaming [1]. This architecture is cost effective because restricts bandwidth utilization of CDN servers. A CDN server has two main roles in the system: media server and P2P index server.

When a media file is requested to distribute, at first it distributes to requester peers just by a CDN server. Now requested peer becomes supplier peers, the CDN server registers supplier peers and their contributions in the system. After that both of CDN and P2P contribute to video distribution, now the CDN node uses second roles as P2P index server. The CDN node checks supplying peer in index server, if they could serve to new arrival peer with acceptable streaming rate, the request serves by subset of peers, in this part. If not so, the CDN node serves to new requester itself. When numbers of peers which contain the media file increase, P2P network can work properly without wasting capacity of the CDN node; so it stops serving and capacity of the CDN node is released for new file. This process is called handoff. This scenario can extend to all CDN servers in system.

X.liu and et al. [2] have used a real deployment of a commercial CDN companies in china (Chinacache). This architecture consists of streaming source site, edge servers that join by multi tree topology and users that can use traditional client-Server or P2P applications for streaming. Clients or peers are divided into edge regions; they are redirected by SLB (Server load balancing) system to one of the closest edge server located in ISP according to DNS address. Scalability of the system to decrees cost of CDN server is controlled by push-num(x) function, the maximum amount of x is where edge server reach to a threshold, in this situation system is scalable but quality degradation observes. So there is a tradeoff between quality and scalability of system. By this locality aware approach, multiple district mesh under each edge server creates but peers under each mesh cannot use upload bandwidth of peers in other meshes.

CDN-P2P with connected mesh [3] has been designed to solve above problem by connecting these independent meshes to each other. When a peer wants to join the network, it contacts to tracker to get list of random neighbors. The tracker checks Connected Mesh parameter, if this parameter becomes true, it means that new arrived peer doesn’t connect to two CDN servers directly. Tracker returns neighbor from other meshes to this node after a period of time, this period is called connection satisfier. After that a single big mesh under all CDN servers creates. Peers use Scheduling algorithm like [4] to requesting chunks from neighbor. They send all requesting frame to one supplier by a single message.

The paper compared the performance of the CDN-P2P connected mesh to CDN-P2P unconnected mesh [2] and pure P2P system based on evaluation parameters. The results have shown that their method works better in distortion and end to end delay, since the method connects more peers to each other and peers provide more resources for the network. As a result, resource of peers (bandwidth) is used more efficient than other methods to distribution of video in CDN-P2P connected mesh. However, it has less start up delay in compared to unconnected mesh but this time is tolerable.

Most of the currently hybrid CDN-P2P system do not consider scalable video streaming. When the requirements and conditions differ among peers, all of these architectures send single version of video to all peers. Present scalable streaming methods try to combine scalable streaming with P2P network. Heterogeneity of peers in a P2P system is caused; maintenance of QoS in a certain level becomes a challenging problem.

The Proposed method in [5] has solved the problem by adding scalable video streaming to P2P network. Authors have proposed a method which works based on Bittorent [6] system by considering requirements for live video streaming. This architecture considers spatial layers of video; the spatial layer can be located in temporal or quality layers. At first, video is divided into chunks; each chunk is divided into layers. Base layer is first layer which provide acceptable quality, higher layers refine the quality with more download time, and finally, they send to seeder node (it has copy of all chunks and layers) to distribute them.

When a peer joins to this network, it contacts to tracker and asks information about peers in the same session, the tracker returns list of neighbors. Leecher nodes (just has some part off video or chunk) download video content from seeder node in Bittorent [6] network if it can get all layers of the chunk, it can act as re-seeder. After a peer gets all layers or some parts of it, it assembles them. Depending on number of layers that leecher node can obtain, quality of video is specified (more layers provide better quality). Playing starts after these layers assemble in player buffer. They use strategy which is called PSW (Priority Sliding Window) algorithm to get the data in order.

In Layer P2P method [7] authors have considered different types of peers; a Peer with more contributions in P2P overlay receives better quality. It uses layered video, it encodes video into L layers, and each layer consists of packets. Layered Chunk’s (LC) distribute in P2P overlay. Layer P2P uses mesh pull system, but here each peer maintains L buffers, each buffer for each layer. At first similar to single layer peers request for their missing LC’s. The supplier gives priority to LC’s request of neighbors, according to it for tat strategy; peers with higher contribution receive better quality.

Differentiated services provide with respect to heterogeneity bandwidth of peers; when the system has enough bandwidth, each peer can use all layers and watch the video with high quality. When the bandwidth of system is insufficient, users commensurate with their upload receives quality. When upload bandwidth is lower than video rate, peer receives different quality according to their upload bandwidth. Layer video coding accomplishes by providing multiple versions of video. It encodes video in multiple layers, higher layers have better quality.

Among the Hybrid CDN-P2P solutions, we take up our option on connected mesh CDN-P2P [3] while results have shown better performance in compare to other methods. However we modify it in order to comply with temporal scalable requirements. On the other hand most of scalable streaming are implement in P2P system. We employ multiple layers over the CDN-P2P system.
3. THE PROPOSED TEMPORAL SCALABLE METHOD OVER HYBRID CDN-P2P

3.1 System Assumption
We assume that peers can download whatever they want, but have finite upload bandwidth. Peers contribute all of upload bandwidth and are cooperative (healthy system). Since a peer has multiple uplinks and one downlink, these assumptions are common in the most of papers and don’t cause bottleneck in the downloading of the video. However, peers cannot serve neighbors more than what it downloaded and the network has finite capacity. We also assume that size of each chunk is equal to a frame, similar to [3].

3.2 Overview
Peer’s bandwidths differ extensively in the network. If they don’t take into account in this system, it leads to transmission loss and quality degradation in the system.

To solve above problem, the idea behind scalable video streaming comes of help by adaptation to different uplink of peers and helps to provide quality for peers commensurate with heterogeneity of bandwidths. Two bit streams generate; one carrying the most vital information and called base layer, remaining carry more information to improve quality of the base layer and called enhancement layer. We provide the capability between peers to change transmission policy to send layers in regard to their bandwidth. On the other hand neighbors change requesting policy to just request frames of lower layers for neighbors with narrower bandwidth.

Basic layer is the most important layer. With single layer video, video content loss that is belong to Intra (I) or Inter (P) frames not only affect on the video frame it belongs to, but also affect on the successive Bidirectional frame (B). So error propagation due to forward prediction is caused severe video quality degradation.

In the proposed method, we try to increase the number of I and P frames for two main reasons; improve video quality of system and provide a basic quality for all peers in system. There are three versions of video and three frame rates; frame rate is specified commensurate with heterogeneous bandwidth of peers. Peers with lowest upload bandwidth send only base layer, peers with higher bandwidth increase frame rate and send one or two enhancement layers. For a peer basic layer is transmitted by all neighbors and two remaining enhancement layers are transmitted by neighbors with higher bandwidth.

By this approach, video distortion decreases and all peers can watch video with higher quality without considering their upload bandwidth, on the other hand peers send layers according to their upload bandwidth.

3.3 Description of the Proposed Method
In the method, CDN servers form a single tree topology. Source node performs video encoding; camera module in this node sends the video frames to peers. Since the CDN part optimized in commercial approach, we are not working on this part and it is same as [1].

3.3.1 Tracker
Managing of upload bandwidths are commensurate with frames rate and they are not caused quality degradation in event of lack of bandwidth.

In the method, a tracker divides upload bandwidth of peers into three range of numbers, group three is for highest upload bandwidth, group two for medium upload bandwidth, group one for narrower upload bandwidth. It compares amount of upload bandwidth to the three ranges and adds it to each group according to this amount. Table 1 shows the range of three groups of upload bandwidth.

<table>
<thead>
<tr>
<th>Type</th>
<th>Range of Upload bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>128 Kbps-256 Kbps</td>
</tr>
<tr>
<td>Group 2</td>
<td>256Kbps-768 Kbps</td>
</tr>
<tr>
<td>Group 3</td>
<td>&gt;768Kbps</td>
</tr>
</tbody>
</table>

New arrival peer has address of the tracker; it requests neighbors in the same video session from tracker and reports its own information including upload bandwidth to the tracker. The tracker has two more responsibilities:

1. The tracker compares amount of upload bandwidth of a new arrival peer with three ranges of numbers and map it to one group.
2. The tracker return a list includes random neighbors but it also sends one neighbor from group one; we will describe the reason of this function in the following.

When the peer gets neighbor list, it also checks connection satisfier parameter to prevent connecting to two CDN servers directly. After that it joins to a single big mesh and sends notification messages to the tracker to inform it about remaining capacity to getting new neighbors.

3.3.2 Mesh construction and data exchanges
The peer tries to make connection with neighbors and sends a JOIN-REQ message to them. If a neighbor can accept the request based on upload bandwidth, it returns JOIN-RSP message to the peer. The peer sends JOIN-ACK to neighbors and adds it to neighbor list, otherwise returns JOIN-DNY message. After getting enough neighbors, it starts to exchange video content to them. Figure 3 shows process of mesh construction.

Peers which are connected to CDN servers directly, can get video with high quality from CDN servers. In the method each peer delivers video content to neighbors according to its upload bandwidth. Nodes with narrower bandwidth cannot send all missing chunks to their neighbors because frame losses as explained in previous section, induce video quality degradation. These nodes exchange video content with lower rate for example base layer or base layer and one of the enhancement layers to the neighbors. In Mesh pull system, peers exchange chunk availability with neighbors by buffer maps, buffer maps are streams of zeros and ones which indicate frames which currently has cached in the buffer.
If a peer belongs to Group 1 with narrower bandwidth, it should send only the base layer to its neighbors thus its buffer map should indicate frames belong to base layer. If this peer sends more frames, video quality degradation occurs as described in the previous section. So it is better to send more important frames to prevent error propagation for B1 and B2 frames and provide basic quality for neighbors. The peer inserts zeros for all B1 and B2 frames, without considering its availability in buffer.

If a peer belongs to Group 2 with higher bandwidth than Group 1 member, it is able to send more frames to neighbors. It should send the base layer (I and P) and one of the enhancement layer (first B we called them B1). It inserts zeros for all B2 (Second B) frames, without considering its availability in buffer.

If a peer has enough bandwidth to send full rate video to neighbors, it should send all frames (I, P, B1 and B2).

By this approach, peers change transmission policy to send layers which by current bandwidth can transmit. On the other hand, buffer maps are modified by suppliers, it could be said that requesting policy of neighbors is changed by suppliers.

Due to asymmetric characteristic of network and prevention from bottleneck in downloading chunks, we assume that all peers have infinite download. Thus, it is necessary to provide all frames (I, P, B1 and B2) for them and they can download it based on their download bandwidth. The tracker chooses neighbors randomly, if peers haven’t neighbor from Group 1 they cannot get B2 frame from neighbors. To prevent this problem, tracker should return one neighbor from Group 1 to all peers.

4. EVALUATION

In this section the proposed method being implemented on selected Hybrid CDN-P2P system [3]. The Connected mesh with layered method and without it [3] are compared based on evaluation criterions for live media streaming such as Distortion, Hop count, Startup delay and End to End delay.

4.1 Simulation Environment

Hybrid CDN-P2P [3] architecture is implemented in OMNET++ v.4 [8]. OMNET++ is a modular discrete event network simulation framework. INET framework [9] is a collection of network units; it contains models for Internet protocols such as IP, TCP, UDP and Data Link layer which work in OMNET++. OverSim [10] is an open source framework for implementation of P2P network in OMNET++; it contains models for unstructured network (Mesh network). Connected CDN-P2P architecture is implemented in Denacast [11]. Dena cast is used as a hybrid CDN-P2P to simulate the proposed method.

4.2 Experimental Scenario

The simulation is implemented by using actual video streams, a video trace file from Video Trace Library in [12]. The setup of our simulation is as follows; the result of simulation is based on Star war IV streams, this file is chosen because of variable bit rate is a range of 25 to 800 Kbps. Each video has 25 FPS (Frame Per Second). The video encodes in “[I|B1|B2|P|B1|B2|B1|B1|B2|B2|B2|B2|B2|B2|B2|B2|B2|B2|B2]” structure, one GOP includes 12 frames. Table 2 illustrates characteristics of video used in scenario.

| Table 2. Characteristics of video used in experiment |
|---------------------------------|-----------------|
| Video Trace file                | Star Wars IV, MPEG 4 Part 2 |
| Video average bit rate          | 256 Kbps         |
| Maximum video bit rate          | 800 Kbps         |
| Video FPS                       | 25               |

Assumptions which described in the previous section are considered in the scenario:
1. Peers contribute all of upload bandwidth and are cooperative (healthy system)
2. CDN servers don’t leave system
3. Peers cannot upload more than what downloaded

Large number of heterogeneous upload bandwidth are considers. The upload bandwidth value of peers are chosen according to overall distribution which are based on actual measurement studies in [13], however the measurement considers the contributed bandwidth of peers. Since assumptions don’t consider peers’ contribution, we have changed these values to narrower bandwidth to create a scenario and get real results. This distribution is given in Figure 2. We set upload bandwidth of nodes 128 Kbps up to 1600 Kbps.

![Figure 4. Upload bandwidth Distribution of peers](image-url)
In the simulation physical topology is generated using Georgia Tech Internet Topology Model (GT-ITM) [14] tools for OMNeT++ v.4 with 28 AS (backbone routers) and 28 access routers per AS in top-down mode. A peer selects a router randomly and connects to it by a random physical link. Each peer selects 5 random neighbors and 1 neighbor from Group 1. Neighbors exchange buffer maps every 1 second. Each Window of Interest includes 120 segments, each one 1 second. Video file is divided into chunks; in our simulation we assume that size of each chunk is equal to a frame. Table 3 shows rest of simulation parameters.

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminal number</td>
<td>150</td>
</tr>
<tr>
<td>Number of neighbors</td>
<td>6</td>
</tr>
<tr>
<td>Window of Interest</td>
<td>120 s</td>
</tr>
<tr>
<td>Buffer map exchange</td>
<td>1 s</td>
</tr>
<tr>
<td>Simulation Time</td>
<td>200 s</td>
</tr>
<tr>
<td>Average chunk Length</td>
<td>11011 bit</td>
</tr>
<tr>
<td>Chunk size</td>
<td>1 Frame</td>
</tr>
<tr>
<td>Packet Unit</td>
<td>240 Byte</td>
</tr>
<tr>
<td>Start up buffering</td>
<td>15 s</td>
</tr>
</tbody>
</table>

4.3 Comparison and Result Analysis

The result in [3] shows Connected mesh improved Distortion and End to end delay in compared to Pure P2P and unconnected CDN-P2P [2]. However, Connected CDN-P2P sends video content with full rate in single layer without considering different upload bandwidth of peers. So, we choose it as a single layer method and compare our proposed method with it based on performance metric. In the following, performance metrics based on upload bandwidth distribution in the Figure 3 will be present. The proposed method is evaluated by some of important Quality of Service (QoS) parameters for live video streaming. These metrics are startup delay, Video Distortion, Hop Count and End to End delay.

4.3.1 Startup delay

In order to maintain continues playback, a peer buffers number of frames before playback can be started. Startup buffering is time interval between a peer buffers the video and playback starts. Startup delay is a time interval between a peer decides to connect to a video session in mesh and playback starts. Startup buffering could be set in configuration; to come up with realistic scenarios as [15] reports startup buffering generally is around 10 to 15 seconds. This time is set in 15 seconds in the simulation; it means client buffer size to 15 seconds is reduced [16]. The startup delay depends on start up buffering; it means that a peer for buffering 15 seconds of video how much time is taken before play back can be stared.

We plot the result for CDN-P2P with single layer and with layered in the Figures 5 and as CDF (Cumulative Distribution Function) with startup buffering of 15 seconds. The result shows that the proposed method on average works same as single layer. It shows 50% of nodes can buffer 15 seconds of video in 15 second which is better than 17 seconds in CDN-P2P. The same behaviors can observe for the 70%, peers can receive 15 seconds of video from neighbors faster than single layer approach, because frame distortion by layered approach decrease as we will present in next section in more details. For remaining 20% of peers, delay will become doubled because the number of Group 3 members which can be neighbor with Group 2 and Group 1 decrease.

4.3.2 Video Distortion

Video Distortion is a percentage of video content that is lost during transmission. We plot the result for CDN-P2P with single layer and with layered in the Figure 6 as a CDF (Cumulative Distribution Function). The Figure shows that the proposed method works better on average rather than single layer. For example, less than 90% of peers obtain less than 10% distortion in compare to 40% in single layer. Distortion is an important metric for video quality. Loss of a frame is not only caused frame destruction but affected on successive frames and finally error propagation.
rate according to their upload bandwidth and finally reduction in distortion observes.

Group3 neighbors of a peer can reduce effect of removed frames by Group 1 neighbors, so the overall quality increases by layered approach. The Figure shows 10% of nodes have 10% to 35% distortion. The lack of Group3 members to be neighbor with Group1 and 2 members in network is caused, growth of distortion for these nodes. The maximum distortion of layered CDN-P2P is less than 35% in compared to 75% in single layer CDN-P2P.

4.3.3 Hop Count
Hop count is the number of hops that a video unit passes from a source node to reaches at destination nodes. For example, this value for nodes that are directly connected to CDN servers is equal to 1. The Figure 8 shows that with layered approach, all peers receive the video after 10 hops. This value in single layer is 14 hops.

More hops indicate that frames from a source node should pass longer paths and different peer’s buffer to reach at destination node. In the proposed method, frames have more ways to reach at destination from servers. Two factors affect on it: Firstly, I and P frames have more paths to reach a peer, because member of all groups supply them and can send them without loss. Secondly, in the mesh formation part we assume that all peers have one neighbor from Group 3. Thus, each node has at least one neighbor that can provide B1 and B2 frames. So, frames don’t need to pass long route to reach destination.

4.3.4 End to End delay
End to end delay is defined as time interval between frame creation in the source nodes and playing at destination node. This metric is one of the most important metric for live video streaming such as IPTV. For example in the broadcast of the soccer play, this represents the delay between when an action of a goal occurs and the time when it receives and playing start at destination [17].

When number of nodes in the network increase, it increases too. The result for 70% of peers show, this time is less than 30 seconds, and is near to single layer. But, it increases for remaining 30% of peers.

As explained in the Hop count section, routes for frames to reach at destination node decrease. But, the time interval to reach at destination for 50% of nodes in our proposed method is more than single layer, because 35% of peers are member of Group 2 and Group1; these members have extra overhead before sending their buffer maps. Modifying buffer map’s structure does incur overhead and introduces additional delays for sending frames to neighbor. Figure 8 shows end to end delay.

While 50% of peers receive frames and played it before 40 seconds, remaining 20% have to wait longer time. This time is tolerable, because it is just a few seconds. However, there is a tradeoff between quality and delay. These peers get the video content with more delay but better quality.

5. CONCLUSION
This paper described the temporal scalable method over hybrid CDN-P2P connected mesh. We try to change transmission policy for heterogeneous peers to provide higher quality. The method provides basic quality for all peers; a peer achieves higher quality by sending more layers regard with heterogenous upload bandwidth of neighbors. Our simulation showed that by applying our method, distortion and hop count significantly reduced, and startup delay are same as single layer. The proposed method is slightly weaker than single layer method in End to end delay, but, it can satisfy our purpose to provide higher quality. The system can be further enhanced by adding an incentive mechanism to send layers based on nodes’ contribution in the network. We can also extend the method by providing roles for CDN servers to distribute layers in hybrid CDN-P2P system.

6. REFERENCES


