CINAPS: A Scalable Solution for VoD over WLANs Using Application-Friendly Association Control

Leonardo Bidese de Pinho and Claudio Luis de Amorim
Parallel Computing Laboratory - Computer and Systems Engineering Program
COPPE - Federal University of Rio de Janeiro
Email:{leopinho, amorim}@lcp.coppe.ufrj.br

Abstract

In recent years, the delivery of video-on-demand (VoD) applications to mobile users has gained increasing attention by the wireless communication industry. In this paper, we introduce a scalable VoD system operating over 802.11 networks, namely CINAPS (Cluster of INexpensive Access PointS), in which collocated commodity Access Points (APs) work in no-overlapping channels orchestrated by the Wireless Channel Manager (WCM), a central unit that performs association control of clients to the APs. Key to CINAPS' scalable performance is the support to two novel application-friendly heuristics we developed, namely Early Released First (ERF) and Bounded ERF (BERF) for association control that guarantee minimum bandwidth through effective reservation mechanisms. The main insight behind ERF and BERF is to exploit the slack of time aka the patience factor that VoD applications provide, to allow the association control unit to accept future requests for using the communication channels. Our simulated results showed that CINAPS outperformed our previous design that used a standard LLF (Least Loaded First) heuristic with minimum bandwidth guarantees, by decreasing significantly the overall blockage rate of video requests.

Keywords: scalable VoD, wireless networks, collocated access points, association control heuristics

1. Introduction

In recent years, the delivery of video-on-demand (VoD) applications to mobile users has gained increasing attention by the wireless communication industry. In particular, there exists a growing need to offer ubiquitous access to video playback, as more and more people wish to have access to such applications at their convenience. The use of wireless networks to video delivery can be the answer to such VoD systems provided that they can support an increasing number of simultaneous user accesses in a scalable manner, while guaranteeing quality of service (QoS) by assuring fast response time and jitter-free video presentation to their clients

From the network perspective, the delivery of video streams on demand can be seen as a regular application whose typical operation consists of three well-defined phases: (i) a client issues a video request from a list of available video titles; (ii) the chosen video stream[1] is transmitted to the client with bit rate equal to the average consumption rate of the client's video decoder; (iii) the network bandwidth[2] the VoD system allocates for transmitting the video stream remains reserved as long as the client watches the video. Here and thereafter, we will refer to client, user, and client machine, as simply the client.

Hypothetically, in true VoD systems client requests are serviced immediately and video packets are transmitted from the server to the client through a jitter-free, zero-latency network. In real systems, however, clients always experience some delay between the video request and the beginning of the video playback. Such a delay or playback latency can be broken down into three components: 1) the service latency, which is the elapsed time between the request and finishing the schedule of resources to delivery the video stream to the client; 2) the network latency, which is the elapsed time the first video block takes to go over the network until it arrives at the client; and 3) prefetch latency, which is the elapsed time to fill up the playout buffer at the client in order to deal with variations on both network latency and

[1] Smoothed through a playout buffer at the client, which allows for variable bit-rate (VBR) videos to be treated as constant bit-rate (CBR) by the VoD system for purposes of both network's bandwidth allocation and hiding the network jitter.

[2] Bandwidth is the main network resource that needs to be distributed to all applications in a way that simultaneously satisfies all QoS requirements [8].

video consumption rate. Assuming that both network and prefetch latencies are at their lowest values, a practical true VoD system can be considered the one which achieves zero service latency.

In the context of wired networks, many research efforts have been made to increase the scalability of VoD systems, i.e., the capacity of supporting an increasing number of simultaneous VoD clients. In particular, peer-to-peer VoD designs based on stream-reuse techniques can be highly scalable for switch-based LANs [14], by taking advantage of plentiful point-to-point bandwidth that switched LANs offer between clients. In case of WANs and MANs, where the access networks usually restrict the available bandwidth among clients, proxies can be placed at the edges of content distribution networks (CDNs) [11] to improve the scalability of VoD applications. On the other hand, with increasing offer of VoD-based applications, there is a growing need of ubiquitous access to them as more and more people use mobile devices capable of video playback (e.g., laptops, PDAs, high-end mobile phones) and wish to have access to those applications at their convenience.

Currently, IEEE 802.11 becomes a popular WLAN technology whose specification provides three main standards for the physical layer: a, b, and g. As reported in [12], each of the three standards supports a multitude of transmission modes - which specifies the data rate, the modulation scheme, and the error control scheme (e.g., FEC), if any. In this paper, we focused on the high-speed “a” and “g” 802.11 variants, working in Access Point (AP) infrastructure based mode [1]. The 802.11g offers only three interference-free, coexistent channels over the 2.4 GHz band, while 802.11a supports thirteen channels that operate over the 5 GHz frequency range. The key idea behind the present work came from the observation that, it is theoretically possible to join channels’ capacities, so that we can have sixteen parallel channels in the coverage area of the *collocated APs*[3]. Although each channel has 54 Mbps of maximum link rate, its effective throughput using UDP is near to 30.7 Mbps, leading to a total aggregate bandwidth of up to 491.2 Mbps [13]. Note that it is proportional to the throughput of a Gigabit Ethernet interface, which is often used in VoD proxy solutions [4, 17]. As a result, 802.11 can be considered as

[3] APs that are positioned at the same point in space, with the same coverage area, using no-overlapping channels frequencies.

a potential candidate for access network in a VoD delivery system, especially in environments where support for mobile and portable stations[4] is most required.

Even though the aggregate bandwidth suffices, the 802.11 variants lack for an efficient mechanism that optimizes the use of the APs. Specifically, 802.11 requires an adequate *association control (AC)* procedure for choosing the AP where a client device must be associated when it moves into the coverage area of multiple APs, whether collocated or not, using non-overlapping channels. The 802.11 basic heuristic for association control, namely the Strongest Signal First (SSF), gives priority of choice to APs based on the Receive Signal Strength Indicator (RSSI). It is well-known that SSF often leads to poor load balance among the APs, while not providing minimum bandwidth guarantees. In particular, when applied to collocated APs, SSF tends to a random behavior since all the APs will be at the same distance to the client. Other widely used heuristic but that is non-interoperable and proprietary, is the Least Loaded First (LLF), where an incoming client is assigned to the AP with the highest available throughput. Although the LLF approach optimizes the use of APs[3], the lack of interoperability among different brands of wireless equipments restricts the appeal for LLF. Also, some wireless equipment companies extended LLF to support applications that require guarantee of a minimum bandwidth, which weakened interoperability even further.

Besides the minimum bandwidth restriction, heuristics for association control we found in the literature were generic in the sense that they were not able to explore singularities of applications to improve their performance. In our previous work [15] we modified slightly the LLF heuristic to provide minimum bandwidth, and applied the resulting LLF heuristic (LLF+) to a VoD system design that used collocated APs. Our experimental evaluation of the VoD system showed that, despite the effective use of the aggregate bandwidth, the LLF+ system could successfully handle only short video requests at low arrival rates, and performed poorly with increasing amount of denials of requests when either the video length or the arrival rate increased. As our main research interests aim to design scalable VoD systems

[4] A portable station can be moved between different locations, but it can only be used while at a fixed location. However, mobile stations access the LAN while in motion. Both station' types are supported by 802.11 [1].

for wireless networks, we developed novel association control heuristics that can take advantages of VoD application characteristics to further increase the scalability of wireless VoD systems.

In this work, we introduce Early Release First (ERF) and Bounded ERF (BERF), two application-friendly heuristics for association control that benefited from the patience factor. The patience factor is a property of VoD systems that express the users' expectation of the service in that the user can tolerate a given service latency, i.e. a waiting time between the video selection and the beginning of video reception. Given that the patience factor provides a certain slack time, both ERF and BERF techniques take advantage of that to allow the association control unit to accept requests, which otherwise would be refused, that will make future use of the channels, with a determined service time. In summary, this paper main contributions are:

1. By focusing on clients' behavior, we model the association control problem of a video-on-demand system with collocated APs for video delivery over WLANs;

2. We propose CINAPS, a new system design for VoD servers operating over 802.11 WLANs that can use efficiently the aggregate bandwidth of collocated APs, while providing minimal bandwidth guarantees through reservation mechanisms closely tied to the association control procedure;

3. We introduce two novel application-friendly heuristics for association control, namely ERF and BERF that take advantage of the patience factor of a VoD application, in order to allow the association control unit to accept requests for future use of wireless communication channels.

4. We provide an in-depth evaluation of the CINAPS system through simulated performance comparison between a generic association control heuristic and the two novel application-friendly heuristics we introduced.

This paper is structured as follows. Section 2 describes briefly related works. In Section 3, we formulate the problem of association control focusing on VoD system properties. In Section 4, we present the

CINAPS design and describe in details ERF and BERF association control heuristics. Section 5 reports a performance analysis of CINAPS based on simulated results. Finally, we draw our conclusions and outline ongoing work in Section 6.

2. Related Work

Several research works addressed the problem of VoD over WLANs, association control, and multi-channel ad hoc networks. However, we could not found any significant work directly related to ours. The work in [18] describes a VoD system design - namely MobiVoD - for mobile ad-hoc clients, which used periodic broadcast to increase the scalability of a VoD system. In contrast with our work, MobiVoD does not explore collocated AP's bandwidth, and reports results only for a single video. Another work is the WiVision system [6] that supports both live and on-demand delivery of video over WLANs located in the last-mile. Even though they provided practical results of the system for multi-channel 802.11b networks, they assumed a best-effort service and measured errors that occurred in different situations, which also is not our case since we support minimal QoS guarantee.

There are many works focusing on association control to achieve load balance among APs with overlapping coverage area, especially those focusing on fairness, where the goal is to provide the same level of service to all clients. In particular, the work in [3] showed that the performance of their max-min fairness scheme outperformed the SSF and LLF heuristics, and that by balancing the load on the APs the overall network throughput could be increased. Although that work did not study the behavior of the system by focusing on application needs, it helped us to better understand the problem of association control and inspired us to apply it to the VoD context.

Several works reported significant improvements on the throughput of ad-hoc networks organized as meshes, in particular those using either multiple channels [16] and/or interfaces [19]. Currently, our work uses infrastructure based WLANs, but it can be extended to deal with mesh networks, as well.

3. Application-Friendly Association Control

In this section, we model the association control problem for VoD systems that deliver videos on demand through WLANs to a set of mobile devices (MD) with multi-band 802.11 interfaces configured to either "a" or "g" through software.

Let $Max_{channels}$ be the maximum number of no-overlapping channels. There is a set of collocated APs (AP_{total}), where $1 <= AP_{total} <= Max_{channels}$. $MinAP_{throughput}$ and $MaxAP_{throughput}$ are the minimum and maximum effective AP throughput, respectively. The effective throughput ($AP_{throughput_i}$) is the throughput (in Kbps) of the ith AP, so that $MinAP_{throughput} <= AP_{throughput_i} <= MaxAP_{throughput}$. Thus we can derive the Aggregate Bandwidth (AB) of the wireless segment of the system, which is equal to $\sum_{i=1}^{AP_{total}} AP_{throughput_i}$.

We consider a set of videos (V_{total}), where $MinV_{rate}$ and $MaxV_{rate}$ are the minimum and the maximum video rate, which correspond to the average consumption rate (in Kbps) of the video at the decoder, respectively. Also, let the minimum and maximum video length measured in seconds be $MinV_{length}$ and $MaxV_{length}$. Thus the ith video has a video rate (V_{ratei}) and a length (V_{length_i}), so that $MinV_{rate} <= V_{ratei} <= MaxV_{rate}$ and $MinV_{length} <= V_{length_i} <= MaxV_{length}$. The system server has an effective throughput ($S_{throughput}$) that can deliver at least $\frac{S_{throughput}}{MaxV_{length}}$ and at most $\frac{S_{throughput}}{MinV_{length}}$ simultaneous video streams.

We assume that the client request rate aka arrival rate follows a Poisson process with a given λ so that we have in average λ video requests per minute. Also, the video popularity follows a general Zipf distribution with α skew. For a given period in seconds of simulation time, i.e., Time of Simulation (TS), the total amount of video requests (R_{total}) is on average $\frac{TS*\lambda}{60}$. Let C_{videoi} be the id of the chosen video for playback by the ith client. The AP where the ith client is associated with is C_{ap_i} and the time it requests C_{videoi} is expressed as $T_{request_i}$, while $T_{servicei}$ is the time when it starts receiving the stream. From these equations, we derive the Service Latency (SL_i) of the ith client (in sec), which is equal to $T_{servicei} - T_{request_i}$. As described before, the system uses a playout buffer at the client, which must be filled up before the playback start. We define this time ($Prefetch_i$) as the prefetching period of the ith client (in sec). Thus the total amount of waiting time between the request and playback start of the ith client, the Playback Latency (PL_i), is calculated by $PL_i = SL_i + Prefetch_i$. Also, we define the Patience Factor (PF_i) as the maximum SL that the ith client considers acceptable [2].

Finally, let $R_{accepted}$ and R_{denied} denote the amount of requests that are accepted and denied, respectively. We aim to minimize R_{denied} to increase the scalability of the system. A new request from the ith client is accepted only if the necessary bandwidth ($V_{rateC_{videoi}}$) can be allocated in one of the APs so that the $SL_i <= PF_i$. If not, the request is denied. Intuitively, we can say that the availability of bandwidth is a function of the λ, V_{length}, V_{rate}, and AB. Thus $R_{accepted}$ and R_{denied} will vary according to the combination of these parameters.

4. CINAPS: Cluster of INexpensive Access PointS

In this section, we describe a new scalable solution for video-on-demand systems operating over 802.11 WLANs, namely Cluster of INexpensive Access PointS (CINAPS). First, we present a general view of CINAPS and a brief description of its software components. After, we describe in details the Wireless Channel Manager (WCM) unit that is responsible for the association control, and introduce the WCM's novel application-friendly heuristics.

4.1. General View of CINAPS

Figure 1 shows the hardware framework CINAPS uses. It consists of a cluster of commodity 802.11 APs in a collocated manner that places the APs practically at the same location, enabling the APs to handle the same coverage area through non-overlapping channel frequencies. The APs are interconnected by a network switch that also supports at least one Gigabit Ethernet port to connect the VoD server to the APs.

The operation of CINAPS involves three software components, each of which is responsible for a specific task, as follows:

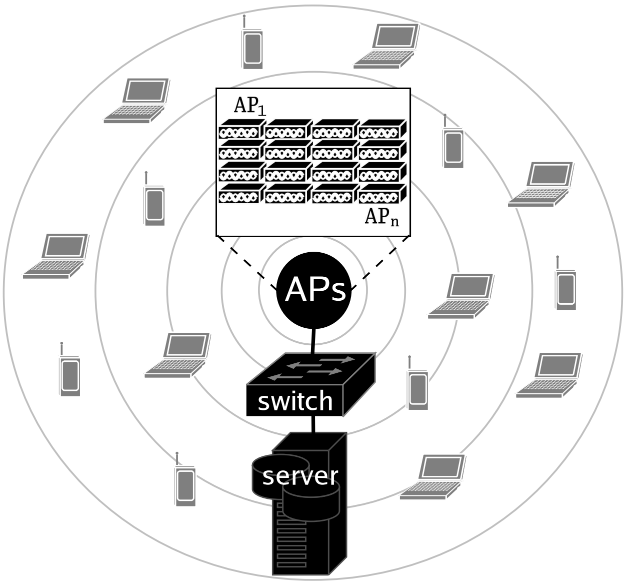


Figure 1. Hardware framework of CINAPS

- Video Player (VP) task. It runs in the client device (e.g.,laptops, PDAs, high-end mobile phones, with 802.11 multi-band interfaces), taking care of both video requests and playbacks;
- Video Server (VS) task. It runs in the system server, handling video requests either as a real server (i.e., as the main storage of video content) or as a proxy.
- Wireless Channel Manager (WCM) task. It runs necessarily in the system server, performing the association control that selects to every new client the appropriate AP it will be associated with.

4.2. Association Control Procedure

The association control procedure follows the protocol presented in Figure 2. Every time a client selects a video to watch, the video player sends a $token1$ (CID, VID) to WCM through any one of the APs, where CID and VID are the client id and the video id, respectively. Depending on the available bandwidth at the collocated APs, WCM accepts or denies the video request. Whenever WCM accepts a request, it sends the $token2$ to CID, in the form of (CID, $T_{serviceCID}$, $APID$), where $T_{serviceCID}$ is the time at which the bandwidth will be actually allocated to CID, and $APID$ is the id of the AP that CID

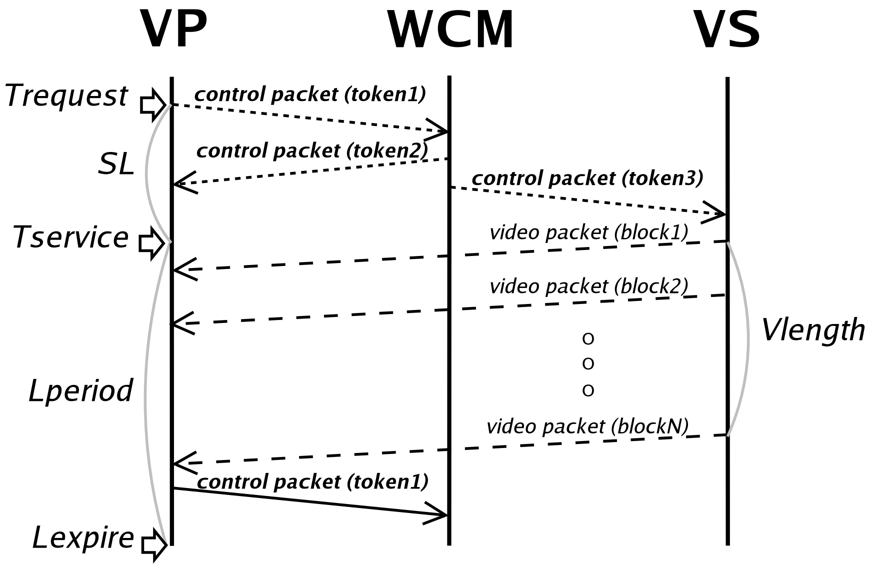


Figure 2. Protocol used for association control

will be associated with. WCM also signalizes the video server to start a video stream VID to CID at the time $T_{serviceCID}$ using the $token3$ (CID, VID, $T_{serviceCID}$). If the request is denied, a $token2$ with $T_{serviceCID}$=$NULL$ and $APID$=$NULL$ is sent to CID. After receiving the last video packet, the client will send a $token1$ with VID=$NULL$ to WCM, just to confirm the release of bandwidth.

4.3. WCM Internal Operation

To accomplish the association control task, WCM adopts the concept of Sub-Channel Lease[5] for accepted requests - the client CID that requests the video VID receives a lease i.e., a fraction of the entire bandwidth of the access point $APID$ (at most $AP_{throughput_{APID}}$) during a specified period of time, namely Lease Period (L_{period}) (Fig. 2). As we mentioned before, due to the regular behavior of VoD applications it is expected that once a client requests a given VID, it will watch it completely. For VoD systems that adopt the transmission rate equal to the video rate ($V_{rateVID}$), it suffices to allocate the fraction of the channel (aka Sub-Channel) proportional to that rate at the time $T_{serviceCID}$, for a period of time that lasts - the L_{period} - proportional to $V_{length_{VID}}$. Thus, the release time of the

[5] A lease is a contract that grants its holder specified rights over property for a limited period of time [10].

Sub-Channel occupied by the CID, hereafter expressed as Lease Expiration Time ($L_{expire_{CID}}$), is set to $T_{serviceCID} + V_{length_{VID}} + 1^6$.

The maintenance of leases is done through a simple structure, the Release List (RL). The entries of RL have the following fields: RL_{cid}, for the client id (CID); RL_{vid}, for the video id (VID) the client requested; RL_{apid}, for the AP id ($APID$) where the client is/or will be associated; and RL_{time}, for the lease expiration time which is the time when the sub-channel allocated to CID will be released ($L_{expire_{CID}}$). Thus, every time a request is accepted from CID to VID, a new entry is created in the RL, with $RL_{time} = L_{expire_{CID}}$, so that the increasing order of RL_{time} is preserved.

As described before, $T_{serviceCID}$ must be set to compute $L_{expire_{CID}}$. To do this, WCM checks if there is an AP with available bandwidth to accommodate $V_{rateVID}$. If it exists, $T_{serviceCID}$ is set to zero. In the other hand, WCM checks the RL to find out the Earliest AP (EAP) that will have available bandwidth to satisfy the bandwidth requirements as requested. To do so, it is enough that WCM traverses the RL adding separately for each AP the bandwidth ($V_{length_{RL_{vid}}}$) of the previous clients until the aggregate bandwidth of any one of the APs is equal or greater than $V_{rateVID}$. When it occurs, EAP is set to RL_{apid} and $T_{serviceCID}$ receives RL_{time}. Hereafter, this procedure is refereed to as DiscoverEAP.

If all the videos have the same rate ($MaxV_{rate}$ = $MinV_{rate}$), the total throughput of each AP is the same ($MinAP_{throughput}$ = $MaxAP_{throughput}$), and $AP_{throughput}$ is a multiple of the V_{rate}, then it is enough to consult just the first entry to discover EAP. Otherwise, the maximum number of entries to be searched for is $\frac{AP_{total} * MaxV_{rate}}{MinV_{rate}}$. For instance, let us assume that a request from client CID is accepted by WCM after consulting the RL with $APID$ = EAP. To keep RL updated, all entries with $RL_{apid} = EAP$ that were traversed must be removed from RL, because their times will be accumulated in the RL_{time} field of the entry created for CID.

The WCM unit was designed with focus on modularity so that it could support different types of association control heuristics, either generic or application-friendly. In this work, WCM supported three

[6] To hide a possible delay in the sub-channel acquire/release process.

heuristics, one generic - the LLF+, which implemented the least loaded first heuristic (LLF) [3] slightly modified to guarantee minimum bandwidth - and two application-friendly - the novel ERF and BERF that benefited from exploiting the patience factor of VoD applications. More specifically, ERF and BERF take into consideration the patience factor to allow the association control unit to accept requests for future use of the channels. Next, we describe ERF and BERF in details.

The Early Release First (ERF) is an application-oriented association control heuristic. It takes advantage of tolerance to service latency of VoD applications[7] that allows us to create a lazy association procedure for busy periods in which channels are fully used. When it happens, client requests are enqueued and serviced at a later time when enough bandwidth becomes available. The key point is that the time at which there will be enough bandwidth is easily predictable by the system. This feature allows a client to know at request time the actual service latency it will experience.

The Bounded Early Release First (BERF) is an extension of ERF that guarantees maximum latency. In this way, the client request will be accepted only if the service latency will not go beyond the patience factor (PF). It is interesting to note that ERF can be seen as a BERF with PF=∞, while LLF+ performs similar to BERF with PF=0.

Besides the release list (RL) structure, WCM maintains a vector - FB[n] - to keep the free bandwidth of each one of the n collocated APs and uses the algorithm in Figure 3 to either accept or deny requests, depending on the heuristic (H) employed.

5. Experimental Analysis

In this section, our main goal is to investigate the impact of association control heuristics on VoD system's performance. In the following subsections, we present the evaluation methodology, the assumptions we made for the simulation environment, and then report the performance results we obtained.

[7] Clients usually accept some delay between the issue of video request and starting the video playback.

```
For $i$ from 1 to $AP_{total}$ do
    FB[$i$] = $AP_{throughput_i}$
For each $token1$ received do
    If $VID$ = $NULL$ then do
        FB[$C_{ap_{CID}}$] += $V_{rate C_{videoCID}}$
    Otherwise, do
        $APID$ = $NULL$; $T_{serviceCID}$ = $NULL$
        For $i$ from 1 to $AP_{total}$ do
            If $V_{rateVID}$ < FB[$i$] then do
                $C_{videoCID}$ = $VID$; $C_{ap_{CID}}$ = $i$; $APID$ = $i$;
                $T_{serviceCID}$ = 0; FB[$i$] -= $V_{rate C_{videoCID}}$
        If $APID$ = $NULL$ and $H$ != $LLF+$ then do
            $APID$ = Discover$EAP$
        If $H$ = $BERF$ and $SL_{CID} > PF_{CID}$ then do
            $APID$ = $NULL$; $T_{serviceCID}$ = $NULL$
        Send $token2$ to $CID$
```

Figure 3. WCM algorithm used to handle video requests

5.1. Evaluation Methodology

To evaluate our proposal for association control heuristics we developed a discrete event simulator [9]. We used two main metrics to quantify the overall system performance. First, the Blockage Rate (BR) metric that hints at the scalability of a given VoD system. We defined (BR), as follows:

$$BR = \frac{R_{denied}}{R_{total}} \tag{1}$$

where R_{total} is the total amount of video requests the VoD system received and R_{denied} is the number of requests that WCM could not service due to the unavailability of enough resources. Thus, the lower is BR, the higher is the VoD system's scalability.

Second, the Average Latency (AL), which measures how far is the system from a true VoD system, as described before. AL is given by:

$$AL = \frac{\sum_{i=1}^{R_{accepted}} SL_i}{R_{accepted}} \tag{2}$$

were $R_{accepted}$ is the total amount of video requests WCM accepted and SL_i is the service latency of request i, which is the elapsed time between the request and the allocation of a sub-channel for i.

5.2. Simulation Environment

We assume a scenario where thousands of potential video clients have mobile devices (MD) distributed over an area near to two hundred square meters. For instance, it may represent sports events in stadiums, hot-spots in airports, among others. Thus the size of the simulated area is equal to 200 m^2, which is further divided in 5 m^2 regions. MDs are randomly placed in those regions. Each MD has its own id, so that the first client that requests a video has id=1, the second has id=2, and so on. As said before, we assumed that the arrival of client requests follows a Poisson process with λ and the choice of videos follows a Zipf distribution with $\alpha = 0.7$ [5].

Our simulation assumptions are as follows:

- Contention-free Dedicated Network: the network's bandwidth is dedicated only to the traffic that the VoD system generates. Also, as we have a centralized main source of traffic - the video server - we assume that in practice the contention at the APs will be insignificant given to the small size and amount of control packets that are exchanged among the units. Thus, the system can use DCF without standard modifications, while guaranteeing QoS to the application.

- Transmission Range and Rates: we consider the coverage area of the APs equal to the simulation area and the use of a single transmission mode. We focused on VoD applications, where the bandwidth each client uses was bounded to the video rate we set at 1024 Kbps[8]. At this client rate, it suffices to set the 802.11a/g operation mode at the lowest throughput.

- Channel Interference: we adopted only non-overlapping channels, using different frequency ranges, and did not investigate interference from other types of signal sources. Regarding to the use of multiple APs near to each other, we assumed that the interference among them will be in an ac-

[8] Today's codecs offer an excellent playback quality at that rate when using a spacial resolution of 320x240 pixels.

ceptable threshold provided that each AP is at least 60 cm far from any other AP, as shown in [16].

- Transmission errors: although error resilience is an important issue in wireless video delivery [7], our current proposal did not tackle such a problem. However, note that VoD allows trivial error treatment at both the network and the application level.
- Equalized Signal-Strength: all APs are located in the same point in the simulated scenario, using the same transmission power. Thus, the signal level is the same for all APs, as perceived by the MD. As there is no strongest-signal AP, the SSF association control heuristic is useless in this context.

5.3. Performance Evaluation

In this subsection, we will evaluate our simulated performance results. Given the restricted space, we will concentrate our analysis on videos of same length and rate, ranging from one to twenty minutes, which we assume will be most popular for environments with mobile and portable devices, especially because wireless devices have battery constraints. Also, we restricted the analysis on APs with the same throughput and defined all clients with the same patience factor. Table 1 summarizes the simulation parameters.

Despite using multiple videos in our simulations, we expect that the simulated behavior will be the same for a single video because we used multiple videos with the same rate and length. In future work we plan to evaluate performance results for multiple videos with different lengths and transmission rates.

Figure 4 shows the influence of the association control heuristic (H) on the blockage rate (BR) for two video lengths (60 and 1200 sec), using a patience factor (PF) equal to video length (V_{length}). Note that ERF was not plotted because it always achieved BR=0. The curves show that a single AP was enough to deliver all the video lengths we analyzed for arrival rates lower than two requests/min using any heuristic. As the arrival rate increased, the number of collocated APs had to be expanded to avoid request denials,

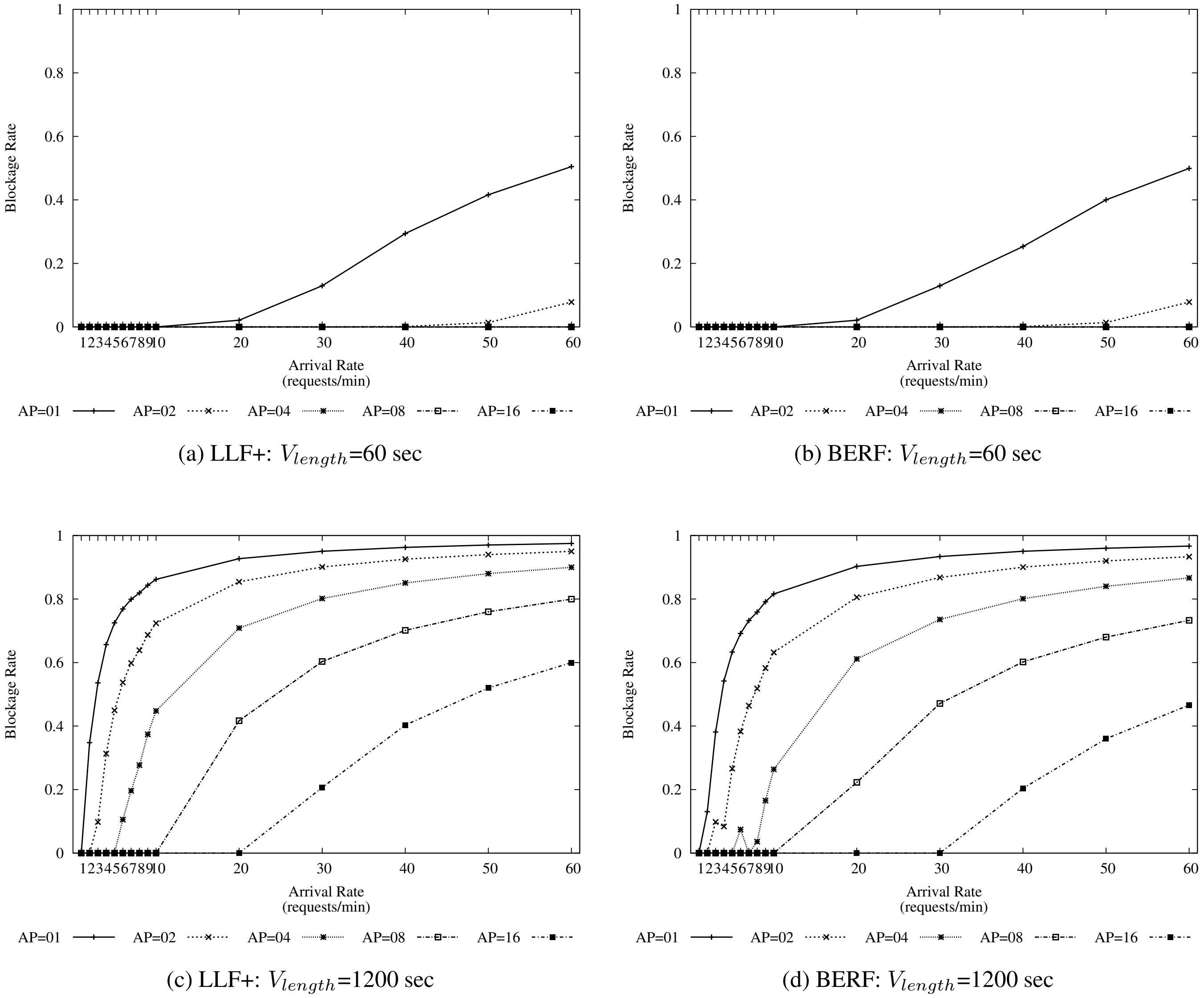
Table 1. Simulation Parameters

Parameter	Values
Simulation Time - TS - (sec)	3600
Collocated APs - AP_{total} - (units)	1, 2, 4, 8, 16
AP Throughput - $AP_{throughput}$ - (Kbps)	30720
Video Length - V_{length} - (sec)	60, 300, 600, 900, 1200
Video Rate - V_{rate} - (Kbps)	1024
Videos - V_{total} - (units)	100
Arrival Rate - λ - (requests/min)	1, ..., 9, 10, ..., 60
Patience Factor - PF - (sec)	60, 600, 1200, V_{length}

except for ERF that always accepted requests for future use of sub-channels, while increasing the average service latency. For the arrival rates we measured and short-length videos, less equal to 300 sec, LLF+ and BERF sufficed. However, the VoD system using the heuristics with maximum number of collocated APs had to deny requests for long videos. Nevertheless, BERF significantly decreased the blockage rate when compared to LLF+, especially when the patience factor was equal to or greater than the video length. In the worst case (V_{length}=1200), BERF generated near to 25% less requests denials than LLF+.

Figure 5 depicts the impact of the association control heuristic (H) on the average service latency (AL) experienced by the clients for different video sizes, using PF=V_{length}. As LLF+ only accepts requests if there is a sub-channel available at the time of request, its AL is always minimal. For ERF and BERF, the minimal AL is preserved until the aggregate bandwidth is completely allocated. For higher arrival rates, ERF showed an increase on AL almost linear with the value of λ, limited by the simulation time. In the other hand, BERF kept the AR lower than the patience factor, for all video lengths.

Figure 6 presents the influence of combinations of patience factor (PF) and amount of APs (AP_{total}) on the blockage rate (BR) achieved with the BERF heuristic for different video lengths. As expected, the influence grows with V_{length} and, the smaller is the aggregate bandwitdth, the higher is the influence. Moreover, the use of higher PF is more effective for middle range λ, especially for arrival rates slightly



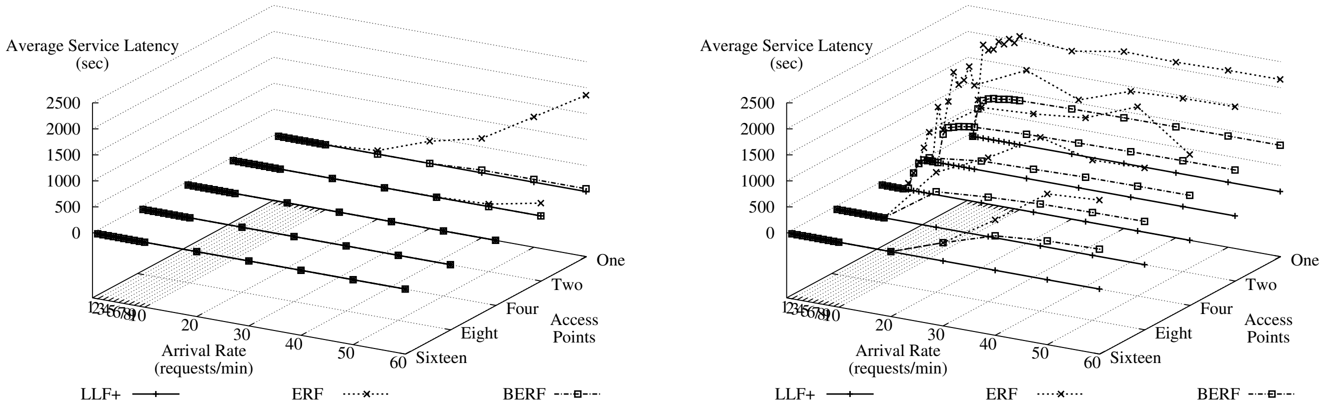
(a) LLF+: V_{length}=60 sec

(b) BERF: V_{length}=60 sec

(c) LLF+: V_{length}=1200 sec

(d) BERF: V_{length}=1200 sec

Figure 4. Influence of the heuristic (H) on the blockage rate (BR) for the shortest and longest video length



(a) V_{length}=60 sec

(b) V_{length}=1200 sec

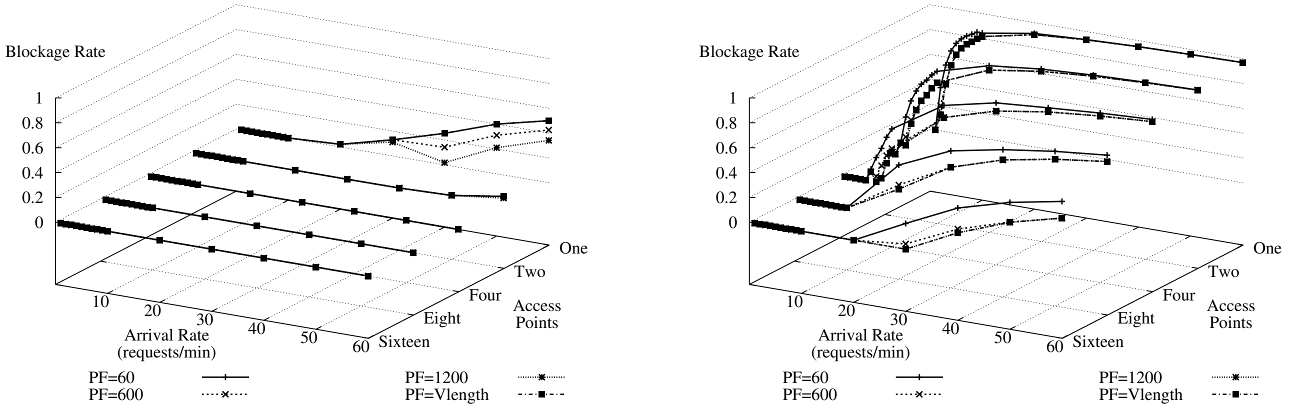
Figure 5. Impact of the heuristic (H) on the average service latency (AL) for different V_{length}

greater than the minimal rate that achieves full aggregate bandwidth use.

Table 2 summarizes the maximum blockage rate ($MaxBR$) for each combination of V_{length} and AP_{total} measured for the LLF+ and BERF (PF=V_{length}) heuristics. ERF was not placed in the table because its $MaxBR$ was zero for all combinations we tested, since ERF does not deny requests. Comparing BERF and LLF+, we clearly see that BERF outperformed LLF+ for all cases where the aggregate bandwidth is not enough to let WCM to immediately allocate a sub-channel in response to a video request. Moreover, their performance difference increased almost linearly with the augment of aggregate bandwidth the APs offered. Overall the simulated results indicated that BERF is significantly more scalable than LLF+.

6. Conclusions and Ongoing Work

In this paper, we proposed CINAPS, a new solution for video-on-demand (VoD) systems over WLANs with collocated access points (APs). Key to CINAPS' scalable performance is the support to application-friendly heuristics for association control that guarantee minimum bandwidth through reservation mechanisms. In particular, we developed two novel heuristics, the Early Released First (ERF) and the



(a) V_{length}=60 sec

(b) V_{length}=1200 sec

Figure 6. Influence of the patience factor (PF) and AP_{total} on blockage rate (BR) for different V_{length} using BERF

Table 2. $MaxBR$ of LLF+ and BERF (PF=V_{length}) for combinations of V_{length} and AP_{total}

AP_{total}	H	V_{length} (sec)				
		60	300	600	900	1200
01	LLF+	0.505	0.900	0.950	0.967	0.975
	BERF	0.499	0.892	0.942	0.958	0.967
02	LLF+	0.078	0.800	0.900	0.933	0.950
	BERF	0.078	0.783	0.883	0.917	0.933
04	LLF+	0.000	0.600	0.800	0.867	0.900
	BERF	0.000	0.566	0.766	0.833	0.867
08	LLF+	0.000	0.204	0.600	0.733	0.800
	BERF	0.000	0.132	0.533	0.666	0.733
16	LLF+	0.000	0.000	0.199	0.466	0.600
	BERF	0.000	0.000	0.066	0.333	0.466

Bounded ERF (BERF) that exploited the slack of time aka the patience factor that VoD applications provides, to allow the association control unit to accept future requests for using the communication channels. Our simulated results revealed that CINAPS could decrease the blockage rate of video requests substantially, so that CINAPS outperformed significantly our previous VoD system that used the generic heuristic LLF (Least Loaded First) slightly modified to assure minimum bandwidth guarantees.

Currently, we plan to evaluate VoD system's performance in heterogeneous 802.11 networks, where mobile client devices can have either multi-band or single-band interfaces. This and future works are part of the ongoing TRAVIS-QoS project (http://www.lcp.coppe.ufrj.br) aiming to build a wireless proof-of-concept prototype to validate the results we obtained so far through simulations.

Acknowledgments

The authors would like to thank the researchers of the Parallel Computing Lab, especially the multimedia team, for their helpful comments, our colleagues from TRAVIS-QoS project that are hardworking to create a wireless environment for our tests, and the Brazilian agencies CAPES, FINEP, and CNPq for their support.

References

[1] Wireless LAN medium access control (MAC) and physical layer (PHY) specification, 1999 edition, 1999.

[2] E. L. Abram-Profeta and K. G. Shin. Scheduling video programs in near video-on-demand systems. In *Proceedings of the ACM Multimedia*, pages 359 – 369, 1997.

[3] Y. Bejerano, S.-J. Han, and L. E. Li. Fairness and Load Balancing in Wireless LANs Using Association Control. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom)*, pages 315–329, Philadelphia, PA, September 2004.

[4] C-COR (formerly nCUBE), 2005. http://www.ccor.com/.

[5] A. Dan, D. Sitaram, and P. Shahabuddin. Dynamic Batching Policies for an On-Demand Video Server. *Multimedia Systems*, 4(3):112–121, 1996.

[6] P. De, S. Sharma, A. Shuvalov, and T. Chiueh. WiVision: A Wireless Video System for Real-Time Distribution and On-Demand Playback. In *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, January 2004.

[7] M. Etoh and T. Yoshimura. Advances in wireless video delivery. *Proceedings of the IEEE*, 93(1):111–122, January 2005.

[8] A. Ganz, Z. Ganz, and K. Wongthavarawat. *Multimedia Wireless Networks: Technologies, Standards, and QoS*. Prentice Hall, 2004.

[9] J. Garrido. *Practical process simulation using object-oriented techniques and C++*. Artech House, 1998.

[10] C. G. Gray and D. R. Cheriton. Leases: An efficient fault-tolerant mechanism for distributed file cache consistency. In *Proceedings of the Twelfth ACM Symposium on Operating System Principles (SOSP)*, pages 202–210, Litchfield Park, AZ, December 1989.

[11] E. Ishikawa and C. L. Amorim. Collapsed Cooperative Video Cache for Content Distribution Networks. In *Proceedings of the Brazilian Simposium on Computer Networks (SBRC)*, pages 249–264, Natal, RN, Brazil, May 2003.

[12] M. Manshaei, T. Turletti, and M. Krunz. A media-oriented transmission mode selection in 802.11 wireless LANs. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, Atlanta, GA, March 2004.

[13] B. McFarland and M. Wong. The family dynamics of 802.11. *ACM Queue*, 1(4), May 2003.

[14] L. B. Pinho and C. L. Amorim. Assessing the efficiency of stream reuse techniques in P2P video-on-demand systems. *Journal of Network Computing Applications (JNCA)*, 2004. (article in press).

[15] L. B. Pinho and C. L. Amorim. Investigating the Performance of Video-on-Demand Systems over WLANs Using Generic Association Control. Technical Report ES-677/05, COPPE/UFRJ Systems Engineering Program, May 2005.

[16] A. Raniwala and T. Chiueh. Architecture and Algorithms for an IEEE 802.11-based Multi-channel Wireless Mesh Network. In *Proceedings of the INFOCOM*, Miami, FL, March 2005.

[17] SeaChange, 2005. http://www.schange.com/.

[18] D. A. Tran, M. Le, and K. A. Hua. MobiVoD: A Video-on-Demand System Design for Mobile Ad Hoc Networks. In *Proceedings of the IEEE International Conference on Mobile Data Management (MDM)*, pages 212–223, Berkeley, CA, January 2004.

[19] J. Zhu and S. Roy. 802.11 Mesh Networks with Two Radio Access Points. In *Proceedings of the IEEE International Conference on Communications (ICC)*, Seoul, Korea, May 2005.