

CPC Technical Draft

Layered Media Multicast Control (LMMC): Rate Allocation and Partitioning

Homayoun Yousefi'zadeh Hamid Jafarkhani Amir Habibi

Center for Pervasive Communications
Electrical and Computer Engineering Department
University of California, Irvine
Irvine, CA 92697

[hyousefi,hamidj]@uci.edu

Abstract

In the recent years, layering techniques of distributing multimedia traffic over multicast IP networks have received a growing attention. The objective of such techniques is to effectively cope with the challenges in continuous media applications such as heterogeneity, fairness, real-time constraints, and quality of service. In this paper, we present an analytical solution to the general problem of rate allocation and receiver partitioning in layered and replicated media systems. Our optimal Layered Media Multicast Control (LMMC) solution to a formulation of the rate allocation problem analytically determines the layer rates and the corresponding partitioning of the receivers such that a close approximation of the so-called max-min fairness metric is maximized.

Index Terms

Layered Media, Replicated Media, Multicast IP Networks, Heterogeneity, Optimality, Fairness Extrapolation, Rate Allocation, Receiver Partitioning.

I. INTRODUCTION

TRANSMITTING realtime compressed digital media over multicast IP networks has been the subject of heavy research in the recent years as surveyed by Li. et al. in [15] and the references cited therein. There have been three different adaptive bit-rate media multicasting schemes for the transmission of digital media in the literature:

- Single stream adaptive approach first presented by Bolot et al. [4] and Ammar [2] in which a single encoded video stream is transmitted by the source with feedback returned from the receivers to the source. The source uses the feedback information to adapt its data rate. One of the potential problems with this approach is the problem of feedback implosion for a large number of receivers attempting to return feedback to the source. Practical video multicast protocols targeting a large number of receivers require to address this issue. The single stream adaptive approach while being straightforward is unable to properly accommodate different bandwidth requirements of the target session receivers, an issue also known as receiver heterogeneity problem.
- Replicated media streams approach was first presented by Cheung et al. [5] within the context of DSG protocol as an extension to the single stream approach that is capable of addressing the heterogeneity issue. In this approach the source sends multiple streams carrying the same video with different quality and bit rate obtained by encoding different streams with different compression parameters. Each stream is multicast for a different multicast group with receivers being able to join and change the groups according to their capacities. While the simplicity of this scheme in addressing the heterogeneity issue is attractive, it has the drawback of requiring the network to carry redundant information of replicated media streams.

- Layered media streams approach was first proposed by Deering et al. [6] in the context of multicast routing and further enhanced by McCanne et al. [18] in the context of RLM protocol, Amir et al. [1] in the context of SCUBA protocol, and Li et al. [16] in the context of rate control aspect of LVMR protocol. The approach relies on the ability of many video compression schemes to divide their output bit stream into layers; a base layer and one or more enhancement layers. The base layer can be independently decoded providing a basic level of video quality. The enhancement layers can only be decoded together with the base layer providing improvements to video quality. This approach is also known as successive refinability approach in the context of source coding literature and is discussed by Jafarkhani et al. in [9] and references therein. Using this capability, a video multicast source could send each layer to a different multicast group. Receivers would then join at least the base layer group and join as many enhancement layer groups as their capacities allow. A valid question within the context of layered media systems is that what is the optimum number of the layers? Although the answer to this question is rather complicated, it has to be mentioned that the trade off is between accommodating wide receiver heterogeneity and incurring excessive overhead in source encoding, receiver decoding, and multicast addressing. Layered media approach provides an elegant and efficient way to deal with the heterogeneity issue at the expense of protocol complexity.

In a typical multicasting transmission scenario, a source generates realtime media traffic following a periodic pattern. The periodic pattern of realtime media traffic generated at a source consists of many frames in a unit of time at a variable bit rate, i.e., the number of bits per frame varies for individual frames. The receivers rely on a preserved frame periodicity at the time of play back. Data not available at the play back time is considered lost. In addition, the delay jitter or the difference in the delay of packets arrived at the receivers has to be small. In order to accommodate the latter need, buffering techniques at the receiver can be employed.

The main objective of the current research work is to provide an analytical framework for the partitioning strategy and rate allocation of layered media systems over multicast IP networks in the context of Layered Media Multicast Control (LMMC) protocol. In this study, we assume the existence of congestion and flow control mechanisms capable of dynamically addressing inter-session fairness issue, i.e., a fair distribution of available bandwidth among multiple media and other sessions such as TCP sessions. Typical examples of such mechanisms are given in [24], [17], [26], [20], and [25]. Hence, the result of our work manifests in dynamic distribution of an assigned fair bandwidth among individual layers of a media session considering intra-session fairness and receiver heterogeneity issues. The main contribution of this paper is in three areas. First, the paper introduces an analytical approach in which the non-continuously differentiable max-min fairness function of individual receivers as described in [11] and [10] is extrapolated by a class of mathematically well-behaved continuously differentiable functions satisfying the conditions required for applicability of traditional optimal control techniques. Second, the paper provides an analytical solution to a formulation of the optimal rate allocation problem of the replicated and layered media systems. Third, the paper offers a near optimal receiver partitioning strategy maximizing the enhanced fairness utility metric for any set of allocated layer rates. It is important to note that the technique proposed in this paper can be independently applied to both replicated media streams as well as layered media streams.

Specifically, we formulate a two-phase optimal control rate allocation problem. In the first phase, we start from an initial partitioning of the receivers and solve the optimal rate allocation problem for individual layers of the media session assuming the number of layers is given. We are able to provide an analytical solution to this first phase problem as the result of applying our fairness extrapolation technique in which a mathematically well-behaved function represents the fairness metric. In solving the optimal control problem, we maximize the session fairness utility function defined by Jiang et al. in [13], [11], and Yang et al. in [27] with an extra constraint. The extra constraint places an upper bound on the overall available bandwidth to the session. The upper bound is provided by a flow control mechanism as the result of enforcing an inter-session fairness algorithm. The solution to this first problem considers receiver heterogeneity, i.e., the variation of the bandwidth among different receivers of the target session by means of maximizing an enhanced inter-receiver fairness metric. In the second phase, we provide a near optimal partitioning strategy for the layered media session based on the allocation rates of the first phase. The solution to the second problem maximizes the overall fairness utility function of the media session.

Considering the phasing approach of our solution, we introduce an iterative approach that can reach a near-optimal solution by iteratively applying the partitioning result of the second phase to the first phase and solving the optimal rate allocation problem with the new partitioning strategy. This is equivalent to employing steepest descent optimal

control strategy and is guaranteed to reach an ϵ -neighborhood of a local optimum point if such a point exists.

In summary given the overall available bandwidth to a media session, the LMMC solution to the formulation of the problem identifies the optimum rates for each individual layer and the corresponding receiver partitioning such that the fairness utility function of the session is maximized while satisfying the problem constraints. To the best of our knowledge, this is a unique approach providing an analytical solution to the rate allocation problem of layered media in multicast networks.

An outline of the paper follows. In Section II, we formulate the two-phase receiver partitioning and rate allocation problem considering individual receivers max-min fairness. In Section III, we analytically solve the optimal rate allocation problem of the first phase assuming a given partitioning. In Section IV, we use the allocated rates of Section III to obtain a near-optimal partitioning strategy. In Section V, we introduce an iterative approach relying on the solution of Sections III and IV to reach a near-optimal solution. Section VI focuses on performance evaluation and includes the simulation results along with practical considerations. Finally, Section VII contains a discussion of the future work and concluding remarks.

II. FORMULATION OF THE PROBLEM BY MEANS OF FAIRNESS EXTRAPOLATION

In this section, we formulate the general rate allocation problem of the layered media sessions in a manner similar to the formulation of [13], [11], and [27] with an extra constraint on the overall available bandwidth to the session. The previous problems without any constraints can hence be considered as a specific case of our problem.

Consider a multicast media session with a partitioning of the receivers into K groups. Remind that for a media session with N receivers and K groups, a set $P = \{G_1 | \dots | G_K\}$ is called a partitioning of the receiver set $R = \{1, \dots, N\}$ if P is a decomposition of the set R into a family of disjoint sets. Make note of the fact that we are formulating the problem for a given number of groups. The impact of the changes in the number of groups K is investigated in Section VI. The term group rate is used to denote the aggregated receiving rate of a receiver in the group while the term layer rate is used to denote the transmission rate to a specific layer. For an ordered partitioning of the receivers into K groups with ordered group rates of g_1, g_2, \dots, g_K such that $g_1 \leq g_2 \leq \dots \leq g_K$, the layer rates of a layered media session are calculated in the form of,

$$g_1, g_2 - g_1, g_3 - g_2, \dots, g_K - g_{K-1} \quad (1)$$

A receiver in group k subscribes to layers 1 through k receiving an aggregated rate of g_k .

Interpretation of our formulation in case of replicated media streams is also straight forward. For an ordered partitioning of the receivers into K groups G_1, G_2, \dots, G_K with ordered group rates of g_1, g_2, \dots, g_K such that $g_1 \leq g_2 \leq \dots \leq g_K$, the layer rates are the same as the group rates. A receiver in group k only subscribes to layer k receiving a rate of g_k . The interpretation difference has a minor impact on the formulation and consequently the solution of the problem in special cases which will be discussed in Section III.

The optimization problem is formulated by means of defining a per receiver max-min fairness utility with the objective of maximizing the session utility defined as the sum of receiver utilities over the layered media session. Each receiver is assumed to have an isolated multi-rate max-min fair rate of r_i as described in [21]. This is the reception rate of the receiver and is typically determined by a network bottleneck link from the source to the receiver or the receiver itself. For the clarity of representation, we also assume that the receivers are numbered such that their isolated rates are in non-decreasing order, i.e., $r_1 \leq r_2 \leq \dots \leq r_N$. In addition, each receiver i is assumed to have a loss tolerance L_i identified as its largest acceptable loss rate. Therefore, the group rates g_k should satisfy the following inequality for individual receivers of the group

$$g_k \leq \frac{r_i}{1 - L_i} \quad (2)$$

$$\forall i \in G_k \quad k = 1, \dots, K$$

In [12] Jiang et al. define the max-min fairness utility for receiver i of group G_k over the receiver isolated rate r_i and its group rate g_k as

$$F(r_i, g_k) = \frac{\min(r_i, g_k)}{\max(r_i, g_k)} = \begin{cases} \frac{g_k}{r_i} & : g_k \leq r_i \\ \frac{r_i}{g_k} & : g_k \geq r_i \end{cases} \quad (3)$$

They define the group utility for the group G_k with a group rate g_k as,

$$IRF_k = \sum_{i \in G_k} F(r_i, g_k) = \sum_{i \in G_k} \frac{\min(r_i, g_k)}{\max(r_i, g_k)} \quad (4)$$

In order to assign priorities to the different receivers of a group, the fairness utilities of the receivers can be multiplied by a parameter α_i with the following characteristics,

$$\begin{aligned} \sum_{i=1}^N \alpha_i &= 1 \\ 0 \leq \alpha_i \leq 1 &\quad \text{for } i = 1, \dots, N \\ \alpha_i = 0 &\quad \text{for } i \notin G_k \end{aligned} \quad (5)$$

The choice of parameters α_i is typically a design decision and in general does not have any significant impact in our study. The session utility of the partitioning $P = \{G_1 | \dots | G_K\}$ is defined as

$$\begin{aligned} IRF_{Total} &= \sum_{k=1}^K IRF_k \\ &= \sum_{k=1}^K \sum_{i \in G_k} F(r_i, g_k) \\ &= \sum_{k=1}^K \sum_{i \in G_k} \frac{\min(r_i, g_k)}{\max(r_i, g_k)} \end{aligned} \quad (6)$$

The objective of both heuristics given in [11] and dynamic programming algorithm given in [27] is to determine the optimal partitioning and the optimal layer rate allocations such that the function defined in (6) is maximized considering receivers loss constraints. The rate allocation optimization problem is, then, formulated as

$$\begin{aligned} \max_{g_1, \dots, g_K} IRF_{Total} &= \max_{g_1, \dots, g_K} \sum_{k=1}^K IRF_k \\ &= \max_{g_1, \dots, g_K} \sum_{k=1}^K \sum_{i \in G_k} \frac{\min(r_i, g_k)}{\max(r_i, g_k)} \\ \text{Subject To: } g_k &\leq \frac{r_i}{1 - L_i} \\ &i \in G_k \quad k = 1, \dots, K \end{aligned} \quad (7)$$

$$\quad (8)$$

for the optimal partitioning $P^* = \{G_1^* | G_2^* | \dots | G_K^*\}$ leading to the calculation of the optimal rates $g_1^*, g_2^*, \dots, g_K^*$.

In **Theorem (1)** of [27] the existence of an ordered receiver partitioning that maximizes the function defined in (6) is proven assuming the receiver utility function $F(r, g)$ satisfies a Receiver Utility Property (RUP). The RUP holds for a receiver with an isolated rate r in a group G with a group rate g if,

- $F(r, g)$ is non-decreasing in the interval $[0, g]$ and non-increasing in the interval $[g, \infty)$ for a fixed r ; and
- $F(r, g)$ is non-decreasing in the interval $[0, r]$ and non-increasing in the interval $[r, \infty)$ for a fixed g .

We now introduce an extrapolation technique to replace the non-continuously differentiable max-min fairness utility for the receiver i of group G_k defined in (3) with a mathematically well-behaved function over the real numbers axis while satisfying RUP. By mathematically well-behaved, we mean that our so-called extrapolated rational function $E(r_i, g_k)$ is continuously differentiable and has no poles over the real numbers axis. We select a function $E(r_i, g_k)$ in the form of

$$E(r_i, g_k) = \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \quad (9)$$

and note that not only $E(r_i, g_k)$ is well behaved for parameter a satisfying the boundary condition $-2 < a < 2$, but it satisfies the boundary and maximum conditions of function $F(r_i, g_k)$. The matter is best explained by a graphical illustration. Figure (1) shows generic sample plots of $F(r, g)$ and $E(r, g)$ versus g for a fixed r . It is important to note that since both $F(r, g)$ and $E(r, g)$ functions can transparently interchange the variables r and g we could consider the plots $F(r, g)$ and $E(r, g)$ versus r for a fixed g instead. Next, we employ least square error estimation technique to

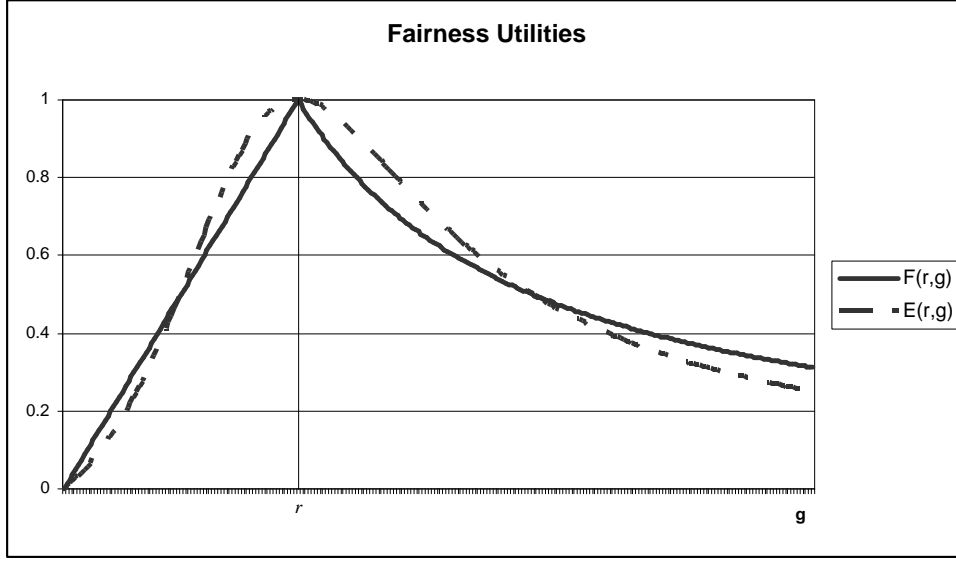


Fig. 1. Plots of $F(r, g)$ and $E(r, g)$ versus g for a fixed r .

find the optimum value of the parameter a within the interval of interest $[0, \frac{r_i}{1-L_i}]$ considering the constraint function of (8) and as shown below.

$$\begin{aligned} \min_a [\text{MLS}(a, r_i, L_i)] &\equiv & (10) \\ \min_a [&\int_0^{r_i} (\frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{g_k}{r_i})^2 dg_k \\ &+ \int_{r_i}^{\frac{r_i}{1-L_i}} (\frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{r_i}{g_k})^2 dg_k] \end{aligned}$$

Solving Equation (10) for different values of r_i and L_i in the intervals of interest reveals the range $[-1.6012, -1.5153]$ for the optimal value of parameter a . In our calculations, we perform a table look up operation to extract the optimal value of parameter a . Appendix I describes the details of the extrapolation technique.

We now formulate the new rate allocation problem with an extra constraint on the available bandwidth to individual groups of the session as

$$\max_{g_1, \dots, g_K} IRFA_{Total} \equiv \max_{g_1, \dots, g_K} \sum_{k=1}^K IRFA_k \quad (11)$$

$$= \max_{g_1, \dots, g_K} \sum_{k=1}^K \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2}$$

$$\text{Subject To: } g_k \leq \text{BWL}_k \quad k = 1, \dots, K \quad (12)$$

$$g_k \leq \text{BWF}_k \quad k = 1, \dots, K \quad (13)$$

where BWL_k in the constraint of Equation (12) is defined as $\text{BWL}_k \equiv \min_{i \in G_k} \frac{r_i}{1-L_i}$, the same as that of (8), and the constraint of Equation (13) indicates the available group bandwidth as the result of enforcing a per group inter-session fairness algorithm. Further, the function $IRFA_k$ is the group fairness utility defined as

$$IRFA_k \equiv \sum_{i \in G_k} E(r_i, g_k) = \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \quad (14)$$

By defining $BWA_k \equiv \min(\text{BWL}_k, \text{BWF}_k)$, we convert the rate allocation problem to

$$\begin{aligned} \max_{g_1, \dots, g_K} IRFA_{Total} &= \max_{g_1, \dots, g_K} \sum_{k=1}^K IRFA_k \\ &= \max_{g_1, \dots, g_K} \sum_{k=1}^K \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \end{aligned} \quad (15)$$

$$\text{Subject To: } g_k \leq BWA_k \quad k = 1, \dots, K \quad (16)$$

We note that it is important to distinguish between the loss tolerance constraints BWL_k and the group bandwidth upper bounds BWF_k . While the former reflects the receivers bandwidth processing capabilities, the latter is the result of employing a flow control mechanism with the objective of enforcing inter-session fairness among different flows.

III. PHASE 1: LMMC OPTIMAL SOLUTION TO THE RATE ALLOCATION PROBLEM

In this section, we provide an analytical solution to the optimal rate allocation problem formulated by Equation (15) and Constraint (16) that can be applied to both layered media and replicated media sessions. Appendix II includes the solution for another case in which an overall available bandwidth for the session is given instead of the available bandwidth to individual groups of the session. As a feasible approach, the general problem of Equation (15) and Constraint (16) can be converted to an optimization problem without constraints by defining a Lagrangian function in the form of

$$\begin{aligned} LG_{IRF} &= IRFA_{Total} + \sum_{k=1}^K \mu_k (g_k - BWA_k) \\ &= \sum_{k=1}^K IRFA_k + \sum_{k=1}^K \mu_k (g_k - BWA_k) \end{aligned} \quad (17)$$

where the parameters μ_k for $k = 1, \dots, K$ are the Lagrange multipliers in the Lagrangian Equation (17). The solution to the unconstrained problem can then be obtained by solving $\nabla LG_{IRF}|_{g^*} = 0$. However considering the specific form of the function $IRFA_{Total}$ and the constraint set of (16), the most straight forward way of solving for the optimal solution is to decompose the system of $2K$ equations and $2K$ unknowns obtained from $\nabla IRFA_{Total}(g^*) = 0$ and the constraints (16) into K pairs of independent equations. This is in essence equivalent to solving the set of K individual unconstrained problems of $\nabla IRFA_k(g_k^*) = 0$ and then investigating the impact of applying the corresponding inequality constraint $g_k \leq BWA_k$ on individual results. Equation (18) shows the simplified formulation applied to the set of K independent problems.

$$\begin{aligned} \max_{g_k} IRFA_k &= \max_{g_k} \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \\ \text{Subject To: } &g_k \leq BWA_k \end{aligned} \quad (18)$$

where $k = 1, \dots, K$. The set of optimal control problems of Equation (18) can be solved by finding the roots of the following equations,

$$\frac{\partial IRFA_k}{\partial g_k} = \sum_{i \in G_k} \frac{(2+a)r_i (r_i^2 - g_k^2)}{(g_k^2 + ar_i g_k + r_i^2)^2} = 0 \quad (19)$$

and extracting the global maximum from among the set of local optimum points satisfying Constraint (16) and

$$\frac{\partial^2 IRFA_k}{\partial g_k^2} = \sum_{i \in G_k} \frac{2(2+a)r_i (g_k^3 - 3r_i^2 g_k - ar_i^3)}{(g_k^2 + ar_i g_k + r_i^2)^3} \leq 0 \quad (20)$$

Prior to proceeding with the solution to individual optimization problems, we review the mathematical characteristics of the function $IRFA_k$. We first note that the function $IRFA_k$ is non-decreasing in the interval $[0, r_{k_{min}}]$ and non-increasing in the interval $[r_{k_{max}}, \infty]$ where $r_{k_{min}}$ indicates the minimum isolated rate and $r_{k_{max}}$ indicates the maximum

isolated rate of the receivers belonging to group G_k . This is true because the function $IRFA_k$ consists of a sum of a number of the receiver utility functions $E(r_i, g_k)$ which are all non-decreasing in the interval $[0, r_{k_{min}}]$ and non-increasing in the interval $[r_{k_{max}}, \infty]$. Consequently, Equation (19) has no roots in the intervals $[0, r_{k_{min}}]$ and $[r_{k_{max}}, \infty]$. We also remind that any acceptable optimum point has to satisfy Constraint (16). Combining the above conditions, we can argue that for $r_{k_{min}} \geq BW A_k$ the optimal solution equals to $BW A_k$ and for $r_{k_{min}} < BW A_k < r_{k_{max}}$ any acceptable maximum point falls into the interval

$$[r_{k_{min}}, BW A_k] \quad (21)$$

For $r_{k_{max}} < BW A_k$, Constraint (16) has no impact on the optimal solution.

Generally speaking, the function $IRFA_k$ can have up to N_k maximum points and N_{k-1} minimum points with N_k indicating the number of the receivers in group G_k . Finding the global maximum of the function $IRFA_k$ is hence equivalent to applying a root finding algorithm on Equation (19) and extracting the global maximum from the set of optimum points satisfying (20) and Constraint (16).

In practice and consistent with the extensive set of the numerical results reported in Section VI, the function $IRFA_k$ only consists of a single global maximum point for the sum of individual receiver utilities distributed in such a way that every two consecutive isolated rates r_i and r_{i+1} satisfy the relationship $r_{i+1} \leq 2r_i$. Figure (2) shows a typical $IRFA_k$ function. Finding the global maximum of the function $IRFA_k$ in such a case is hence equivalent to applying a single root finding algorithm such as bisection or Newton algorithms to Equation (19). These algorithms can identify the single root of Equation (19) with a time complexity of $\mathcal{O}(N \log N)$. We argue that if the media session has the

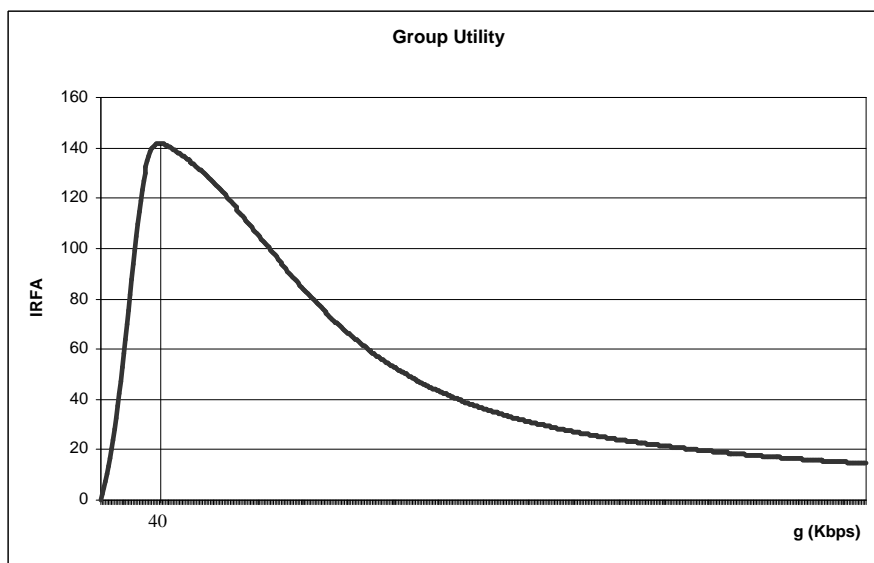


Fig. 2. A sample plot of the group utility $IRFA_k$ versus g_k for a group including 200 receivers with isolated rates in the range of $[32Kbps, 128Kbps]$ and every two consecutive isolated rates r_i and r_{i+1} satisfying $r_{i+1} \leq 2r_i$.

luxury of choosing the proper number of groups according to the distribution of receivers, all of the corresponding $IRFA_k$ functions will only have one maximum point. We also argue that having a limited number of groups can only impact the number of optimum points for the function $IRFA_K$ of the last group. To explain the latter claim, consider a scenario in which the receivers are distributed around S major categories of bandwidth while there are only K groups ($K < S$) are available to accommodate the receivers. A real example of this situation is when you have receivers belonging to the bandwidth range of dial-up, cable, $10Mbps$ LAN, and $100Mbps$ LAN while there are only 3 groups available due to multicasting constraints. In such a scenario, the bandwidth and loss characteristics of the receivers in the lower bandwidth ranges maps the first $K - 1$ bandwidth categories to the first $K - 1$ groups while combining the rest of $S - K + 1$ bandwidth categories in the last group. This creates a situation in which only the last group consists of a mix of receivers with significantly different bandwidth characteristics resulting in an $IRFA_K$ function with multiple optimum points. Additionally even in case of observing multiple maximum points for the function $IRFA_k$, our numerical results have only shown one maximum for any subset of receivers with isolated

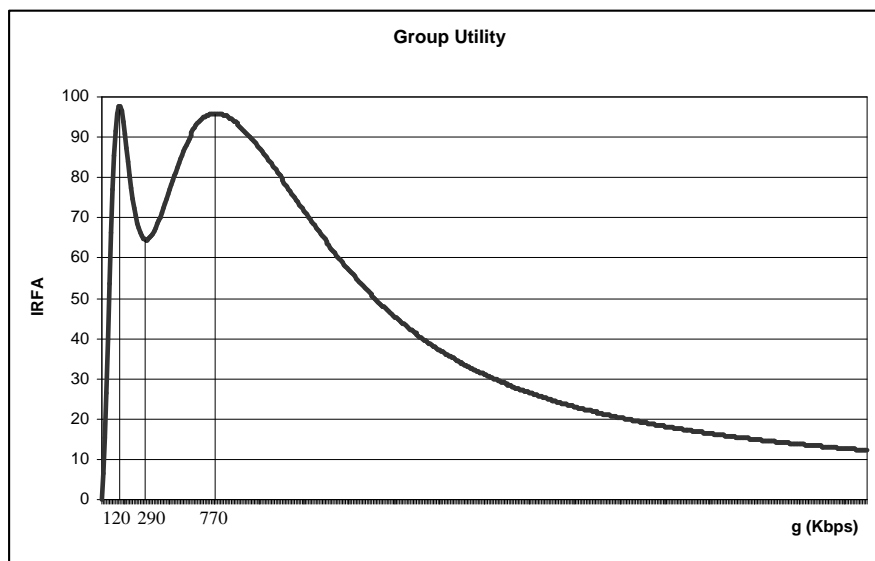


Fig. 3. A sample plot of the group utility $IRFA_k$ versus g_k for a group including 200 receivers with isolated rates in the range of $[64Kbps, 128Kbps]$ and $[640Kbps, 1280Kbps]$. In each interval, every two consecutive isolated rates r_i and r_{i+1} satisfy $r_{i+1} \leq 2r_i$.

rates satisfying $r_{i+1} \leq 2r_i$. Figure (3) shows an example of such an $IRFA_k$ function. We also observe that the maximum acceptable loss tolerance of a receiver typically does not exceed 50%. This implicitly means that BWA_k defined as $\min(BWL_k, BWF_k)$ with BWL_k defined as $\min_{i \in G_k} \frac{r_i}{1-L_i}$ will typically not exceed $2r_{kmin}$ where r_{kmin} indicates the minimum isolated rate of the receivers belonging to group G_k . Combining these observations, we come to the conclusion that in practical cases applying Constraint (16) limits the search to find the first optimum point of the function $IRFA_k$. Applying the interval of Equation (21), Newton, bisection or a similar numerical technique can be employed to find the first positive real maximum of $IRFA_k$ function.

As an important special case and by substituting BWA_k with BWL_k , the general formulation of our problem reduces to the no flow constraint problem formulated in [27] and [11]. The problem can then be solved using the same technique as the one used to solve the general problem. It is now relevant to compare the time complexity of our algorithm with that of [27]. In practice, the time complexity of solving for optimum point of equation set (19) over all of the existing groups is $\mathcal{O}(KN \log N)$. The search for the root of Equation (19) determines the overall time complexity of the solution considering the fact that the rest of calculations are in the complexity order of $\mathcal{O}(N)$. The time complexity of the algorithm is by far better than $\mathcal{O}(N^2)$ the complexity of the dynamic programming algorithm offered by [27]. This is aside from the fact that a dynamic programming approach in general does not provide an analytical solution to an optimization problem and the algorithm of [27] needs minor modifications to be able to solve the formulation of the general problem of Equation (15) considering the impact of enforcing a flow control algorithm.

Before we proceed to phase 2 of our solution, it is also relevant to investigate the impacts of facing some of the source and the receiver limitation scenarios when solving LMMC optimal control problem. First we consider a source limitation scenario that appears in the form of discrete sending rates. Up until now, we have assumed that there is no limitation on the source sending rates, i.e., the source can control the group rates with fine granularity. In practice, encoding techniques may limit the source to some pre-determined quantized discrete group rates. There are two ways to cope with this issue in our rate allocation problem. The first approach is to change the formulation of our optimization problem from a NonLinear Programming (NLP) to a Mixed Integer NonLinear Programming (MINLP) in which the group rates can only take on discrete values. The solution to the new problem will then satisfy the discrete constraints. The second approach is to rely on the continuous optimal solution of the existing formulation and approximate it with the closest discrete rate. Although the approximated solution is sub-optimal in this case, it reduces the complexity of the problem to a great extent and yields acceptable results for as long as the discrete rates are closely distributed. Considering distribution of the discrete group rates, we choose the second approach as the practical way of coping with this issue in our optimization problem. This method is also of special interest, considering the iterative nature of

our two-phase solution as described in Section V.

Next, we consider a scenario in which the receivers introduce a zero loss tolerance. The only impact of facing a zero loss tolerance scenario with $L_i = 0$ for $i = 1, \dots, N$ in our optimization algorithm is to change the definition of BWL_k from $\text{BWL}_k \equiv \min_{i \in G_k} \frac{r_i}{1-L_i}$ to $\text{BWL}_k \equiv \min_{i \in G_k} r_i$ for $k = 1, \dots, K$. Since the previous constraint qualifications hold for $0 \leq L_i < 1$ with $i = 1, \dots, N$ we do not foresee any changes on the method of obtaining our optimal solution. However, we make note that this scenario greatly simplifies the results considering the fact that the function of Equation (15) would have no zero slope point satisfying Constraint (16) for $N_k > 1$. Intuitively, we anticipate that the optimal rate of each group is always less than or equal to the lowest isolated rate of the group.

IV. PHASE 2: LMMC NEAR-OPTIMAL PARTITIONING STRATEGY

In [11], a heuristic approach to partition the receivers of a layered media session is proposed without introducing a formal algorithm. The heuristic method is well categorized under probabilistic classification and clustering methods for non-convex optimization problems. In Appendix III, we provide a formal classification method that is closely related to the partitioning heuristic rules. However, it is worth mentioning that the general short coming of probabilistic classification methods lies in the fact that they are typically appropriate for deduction techniques on the properties of mathematical concepts and closely related computational algorithms concepts rather than being useful for approximate or exact solutions to the optimization problems. Nevertheless, these techniques come handy in case of solving optimization problems and in the absence of a formal solution.

In addition, the dynamic programming algorithm of [27] provides an optimal receiver partitioning strategy for a media session while computing the optimal layer rates. The general problem of dynamic programming approaches is the lack of providing an analytical answer to an optimization problem and the relatively high degree of complexity. However, we make note of the fact that dynamic programming is one of the best optimization and in many cases the only available tool for solving an optimization problem. Fortunately, this is not the case for a typical rate allocation problem.

Rather than relying on a dynamic programming approach, we introduce a near-optimal partitioning strategy with time complexity of $\mathcal{O}(\mathcal{N})$ for a layered media or a replicated media session and show that our partitioning strategy maximizes the session utility for a set of given group rates.

The fact that the extrapolated receiver fairness function $E(r, g)$ satisfies RUP defined in Section II keeps the order of the resulting partitioning of this section.

Considering the general objective of maximizing the session utility of Equation (15), it is imperative that a receiver with isolated rate r_i belongs to the group G_k with the group rate g_k for a set of given group rates $\{g_1, \dots, g_K\}$ if the receiver utility defined in (9) is maximized for the choice of g_k . As the result, we make the observation that the optimal receiver partitioning strategy has to assign the receiver with the isolated rate r_i to the group G_k with the group rate g_k such that

$$E(r_i, g_k) \geq E(r_i, g_l) \quad l \in \{1, \dots, K\} \quad (22)$$

We now translate the latter observation to a simple group assignment mechanism. Let us first consider the fairness function of Equation (9) with parameter g_k and variable r_i . We note that in Sections II and III, the function of Equation (9) with parameter r_i and variable g_k was considered instead. Given the group rates $\{g_1, \dots, g_K\}$, we first plot the family of functions $E(r_i, g_k)$ versus r_i for different parameter values of g_k where $k = 1, \dots, K$. Figure (4) shows the sample plots for $K = 3$. Next, we find the intersection points of every two functions with consecutive group rates g_k and g_{k+1} . The values of r_i at the intersection points are obtained by finding the roots of the following equations for the variables r_i and parameters g_k, g_{k+1} where $k = 1, \dots, K$.

$$E(r_i, g_k) = E(r_i, g_{k+1}) \quad (23)$$

which yields

$$\frac{(2 + a(r_i))r_i g_k}{g_k^2 + a(r_i)r_i g_k + r_i^2} = \frac{(2 + a(r_i))r_i g_{k+1}}{g_{k+1}^2 + a(r_i)r_i g_{k+1} + r_i^2} \quad (24)$$

Although in the general form of Equation (24) the parameter a is a function of the variable r_i , the solution to the equation can nevertheless be expressed in the following form after a bit of algebraic manipulation as

$$r_i = \sqrt{g_k g_{k+1}} \quad (25)$$

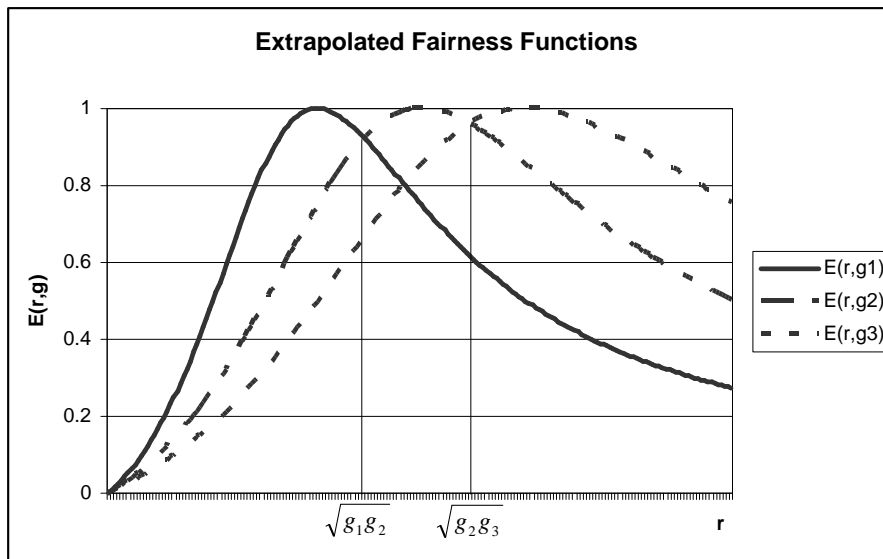


Fig. 4. Sample plots of $E(r_i, g_k)$ versus r_i for three given values of g_k .

We now pay special attention to the key characteristic of the intersection points of the curves that we refer to as partitioning thresholds.

Theorem 4.1: The value of the receiver utility as defined in Equation (9) is maximized for the choice of the group rate g_k for $k > 1$ and $k < K$ over the set of given group rates $\{g_1, \dots, g_K\}$ if $\sqrt{g_{k-1}g_k} < r_i \leq \sqrt{g_k g_{k+1}}$. The receiver utility is maximized for the choice of the group rate g_1 if $r_i \leq \sqrt{g_1 g_2}$ and for the choice of the group rate g_K if $r_i > \sqrt{g_{K-1}g_K}$.

Proof: As graphically observed in Figure (4), among the three functions $E(r_i, g_1)$, $E(r_i, g_2)$, $E(r_i, g_3)$ the value of the function $E(r_i, g_1)$ is the maximum if $r_i \leq \sqrt{g_1 g_2}$, the value of the function $E(r_i, g_2)$ is the maximum if $\sqrt{g_1 g_2} < r_i \leq \sqrt{g_2 g_3}$, and finally the value of the function $E(r_i, g_3)$ is the maximum if $r_i \geq \sqrt{g_2 g_3}$. The above observation graphically proves our claim for the partitioning of the receivers in case of three groups. The graphical proof remains the same by expanding partitioning thresholds from $\sqrt{g_1 g_2}$ to $\sqrt{g_{K-1} g_K}$ for any number of given groups K . **QED**

We now realize that Theorem (4.1) provides the best overall repartitioning strategy for an unconstrained problem. There is also another issue that needs to be addressed in case of solving the constrained problem of Equation (16). Considering the definitions of Equation (16) and Equation (12), the issue has to do with the fact that moving a receiver from group $k - 1$ to group k can potentially introduce a new constraint for group k . If the new constraint is far from the existing optimal group rate g_k^* , it can cause a reduction in the utility sum of groups $k - 1$ and k after repartitioning. There are two ways to resolve this issue. First, we can rely on statistical bounds to control the move of a receiver from group $k - 1$ to group k . In this case a receiver is allowed to move from group $k - 1$ to group k if **one** of the following conditions holds

$$\frac{r_i}{1 - L_i} \geq g_k^* \quad (26)$$

$$(\mu - C_1 \sigma < r_i) \quad \& \quad \left(\frac{r_i}{1 - L_i} < g_k^*\right)$$

where μ and σ are the mean and standard deviation of the receivers in group k . In practice, we have observed that setting $C_1 \in [0.9, 1]$ yields good results for different values of receivers' loss tolerance. Second, we can allow for moving a receiver from group $k - 1$ to group k only if the newly introduced constraint is satisfying a deviation from the existing group k optimal rate. In the second case a receiver is allowed to move from group $k - 1$ to group k if **one**

of the following conditions holds

$$\begin{aligned} \frac{r_i}{1-L_i} &\geq g_k^* \\ C_2 g_k^* < \frac{r_i}{1-L_i} < g_k^* \end{aligned} \quad (27)$$

In practice, we have observed that setting $C_2 \in [0.5, 0.9]$ yields good results for different values of receivers' loss tolerance. Note that, although it is unlikely for the same issue to reveal when moving a receiver from group k to $k-1$, a similar approach can be used to avoid the problem.

LMMC near-optimal partitioning algorithm¹ then reorders the receivers such that each receiver is moved to a group maximizing its individual utility according to Theorem (4.1) and one of the conditions (26) or (27). Such an algorithm introduces a time complexity order of $\mathcal{O}(KN)$. As an alternative and to achieve more rapid convergence, we can also obtain the new optimal rate of the corresponding group of receivers while repartitioning. This is due to the fact that changing the partitioning thresholds yields a different optimal group rate for the group of receivers affected by the change in the sequence. Considering the added complexity for solving yet another optimal control problem, this version of the algorithm introduces a time complexity order of $\mathcal{O}(KN \log N)$. The trade off between the two versions of the algorithm is the speed of convergence versus increased complexity. In practice, one selects the latter over the former if the higher speed of convergence justifies the increased complexity of the latter version. Otherwise, the former version is preferred. The second version of the optimal partitioning algorithm is summarized below. The first version is simply obtained by eliminating the last step of the loop.

LMMC Near-Optimal Partitioning Algorithm:

For every group of the media session and assuming the group rates $\{g_1, \dots, g_K\}$ are given,

for ($k = 2$ to K) {

- Calculate the partitioning threshold $\sqrt{g_{k-1}g_k}$.
- Repartition groups $k-1$ and k . For every receiver belonging to groups $k-1$ or k and isolated rate r_i , assign the receiver to group k if $r_i > \sqrt{g_{k-1}g_k}$ and one of the conditions (26) or (27) hold. Otherwise, assign the receiver to group $k-1$.
- Calculate the new optimal sending rate of group k according to the new partitioning.

} /* for ($k = 2$ to K) */

The other interesting characteristic of the intersection points of Equation (23) is that they remain the same for both the approximate and original fairness functions of Equation (9) and Equation (3). The latter is verified by observing that the partitioning thresholds of Equation (23) are also the intersection points of the fairness functions of Equation (3) for different values of g_k from the following equation

$$\frac{\min(r_i, g_k)}{\max(r_i, g_k)} = \frac{\min(r_i, g_{k+1})}{\max(r_i, g_{k+1})} \quad (28)$$

We conclude that the general algorithm of this section can be used in conjunction with any rate allocation algorithm by properly identifying partitioning thresholds. In specific, the algorithm of this section can also be used with a rate allocation algorithm relying on the fairness function of (4) in order to reach the optimal partitioning assuming a given set of group rates.

V. LMMC NEAR-OPTIMAL ITERATIVE SOLUTION

In this section, we introduce an iterative approach that can reach a near-optimal solution considering the fact that the solution to our two-phase optimal problem is sub-optimal due to the impact of our phasing approach. A near-optimal solution can be achieved by iteratively applying the results of each phase as an existing condition to obtain the solution of the other phase. This is equivalent to applying the partitioning results of the second phase to the first phase and

¹We note that LMMC partitioning strategy is optimal in case of solving the unconstrained problem. Although we use the term near-optimal to mathematically consider the effect of solving the constrained problem, LMMC partitioning strategy is a very close approximation of the optimal solution in case of solving the constrained problem.

solving the optimal rate allocation problem again with the alternative partitioning strategy. The optimal layer rates of the first phase can then be applied to the near-optimal partitioning strategy of the second phase to partition the receivers according to the new set of rates. In what follows we propose the formal iterative algorithm of LMMC and prove that it yields a near-optimal solution considering the necessary condition for optimality defined below holds.

Remind that for a media session with N receivers, K groups and the group rate set $g = \{g_1, \dots, g_K\}$, a set $P = \{G_1 | \dots | G_K\}$ is called a partitioning of the receiver set $R = \{1, \dots, N\}$ if P is a decomposition of the set R into a family of disjoint sets. The necessary and sufficient condition for optimality is now defined over the partitioning P^* and the group rate set g^* such that

$$IRFA_{Total}(P^*, g^*) \geq IRFA_{Total}(P, g) \quad (29)$$

for every $P \neq P^*$ and $g \neq g^*$. Considering the impact of LMMC phasing approach, the necessary condition for optimality is defined for the combination of two individual phases. In the first phase, we consider a fixed partitioning P_{fixed} and define the group rate set g^* such that

$$IRFA_{Total}(P_{fixed}, g^*) \geq IRFA_{Total}(P_{fixed}, g) \quad (30)$$

for every $g \neq g^*$. In the second phase, we consider a fixed group rate set g_{fixed} and define the partitioning P^* such that

$$IRFA_{Total}(P^*, g_{fixed}) \geq IRFA_{Total}(P, g_{fixed}) \quad (31)$$

for every $P \neq P^*$.

LMMC Iterative Rate Allocation-Partitioning Algorithm:

- Step 1: Start from an initial ordered partitioning of the receivers by uniformly distributing the receivers among the existing groups. In addition, set the initial iteration number $j = 0$ and the maximum number of iterations j_{max} .
- Step 2: Calculate the optimal group rates $g^* = \{g_1^*, \dots, g_K^*\}$ and the resulting session utility $IRFA_{Total}$ by numerically solving the system of equations (19) while satisfying conditions (20) and (16). Save the previously calculated $IRFA_{Total}$ in variable q_1 and the currently calculated $IRFA_{Total}$ in variable q_2 .
- Step 3: If $\frac{|q_1 - q_2|}{q_1} < \delta$ or $j > j_{max}$ STOP.
- Step 4: *for* ($k = 2$ to K) {
 - Calculate the partitioning threshold $\sqrt{g_{k-1}g_k}$.
 - Repartition groups $k - 1$ and k . For every receiver belonging to groups $k - 1$ or k and isolated rate r_i , assign the receiver to group k if $r_i > \sqrt{g_{k-1}g_k}$ and one of the conditions (26) or (27) hold. Otherwise, assign the receiver to group $k - 1$.
 - Calculate the new optimal sending rate of group k according to the new partitioning.
 } **/ for* ($k = 2$ to K) **/*
- Step 5: Go back to Step 2.

In the algorithm above the initial conditions are chosen in the first step. While the second step solves the optimal rate allocation problem of the first phase in our two-phase approach, the third step merely checks to terminate the algorithm according to the specified conditions. The fourth step includes the solution to the second phase near-optimal partitioning approach while adjusting the optimal rate of the corresponding group according to the new partitioning. We note that the time complexity of our iterative algorithm is $\mathcal{O}(IKN \log N)$ where I indicates the number of iterations. Comparing overall complexity of LMMC algorithm with that of the dynamic programming algorithm of [27] $\mathcal{O}(N^3)$, LMMC algorithm achieves much lower complexity.

Theorem 5.1: The convergence of “LMMC Iterative Rate Allocation-Partitioning Algorithm” mentioned in this section is guaranteed.

Proof: Let us make note of the fact that the session utility of Equation (15) consists of a finite number of fairness functions, one for each receiver. These functions are all positive, minimized at the value of zero, and maximized at the

value of one. Consequently, the positive session utility function of Equation (15) has both a lower bound and an upper bound. Next, we observe that the session utility function of Equation (15) can only increase in each step considering the operating mechanism of the individual phases of our optimization algorithm. Therefore, the sequence of utility function values at each step of the algorithm is a non-decreasing sequence with an upper bound equal to the number of fairness functions. We also note that any non-decreasing sequence with an upper bound would converge to a finite number also known as a fixed point. We, hence, conclude that LMMC iterative approach converges to a fixed point.

QED

Intuitively, LMMC algorithm is employing steepest descent optimal control strategy and is guaranteed to reach a near-optimal point if such a point exists. It is important, however, to note the followings.

First, we note that the “LMMC Iterative Rate Allocation-Partitioning Algorithm” mentioned in this section converges to a local optimum in case of solving the unconstrained problem. The claim is accurate considering the fact that the sequence of session utility functions of Equation (15) converges to a fixed point satisfying necessary conditions of (30) and (31) for optimality. We remind that we only claim reaching a near-optimal solution in case of solving the constrained problem because of applying one of the conditions of (26) or (27). However, we conjecture that the proper choice of the parameters in (26) and/or (27) leads to reaching a local optimal solution as shown by our numerical results.

In practice, the use of the iterative method is a factor of time complexity and the speed of convergence. The iterative method can be effectively deployed in environments with moderate variations of the available per flow bandwidth. As an example, the scenarios encountered in admission control problems can be mentioned in which the assignment of per flow bandwidth is relatively stable. In environments with rapidly varying available bandwidth, the sub-optimal solution with few or no iteration may be deployed.

It is obvious that the initial choice of the partitioning strategy plays a crucial role in the convergence speed of the algorithm. As a practical alternative, the classification method of Appendix III may be deployed as the partitioning strategy. Additionally, it is important to note that considering the convergence speed of our proposed algorithm as proven by the steepest descent approach and supported by the simulation results of Section VI, in most cases the use of our proposed algorithms yields fast converging results.

VI. NUMERICAL PERFORMANCE ANALYSIS

In this section, we present the numerical results of applying LMMC partitioning and rate allocation algorithms to a number of layered media scenarios and compare them with those of dynamic programming algorithm of [27]². We review the performance of both approaches from the stand point of tracking the maximum value of the utility function, time complexity indicated by experiment runtime, and space complexity indicated by memory allocation. Additionally, we review the scalability of the techniques by covering a relatively broad range of multicast group sizes ranging from hundreds to thousands of receivers. In our simulations, we rely on generalizations of normal distribution namely tri-, quad-, and pent-modal distributions to generate receiver isolated rates. We select the means of distributions from the set of $\{128Kbps, 1Mbps, 10Mbps, 100Mbps, 1Gbps\}$ to properly represent ISDN, Cable/DSL, low-speed LAN, high-speed LAN, and Gigabit LAN users. For each distribution, we also set the standard deviation of the distribution at 20% of the mean value. Considering the location of the means, the choice of standard deviations yields successive distributions remain disjoint with a certainty better than 99.7%. In our experiments, we make use of a host server with a 1.8GHz Pentium 4 CPU, 512MB of physical memory and 1GB of virtual memory.

We remind that the time complexity of the iterative optimal control LMMC algorithm is $\mathcal{O}(IKN \log N)$ where I indicates the number of iterations and the time complexity of the dynamic programming (DP) algorithm of [27] is $\mathcal{O}(N^3)$. In addition, the space complexity of the LMMC algorithm in our implementation is $\mathcal{O}(N)$ where as the space complexity of the DP algorithm proposed in [27] is $\mathcal{O}(N^2)$. In our simulations, we ran in excess of 5000 experiments with different number of groups K , different group sizes N , and different receiver loss tolerance values. Figures (5) through (16) compares sample results of the LMMC algorithm with the DP algorithm of [27]. In each experiment, we have considered the same loss tolerance for all of the receivers of the session. Different figures have been obtained for

²In our simulations, we relax the flow control constraint and assume $BWA_k = BWL_k$. The impact of applying the flow control constraint is to only change the value of the constraint BWA_k .

different choices of loss tolerance set at 10%, 20%; and the number of groups set at 3, 4 and 5. The x-axis of each curve is always in logarithmic scale and includes values of N from the set $\{100, 300, 1000, 3000, 10000, 30000, 100000\}$. Each figure consists of two pairs of curves. The first pair of curves compare fairness results of the two techniques.

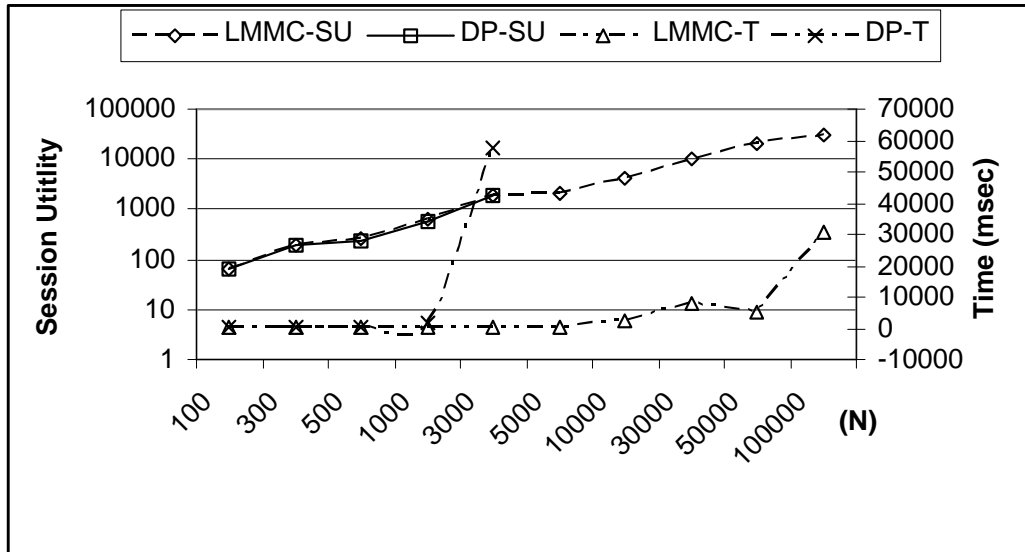


Fig. 5. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 3 categories of receiver isolated rates, $K=2$, and loss tolerance of 10%.

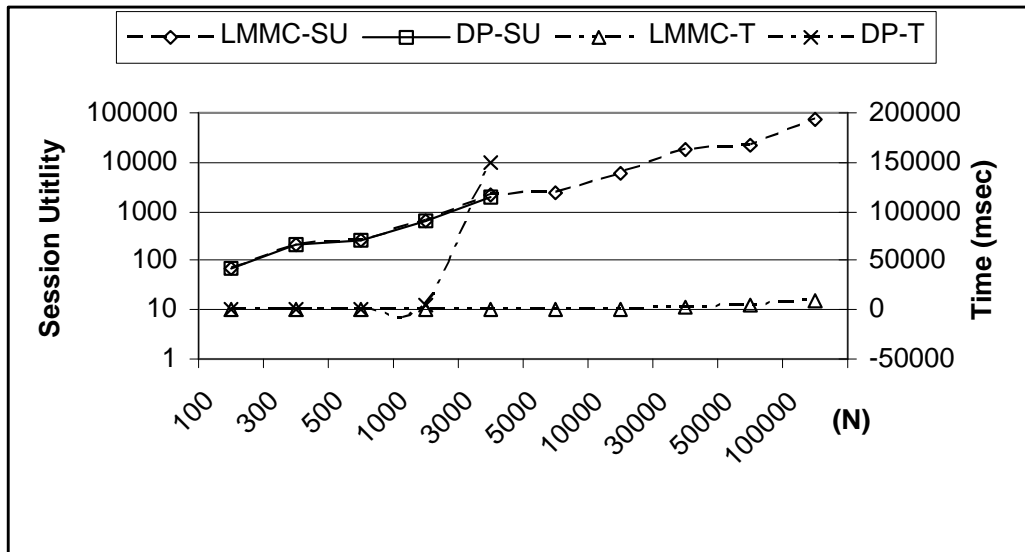


Fig. 6. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 3 categories of receiver isolated rates, $K=2$, and loss tolerance of 20%.

In order to do a fair comparison, we have used the fairness function of Equation (9) for LMMC and the fairness function of Equation (3) for DP. Since the maximum of each individual receiver utility is the value 1, the number N indicates the corresponding upper bound on the fairness for both techniques. A review of the sample results of the figures shows a difference of less than 10% between the raw session utility values of the LMMC and the DP algorithms. Considering the fact that the fairness function of Equation (9) is only an approximation of the fairness function of Equation (3) in the interval of interest, it is in order to mention that the session utility value is only a relative metric of performance comparison. Our overall conclusion is that both of the techniques are capable of tracking a maximum satisfying the existing constraints.

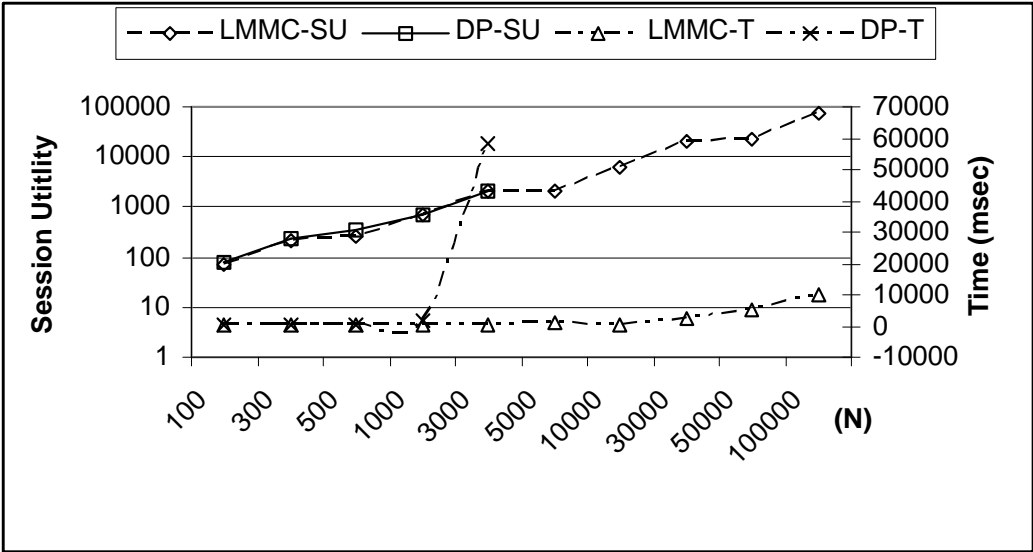


Fig. 7. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 3 categories of receiver isolated rates, K=3, and loss tolerance of 10%.

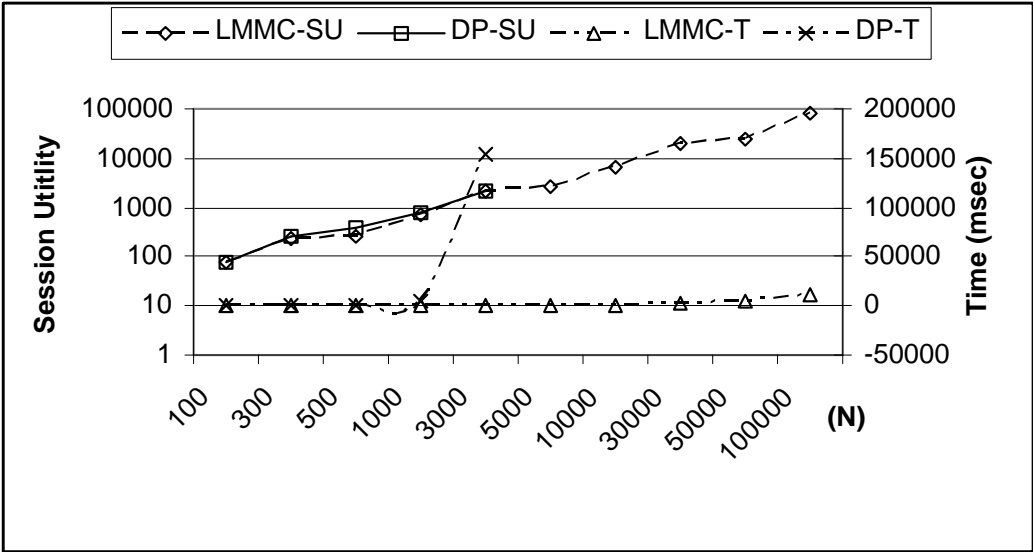


Fig. 8. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 3 categories of receiver isolated rates, K=3, and loss tolerance of 20%.

The second pair of curves display the runtime of the experiments as an indicator of the time complexity of the two techniques. In this area, a review of the results reveals the great performance advantage of LMMC over DP. We observe a nonlinear increase in the runtime of the DP algorithm where as LMMC algorithm curve indicates a linear increase. We also note that in each figure, the pair of the DP algorithm curves end at the value of 3000 receivers.

This is explained in terms of the time complexity and the space complexity of the DP algorithm. We argue that an increase in the value of N increases the runtime of the algorithm proportional with the third power of N and consumes the memory proportional with the second power of N . In our experiments, the impacts of coping with higher time complexity and space complexity become significant for media sessions with more than 1000 receivers. The space complexity analysis also justifies the fact that we have not been able to run any experiment deploying DP algorithm for media sessions with 10000 or more receivers. We argue that although the specific numbers of our experiments are closely related to the capabilities of our host server, the same qualitative behavior is observed in general. It is obvious from our results that Bellman’s **curse of dimensionality** defined in [3] shows its impact much more rapidly in case of

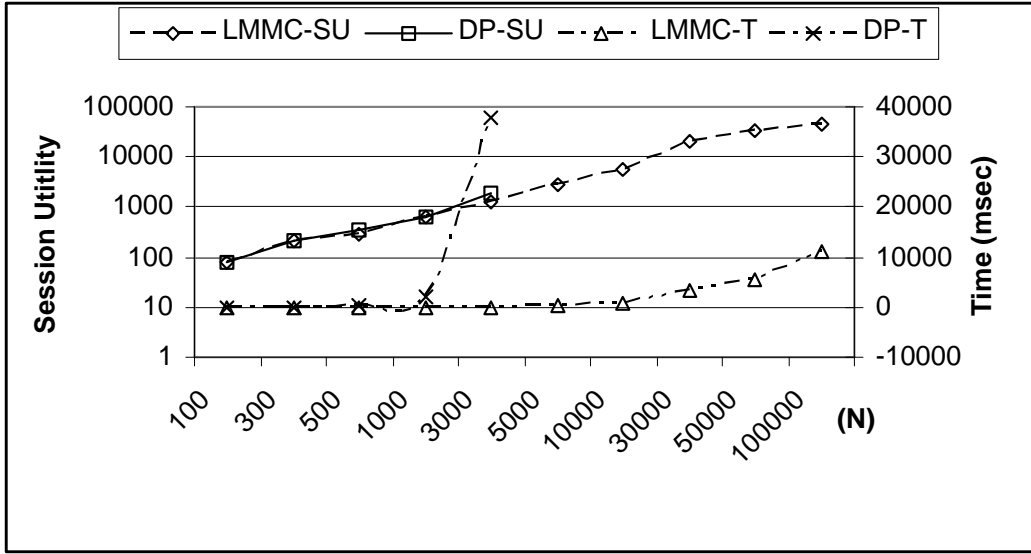


Fig. 9. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 4 categories of receiver isolated rates, K=3, and loss tolerance of 10%.

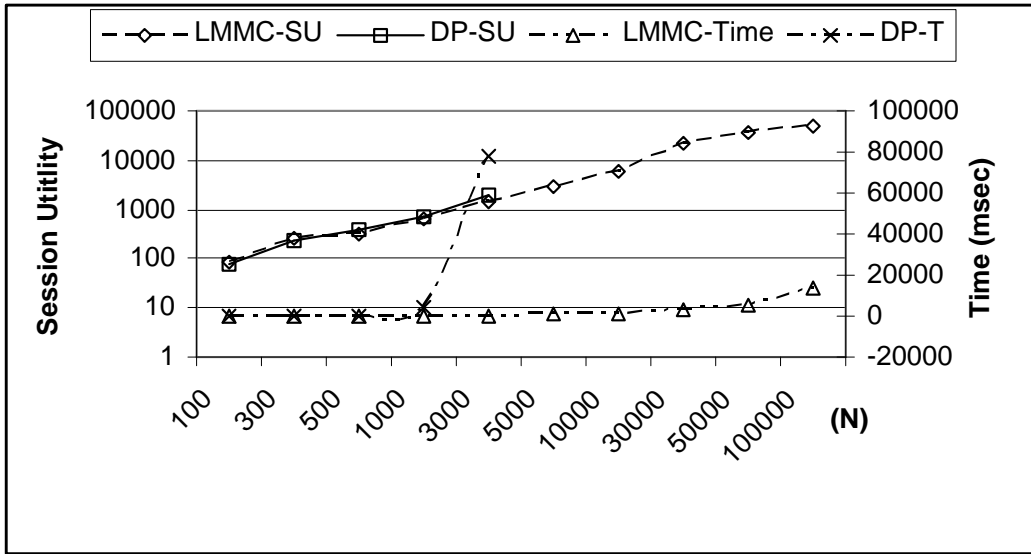


Fig. 10. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 4 categories of receiver isolated rates, K=3, and loss tolerance of 20%.

DP algorithm than the case of LMMC algorithm.

Finally, we would like to review the impacts of using criteria of Equation (26) and Equation (27) in controlling successive groups repartitioning. Generally speaking, we have observed that the proper choice of coefficients C_1 in Equation (26) and C_2 in Equation (27) mostly depends on the loss tolerance. The coefficients have to be chosen such that they enforce a narrower bound for smaller values of loss tolerance and a wider bound for larger values of loss tolerance. In our experiments using criteria of Equation (27) has yielded better results than using criteria of Equation (26). We have experimentally observed that for a loss tolerance of 10% a value of $C_2 = 0.790$ best controls the repartitioning process while for a loss tolerance of 20% a value of $C_2 = 0.885$ provides best repartitioning results. We have also observed that smaller values of loss tolerance increase the number of iterations required for the convergence of LMMC. This is explained considering the fact that smaller values of loss tolerance typically yield narrower bounds in Equation (26) and Equation (27) utilized to control the move of receivers from group $k - 1$ to group k in each iteration. In general, utilizing narrower control bounds results in a higher number of iterations required for convergence. It is also

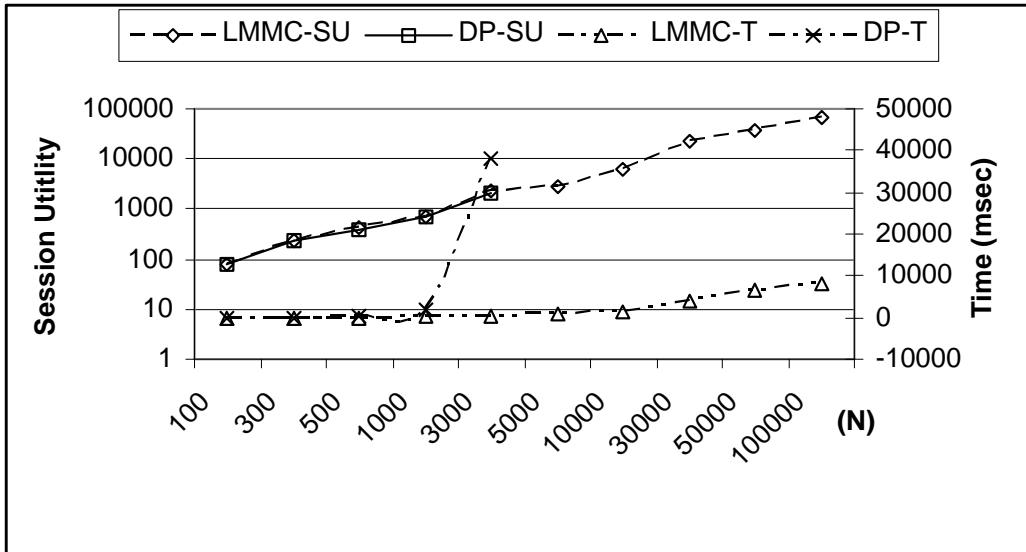


Fig. 11. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 4 categories of receiver isolated rates, $K=4$, and loss tolerance of 10%.

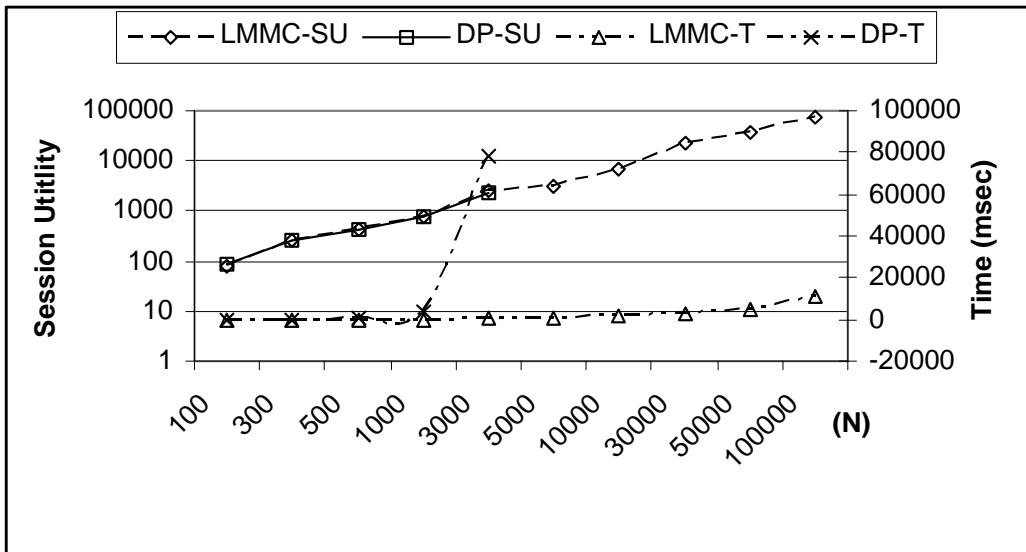


Fig. 12. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 4 categories of receiver isolated rates, $K=4$, and loss tolerance of 20%.

worth mentioning that the distribution of receivers isolated rates plays an important role in the speed of convergence for both LMMC and DP algorithms.

In the rest of this section, we briefly discuss some of the practical issues. Although in this study we did not discuss many of the practical aspects of implementing LMMC technique, we have implicitly assumed the use of most of the known techniques in the course of implementation. First, we need to apply the comparison analysis of source centric and receiver centric methods to LMMC algorithms. Considering the coordination necessary to synchronize the operation between the sender and receivers in LMMC algorithm, it is classified under hybrid algorithms with the main focus on the sender. Next, we need to consider the issue of feedback implosion in the process of collecting the isolated rates and loss tolerance of the receivers of a large multicast group. We can address feedback implosion issue either as an end-to-end or as an intermediate issue. In the former case, we can deploy a selective feedback mechanism from the receivers to the source of the session. In the latter case, we can force the receivers to report their isolated rates and

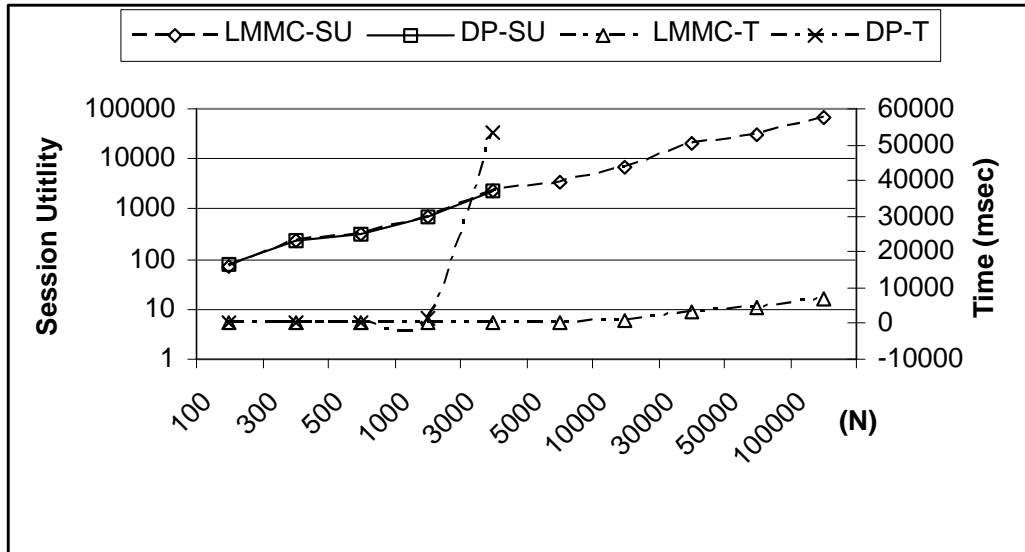


Fig. 13. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 5 categories of receiver isolated rates, $K=4$, and loss tolerance of 10%.

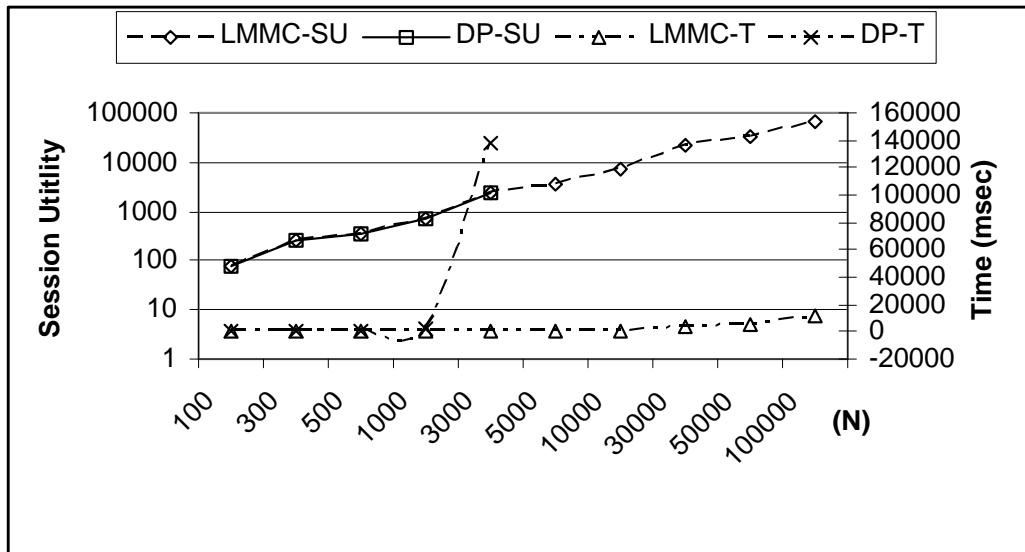


Fig. 14. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 5 categories of receiver isolated rates, $K=4$, and loss tolerance of 20%.

loss tolerance to their parent routers in the multicast tree. The routers can then send aggregated feedback messages to the source in multiple intervals. As an example, the feedback suppression technique proposed in [7] can be used to suppress feedback implosion when practically implementing our algorithms.

Finally, we need to discuss the impact of increasing the number of layers in the extrapolated fairness utility of the overall session. In general, we find consistent results in our numerical analysis with what was reported in [27], i.e., in most cases one can achieve the best combination of receiver heterogeneity accommodation and protocol complexity by choosing 3 to 5 layers. We would also like to add that the best fairness results are typically obtained if the number of groups matches to the number of bandwidth ranges in which receiver isolated rates are distributed. In the latter scenario, each of the ranges can capture the bandwidth characteristics of a group of receivers. For example, receivers with isolated rates distributed in the range of $64Kbps$ indicate dial-up users, receivers with isolated rates distributed in the range of $1Mbps$ indicate Cable/DSL users, and receivers with isolated rates distributed in the range of $100Mbps$

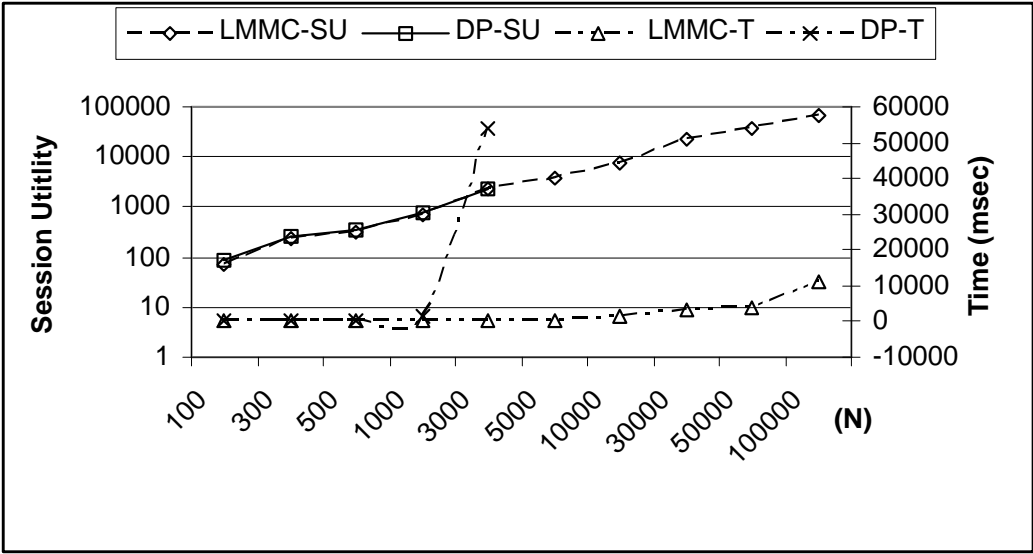


Fig. 15. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 5 categories of receiver isolated rates, K=5, and loss tolerance of 10%.

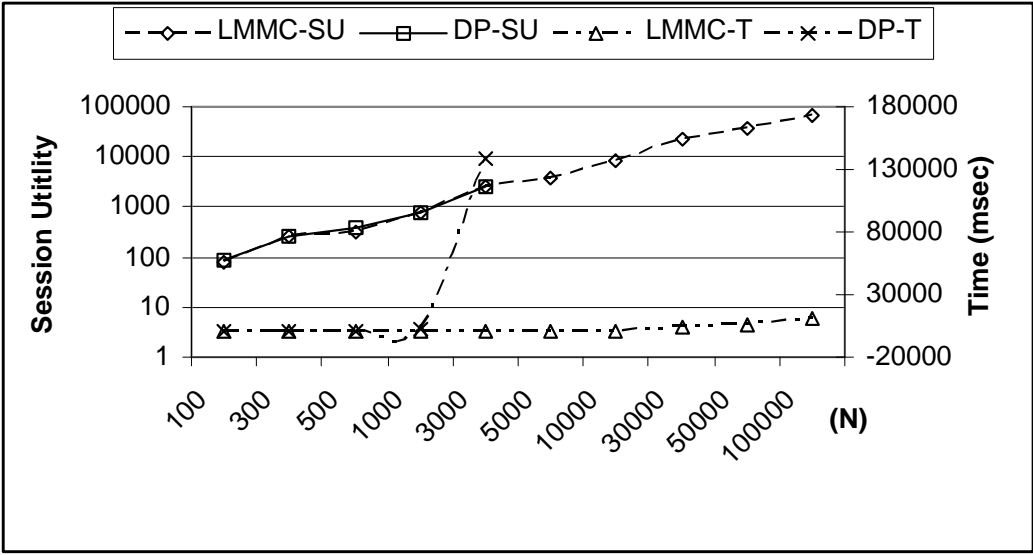


Fig. 16. Session Utility (SU) and Time (T) comparison of LMMC and DP versus number of receivers (N) for 5 categories of receiver isolated rates, K=5, and loss tolerance of 20%.

indicate fast LAN users. We make a practical observation that currently the number of these ranges does not exceed 5 considering the available bandwidths from dial-up, ISDN, Cable/DSL, Ethernet, and fast Ethernet. With the popularity of faster switched network interface such as Gigabit Ethernet and the obsolescence of slower switched network interfaces the number of the groups has to be proportionally adjusted in order for algorithms such as ours to provide best fairness results.

VII. CONCLUSION

In this paper, we studied the Layered Media Multicast Control (LMMC) solution to the general problem of optimal partitioning and rate allocation for layered and replicated media systems over multicast IP networks. We assumed the existence of congestion and flow control mechanisms specifying the fair bandwidth available to the media session. We aimed at providing a two-phase iterative solution converging to a near-optimal solution. We solved the general problem of receiver partitioning and layering rate allocation by means of extrapolating max-min fairness utilities of individual

receivers. We formulated LMMC rate allocation and partitioning problems as a two-phase optimal control problem. In the first phase, we calculated the optimal rates allocated to the individual layers of a media session in an analytical form assuming a given initial partition. In the second phase we introduced a formal approach that obtained the best partitioning of the receivers based on optimal allocated rates of the first phase while adjusting the optimal rates of the resulting groups. We showed that the partitioning strategy of the second phase is near-optimal for the specific allocated optimal layered rates.

Considering the impact of LMMC phasing approach, we introduced an iterative method in which a near-optimal solution could be achieved by iteratively applying the results of each phase to another. We discussed that low complexity LMMC algorithms are ideal for dynamically solving the partitioning and rate allocation problems in media systems where the underlying network environments are subject to frequent changes in available fair bandwidth.

Considering the scalability of LMMC techniques, we showed that the techniques could be effectively adopted in different size point-to-multipoint groups as well as different speeds of load change in the network. We also demonstrated that LMMC techniques were able to achieve a very close approximation of max-min fairness considering the heterogeneity issue and the varying loss characteristics of different receivers. Finally, we evaluated the performance of LMMC solution and illustrated its applicability in realistic network topologies through the use of simulations.

At the end, we would like to briefly review the different issues involved in the distribution of media systems over multicast networks. Generally speaking, the main issues involved with media multicast systems are rate allocation, receiver partitioning, and end-to-end error control. In addition, practical considerations such as scalability and feedback implosion have always played a key role in applicability of any media distribution algorithm. We are currently working on a bundled solution integrating LMMC rate allocation and receiver partitioning technique with a counter part end-to-end error control technique for media systems. In addition, we are working on adding a real-time analysis of receivers isolated rates such that LMMC partitioning technique can dynamically select the best parameters when applying conditions of (26) and/or (27).

APPENDIX I

LEAST SQUARE ERROR EXTRAPOLATION OF THE MAX-MIN FAIRNESS FUNCTION

In this appendix, we introduce a least square error extrapolation technique for the max-min fairness function of Equation (3). The objective of our extrapolation technique is to provide an estimated function $E(r_i, g_k)$ of function $F(r_i, g_k)$ that minimizes the surface between the two curves shown in Figure (1). We select a rational function of g_k and r_i in the form of

$$E(r_i, g_k) = \frac{N(r_i, g_k)}{D(r_i, g_k)} \quad (32)$$

where $N(r_i, g_k)$ and $D(r_i, g_k)$ are polynomials of r_i and g_k . Without loss of generality and to simplify the calculation, let us treat the variable r_i as a parameter and obtain the function $E(g_k) = E(r_i, g_k)$ assuming $\text{Deg}(N(r_i, g_k)) = \text{Deg}(N(g_k)) = M - 1$ and $\text{Deg}(D(r_i, g_k)) = \text{Deg}(D(g_k)) = M$ with respect to g_k and for the parameter r_i . The simplest rational function $E(g_k)$ behaving close to $F(r_i, g_k)$ is resulted by considering $M = 2$ in the form of

$$E(g_k) = \frac{N(g_k)}{D(g_k)} = \frac{bg_k}{a_2g_k^2 + a_1g_k + a_0} \quad (33)$$

In the above equation, the parameters, $b, a_0, a_1,$ and a_2 are obviously functions of the parameter r_i . The following conditions assure that not only $E(g_k) = E(r_i, g_k)$ is well behaved according to the description of Section II, but it satisfies the boundary and maximum conditions of function $F(r_i, g_k)$.

$$\begin{aligned} b &> 0, & a_0 &> 0, a_1 \geq 0, a_2 > 0 \\ E(0) = 0 &\Rightarrow a_0 \neq 0 \\ E(\infty) = 0 &\Rightarrow \text{Deg}(N(r)) < \text{Deg}(D(r)) \\ E(r_i) = 1 &\Rightarrow a_2r_i^2 + (a_1 - b)r_i + a_0 = 0 \\ E'(r_i) = 0 &\Rightarrow a_2r_i^2 + a_1r_i + a_0 - r_i(2a_2r_i + a_1) \end{aligned}$$

$$\begin{aligned}
&= -a_2 r_i^2 + a_0 = 0 \\
&\Rightarrow a_0 = a_2 r_i^2 \\
\Delta D(r_i) < 0 &\Rightarrow a_1^2 - 4a_0 a_2 < 0 \\
&\Rightarrow |a_1| < 2\sqrt{a_2 a_0}
\end{aligned} \tag{34}$$

Without loss of generality, we assume that $a_2 = 1$ and $a_1 = ar_i$. Applying the conditions of Equation (34) to the general form of Equation (33) introduces the specific form of

$$E(r_i, g_k) = \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \tag{35}$$

with the boundary condition $-2 < a < 2$ for the function $E(r_i, g_k)$. We note that the optimum choice of parameter a yields the best least square estimate for the original fairness function $F(r_i, g_k)$ defined in (3). Applying least square estimation technique in the interval of interest $[0, \frac{r_i}{1-L_i}]$ considering the constraint function of (8) yields the optimum value for parameter a in terms of parameters r_i and L_i .

$$\begin{aligned}
\min_a [\text{MLS}(a, r_i, L_i)] &\equiv \\
\min_a [&\int_0^{r_i} \left(\frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{g_k}{r_i} \right)^2 dg_k \\
&+ \int_{r_i}^{\frac{r_i}{1-L_i}} \left(\frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{r_i}{g_k} \right)^2 dg_k]
\end{aligned} \tag{36}$$

Equation (37) provides a closed form for the function $\text{MLS}(a, r_i, L_i)$. The solution to Equation (36) can be obtained by choosing the parameter a resulting in the least value for the function $\text{MLS}(a, r_i, L_i)$ calculated from Equation (37) over a uniform partitioning of the interval $(-2, 2)$. The granularity of the partitioning depends on the desired precision in the numerical algorithm.

$$\begin{aligned}
\text{MLS}(a, r_i, L_i) &= r_i \left(L_i + \frac{1}{2} \right) \\
&+ r_i \frac{a+2}{a-2} \left[(4-a) - \frac{(1-L_i)(a^2+a(1-L_i)-2)}{L_i^2-(a+2)L_i+(a+2)} \right. \\
&\quad \left. + a(a-2) \log(a+2) \right] \\
&- r_i \frac{a+2}{a-2} \left[\frac{2(a^3-2a^2-2a+6)}{\sqrt{4-a^2}} \arctan\left(\sqrt{\frac{2-a}{2+a}}\right) \right. \\
&\quad \left. - \frac{4(a-1)}{\sqrt{4-a^2}} \arctan\left(\frac{-L_i}{2-L_i} \sqrt{\frac{2-a}{2+a}}\right) \right]
\end{aligned} \tag{37}$$

Alternatively, a single non-parametric optimum value for parameter a is the one minimizing the integral of Equation (36) for a fixed value of loss tolerance L_i and calculated over a continuous range of isolated rates from 0 to r_{max} where r_{max} indicates the maximum feasible value of the receivers isolated rates. Considering the available bandwidth ranges, a feasible value for r_{max} is 1 Gbps.

$$\begin{aligned}
\min_a [\text{MLS}(a)] &= \\
\min_a [&\int_0^{r_{max}} \int_0^{r_i} \left(\frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{g_k}{r_i} \right)^2 dg_k dr_i \\
&+ \int_0^{r_{max}} \int_{r_i}^{\frac{r_i}{1-L_i}} \left(\frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{r_i}{g_k} \right)^2 dg_k dr_i]
\end{aligned} \tag{38}$$

In practice, we have observed that the optimal value of parameter a is only a function of parameter L_i . In other words, the optimal value of parameter a remains the same for a fixed value of parameter L_i and different choices of parameter r_i in the interval of interest for receiver isolated rates. Figure (17) plots the optimal value of parameter a versus the loss tolerance percentage L_i . Reviewing the figure in the interval of interest $L_i \in [0\%, 50\%]$ reveals that the optimal value of parameter a is in the range of $[-1.6012, -1.5153]$. In our calculations, we extract the optimum a by performing a simple table look up operation.

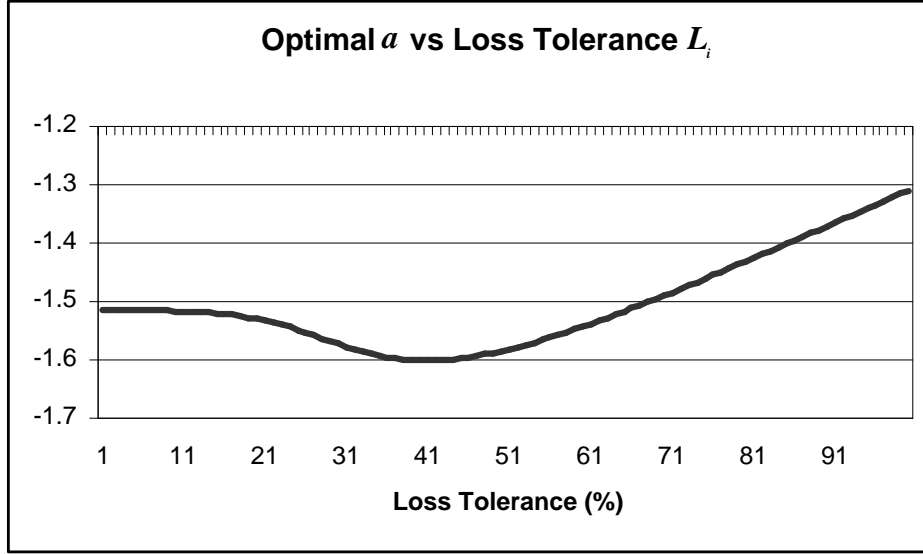


Fig. 17. The plot of optimal a versus loss tolerance L_i .

APPENDIX II

LMMC OPTIMAL SOLUTION TO THE RATE ALLOCATION PROBLEM WITH AN OVERALL AVAILABLE SESSION BANDWIDTH CONSTRAINT

In this appendix, we provide an analytical solution to the optimal rate allocation problem formulated by Equation (11), Constraint (12), and a new constraint replacing Constraint (13). We consider a scenario in which the overall available session bandwidth is given instead of the available bandwidths of the individual groups. We investigate the solution to this problem for both layered media and replicated media sessions. The interpretation of the problem for layered media sessions is fairly straight forward. First, we note that the constraint set of Equation (13) is reduced to a single constraint in the form of

$$g_K \leq \text{BWF}_K \quad (39)$$

considering the fact that the group rate g_K is the aggregated rate of layers $1, \dots, K$ according to Equation (1). The equation sets (15) and (16) can then be solved the same way as described in Section III by simply substituting $\text{BWA}_k = \text{BWL}_k$ for $k = 1, \dots, K - 1$.

In case of replicated media sessions, the constraint set of Equation (13) is reduced to a single constraint in the form of

$$\sum_{k=1}^K g_k \leq \text{BWF} \quad (40)$$

taking into consideration the fact that individual group rates do not include the aggregated sum of the previous layers. First, we convert the rate allocation optimization problem of (15) with inequality constraints to an optimization problem without constraints. We do so by defining the Lagrangian function of Equation (15) as

$$\begin{aligned}
LG_{IRF} &= IRF A_{Total} \\
&\quad + \sum_{k=1}^K \mu_k (g_k - \text{BWL}_k) + \lambda (\sum_{k=1}^K g_k - \text{BWF}) \\
&= \sum_{k=1}^K IRF A_k \\
&\quad + \sum_{k=1}^K \mu_k (g_k - \text{BWL}_k) + \lambda (\sum_{k=1}^K g_k - \text{BWF}) \\
&= \sum_{k=1}^K \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \\
&\quad + \sum_{k=1}^K \mu_k (g_k - \text{BWL}_k) + \lambda (\sum_{k=1}^K g_k - \text{BWF})
\end{aligned} \quad (41)$$

where the parameters λ and μ_k for $k = 1, \dots, K$ are the Lagrange multipliers in the Lagrangian Equation (41). The unconstrained maximization problem is defined as

$$\begin{aligned} \max_{g_1, \dots, g_K} LG_{IRF} &= \max_{g_1, \dots, g_K} \left(\sum_{k=1}^K IRF A_k \right. \\ &\quad \left. + \sum_{k=1}^K \mu_k (g_k - \text{BWL}_k) + \lambda \left(\sum_{k=1}^K g_k - \text{BWF} \right) \right) \\ &= \max_{g_1, \dots, g_K} \left(\sum_{k=1}^K \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \right. \\ &\quad \left. + \sum_{k=1}^K \mu_k (g_k - \text{BWL}_k) + \lambda \left(\sum_{k=1}^K g_k - \text{BWF} \right) \right) \end{aligned} \quad (42)$$

Conditions of Optimality: Constraint Qualifications

We now investigate the existence of necessary and sufficient optimality conditions also known as constraint qualifications. For our unconstrained maximization problem

$$\max_{g_1, \dots, g_K} LG_{IRF} \quad (43)$$

where $g = \{g_1, \dots, g_K\}$ the constraint qualifications are expressed in terms of Lagrange multiplier theory revolving around conditions under which Lagrange multiplier vectors satisfying the following conditions are guaranteed to exist for a local maximum $g^* = \{g_1, \dots, g_K\}$.

$$\nabla LG_{IRF}(g^*) = \quad (44)$$

$$\begin{aligned} \nabla_{(g^*)} &\sum_{k=1}^K \sum_{i \in G_k} \frac{(2+a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} \\ &+ \nabla_{(g^*)} \sum_{k=1}^K \mu_k (g_k - \text{BWL}_k) \\ &+ \nabla_{(g^*)} \sum_{k=1}^K \lambda (g_k - \text{BWF}) = 0 \end{aligned} \quad (45)$$

$$\begin{aligned} \mu_k &\leq 0 & \forall k = 1, \dots, K \\ \mu_k &= 0 & \forall k \notin A(g^*) \end{aligned} \quad (46)$$

for $A(g^*) = \{k \mid g_k^* - \text{BWL}_k = 0\}$ and

$$\begin{aligned} \lambda &\leq 0 \\ \lambda &= 0 & \forall k \notin B(g^*) \end{aligned} \quad (47)$$

for $B(g^*) = \{k \mid g_k^* - \text{BWF} = 0\}$. The constraint qualifications guarantee the existence of Lagrange multipliers for a given local maximum $g^* = \{g_1^*, \dots, g_K^*\}$ if the inequality constraint function of (40) and the inequality constraint functions of (12) are concave³.

Considering the fact that the Lagrangian function LG_{IRF} satisfies all of the conditions mentioned above, finding the optimal solution is equivalent to finding the solutions of the equation set (44) in the appropriate group ranges. The solution to the nonlinear system of $2K + 1$ equations and $2K + 1$ unknowns provides the optimal rates g_k for $k = 1, \dots, K$ as well as the optimal Lagrange multipliers. The system of $2K + 1$ equations consists of the K gradient equations shown below plus $(K + 1)$ constraint equations (12) and (40).

$$\begin{aligned} \frac{\partial LG_{IRF}}{\partial g_k} \Big|_{g_k^*} &= \\ \left(\sum_{i \in G_k} r_i (2+a) \frac{r_i^2 - g_k^2}{(r_i^2 + ar_i g_k + g_k^2)^2} + \mu_k + \lambda \right) \Big|_{g_k^*} &= 0 \end{aligned} \quad (48)$$

where $k = 1, \dots, K$. The solution to the nonlinear system of $2K + 1$ equations and $2K + 1$ unknowns can be obtained by finding the positive real root of (48) such that $r_{k_{min}} \leq g_k^* \leq r_{k_{max}}$ where $r_{k_{min}}$ and $r_{k_{max}}$ indicate the

³The function $f : \mathcal{C} \mapsto \mathcal{R}^n$ defined over the convex set $\mathcal{C} \subseteq \mathcal{R}^n$ is called concave if $\forall x_1, x_2 \in \mathcal{C}$ and $0 \leq \alpha \leq 1$ the inequality $f(\alpha x_1 + (1 - \alpha)x_2) \geq \alpha f(x_1) + (1 - \alpha)f(x_2)$ holds.

minimum and maximum isolated rates of the receivers belonging to group G_k . One can find the region in which the border line second condition of (48) holds. The time complexity of solving for the root of this equation over all of the existing groups is $\mathcal{O}(KN \log N)$ and determines the overall complexity of the solution considering the fact that the rest of calculations are in the time complexity order of $\mathcal{O}(N)$. Note that the system of $2K + 1$ equations and $2K + 1$ unknowns in this case is more complicated than the case of layered media described in Section III, because of the coupling of the constraint (40) with individual gradient equations of (44).

APPENDIX III

APPLICATION OF CLASSIFICATION METHODS IN RECEIVER PARTITIONING

In general, classification is a mapping from a high-dimensional event space onto a low-dimensional feature space. Classification techniques are often used in many real-world examples. The general idea behind probabilistic classification is to allow the possibility of having a different set of numbers in each class with close inter-class statistical properties such that the representation of the numbers as one class is meaningful from the stand point of a target application.

Jafarkhani in [8] and Joshi et al. in [14] proposed the use of gain-based classification algorithms in image coding. They relied on a metric related to the energy of image blocks to assign blocks with close gain values to the same class. One of the classification methods proposed in their work was Equal Mean-Normalized Standard Deviation (EMNSD) approach in which a sorted group of gain values were split into a given number of classes such that the mean-normalized standard deviation of the resulting classes were as close to each other as possible.

Statistically, standard deviation is a measure of dispersion of samples for a distribution. The smaller is the standard deviation of a source, the denser will be the samples about the mean. When one of the classes has a higher dispersion than others, the blocks in that particular class do not have the same level of activity. The coefficient of variation defined as the ratio of standard deviation over the mean is a good measure of dispersion considering the difficulty to compare dispersions in sets with different means.

Jiang et al. in [11] proposed a heuristic approach to partition the receivers of a layered media session. Although they did not introduce a formal algorithm, their heuristic rules were closely related to probabilistic classification methods such as EMNSD. In what follows we propose applying EMNSD technique as a formal probabilistic classification algorithm to the problem of receiver partitioning in a layered/replicated media session assuming the number of groups and the group rates are given. It is important to note that because of the discrete nature of the problem there is no guarantee that the solution proposed by EMNSD algorithm either exists or is unique.

For simplicity, we first consider the case with two classes. When there are N receivers with their isolated rates r_i sorted in an increasing order, EMNSD algorithm looks for an integer N' such that receivers 1 to N' belong to the first class and the remaining receivers belong to the second class. The mean μ and standard deviation σ of each class is defined by

$$\begin{aligned}\mu_1 &= \frac{1}{N'} \sum_{i=1}^{N'} r_i, \\ \mu_2 &= \frac{1}{N-N'} \sum_{i=N'+1}^N r_i, \\ \sigma_1^2 &= \frac{1}{N'} \sum_{i=1}^{N'} (r_i - \mu_1)^2, \\ \sigma_2^2 &= \frac{1}{N-N'} \sum_{i=N'+1}^N (r_i - \mu_2)^2\end{aligned}\tag{49}$$

Here, N' is chosen such that

$$q_1 = \frac{\sigma_1}{\mu_1} = \frac{\sigma_2}{\mu_2} = q_2\tag{50}$$

An iterative algorithm to find N' satisfying (50) is provided below. If there is no integer N' capable of solving (50), the algorithm finds the N' minimizing $|q_1 - q_2|$.

EMNSD Algorithm:

- Step 1: Choose an initial value for N' , e.g., $N' = N/2$, and set the iteration number $i = 0$. Also choose i_{max} as an upper limit on the number of iterations.
- Step 2: Compute q_1 and q_2 using (50) and set $i = i + 1$.

- Step 3: If $\frac{|q_1 - q_2|}{q_1} < \delta$ or $i > i_{max}$ STOP. Otherwise, if $q_1 < q_2$ set $N' = N' + \Delta N'$; else set $N' = N' - \Delta N'$, and go to Step 2.

For fast convergence, a large $\Delta N'$ can be chosen at the beginning of the algorithm and as the iteration number increases $\Delta N'$ should be gradually decreased to one. The same algorithm can be generalized for a larger number of classes. For the case of K classes, K ratios $q_i = \frac{\sigma_i}{\mu_i}$ for $i = 1, \dots, K$, and $(K - 1)$ thresholds are needed. The algorithm is stopped when $\frac{\max_i q_i - \min_i q_i}{\min_i q_i} < \delta$ or when the number of iteration exceeds its maximum. At each step of the algorithm, the thresholds corresponding to the class with the maximum and minimum q_i are alternatively adjusted so as to make q_i 's as close to one another as possible.

REFERENCES

- [1] E. Amir, S. McCanne, R. Katz, "Receiver-Driven Bandwidth Adaptation for Light-Weight Sessions", ACM Multimedia, Nov. 1997.
- [2] M. H. Ammar, "Probabilistic Multicast: Generalizing the Multicast Paradigm to Improve Scalability," IEEE INFOCOM, (Toronto, CANADA), June 1994.
- [3] R. Bellman, "Adaptive Control Processes: A Guided Tour", Princeton University Press, 1961.
- [4] J. Bolot, T. Turetli, "A Rate Control Mechanism for Packet Video in the Internet", Proc. of IEEE INFOCOM'94, June 1994.
- [5] S. Cheung, M. H. Ammar, X. Li, "On the Use of Destination Set Grouping to Improve Fairness in Multicast Video Distribution", Proc. IEEE INFOCOM '96, San Francisco, CA., Mar. 1996, pp. 553-60.
- [6] S.E. Deering, D.R. Cheriton, "Multicast Routing in Datagram Internetworks and Extended LANs", ACM Trans. on Computer Systems, 8:85-110, May 1990.
- [7] D. DeLucia, K. Obraczka "Multicast Feedback Suppression Using Representatives", Proc. of IEEE INFOCOM 97, Apr. 1997.
- [8] H. Jafarkhani, "Adaptive Wavelet Based Image Coding", MS Thesis, EE Dept. of Univ. of Maryland, 1994.
- [9] H. Jafarkhani, V. Tarokh, "Design of Successively Refinable Trellis Coded Quantizers," IEEE Transactions on Information Theory, vol. 45, pp. 1490-1497, July 1999.
- [10] J. M. Jaffe, "Bottleneck Flow Control", IEEE Trans. on Communications, vol. 29, pp. 954-962, July 1981.
- [11] T. Jiang, M. H. Ammar, E. W. Zegura. "On the use of Destination Set Grouping to Improve Inter-Receiver Fairness for Multicast ABR Sessions", Proc. of IEEE INFOCOM 2000, Mar. 2000.
- [12] T. Jiang, M. H. Ammar, E. W. Zegura, "Inter-Receiver Fairness: A Novel Performance Measure for Multicast ABR Sessions", Proc. of ACM SIGMETRICS '98, June 1998.
- [13] T. Jiang, E. W. Zegura, M. Ammar, "Inter-receiver Fair Multicast Communication over the Internet", Proc. of the 9th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), pp. 103-114, June 1999.
- [14] R. L. Joshi, H. Jafarkhani, J. H. Kasner, T. R. Fischer, N. Farvardin, M. W. Marcellin, R. H. Bamberger, "Comparison of Different Methods of Classification in Subband Coding of Images", IEEE Trans. ON Image Processing, vol. 6, NO. 11, Nov. 1997.
- [15] X. Li, M. Ammar, S. Paul, "Video Multicast Over the Internet", IEEE Network Magazine, pp. 46-60, Apr. 1999.
- [16] X. Li, S. Paul, M. H. Ammar, "Layered Video Multicast with Retransmissions (LVMR): Evaluation of Hierarchical Rate Control", Proc. of INFOCOM 98.
- [17] X. Li, S. Paul, M. Ammar, "Multi-Session Rate Control for Layered Video Multicast", Proc. SPIE Multimedia Computing and Networking, San Jose, CA., vol. 3654, pp. 175-89, Jan. 1999.
- [18] S. McCanne, V. Jacobson, M. Vetterli, "Receiver Driven Layered Multicast", Proc. of ACM SIGCOMM '96, Sept. 1996.
- [19] P. Moghe, P., and I. Rubin, "Reserving for Future Clients in Multipoint Application-Why and How?", IEEE JSAC vol. 15, no.3, April 1997.
- [20] W. Ren, K. Siu, H. Suzuki, "On the Performance of Congestion Control Algorithm for Multicast ABR Service in ATM", Proc. of IEEE ATM '96 Workshop, Aug. 1996.
- [21] D. Rubenstein, J. Kurose, D. Towsley, "The impact of Multicast Layering on Network Fairness", SIGCOMM '99, Sept 1999.
- [22] N. Shacham, "Multipoint Communication by Hierarchically Encoded Data", Proc. of IEEE INFOCOM '92, May 1992.
- [23] T. Turetli, S. F. Parisi, J. Bolot, "Experiments with a Layered Transmission Scheme over the Internet", Proc. of IEEE INFOCOM '98, Mar. 1998.
- [24] H. Tzeng, K. Siu, "On Max-Min Fairness Congestion Control for Multicast ABR Service in ATM", JSAC, 15(3), Apr. 1997.
- [25] L. Vicisano, L. Rizzo, J. Crowcroft, "TCP-Like Congestion Control for Layered Multicast Data Transfer", Proc. of IEEE INFOCOM '99, volume 3, Mar. 1999.
- [26] H. A. Wang, M. Schwartz, "Achieving Bounded Fairness for Multicast Traffic and TCP Traffic in the Internet", Proc. of ACM SIGCOMM '98, Sept. 1998.
- [27] Y. R. Yang, M. S. Kim, S. S. Lam, "Optimal Partitioning of Multicast Receivers", Proc. of the 8th International Conference on Network Protocols, Nov. 2000.