Abstract—The objective of layering techniques of distributing multimedia traffic over multicast IP networks is to effectively cope with the challenges in continuous media applications. The challenges include heterogeneity, fairness, real-time constraints, and quality of service. We study the problem of rate allocation and receiver partitioning in layered and replicated media systems. We formulate an optimization problem aimed at maximizing a close approximation of the so-called max-min fairness metric subject to loss and bandwidth constraints. Our optimal Layered Media Multicast Control (LMMC) solution to the problem analytically determines the layer rates and the corresponding partitioning of the receivers. Our simulation results show the effectiveness of our proposed solution in realistic scenarios.

Index Terms—Fairness extrapolation, heterogeneity, layered media, multicast IP networks, optimality, rate allocation, replicated media, receiver partitioning.

I. INTRODUCTION AND RELATED WORK

TRANSMITTING real-time 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 [17] and the references cited therein. In a typical multicasting transmission scenario, a source generates real-time media traffic following a periodic pattern. The periodic pattern of real-time 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. A review of the literature reveals three different adaptive bit-rate media multicasting schemes for the transmission of digital media. The schemes are described below.

1) Single stream adaptive approach was 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 are required to address this issue. While it is straightforward to implement, the single stream adaptive approach is unable to properly address the problem of receiver heterogeneity.

2) 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 qualities and bit rates. Each stream is obtained by encoding the video with different compression parameters and is sent to a different multicast group. Each individual receiver is able to join and change its group according to its capacity. 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.

3) 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. [20] in the context of RLM protocol, Amir et al. [1] in the context of SCUBA protocol, and Li et al. [18] 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 was discussed by Jafarkhani et al. in [12] 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. Layered media approach provides an elegant and efficient way to deal with the heterogeneity issue at the expense of protocol complexity.

As a real-world example of the subject material of this study, one can consider the transmission of a digital video stream to the members of a pay-per-view entertainment club. Club members are typically connected through dial-up, ISDN, Cable/DSL, 10baseT Ethernet, 100baseT Ethernet, and gigabit
Ethernet lines and consequently belong to different bandwidth groups. The differences among processing power, topology, and protocol implementation at the receiving ends typically cause deviations from the nominal bandwidth capabilities of the members in each bandwidth group. Assuming either the replicated or the layered media system approach is used for the transmission of the video stream, an important practical question is what is the optimal number of groups for transmitting the stream? The answer is typically specified by considering the trade off between receiving ends’ bandwidth heterogeneity and the incurring overhead in source encoding, receiver decoding, and multicasting addressing. Without considering coding and multicasting overhead, the number of groups is directly mapped to the number of bandwidth categories. However, it is often required to select a smaller number of groups than the number of bandwidth categories in order to reduce the overhead. In such cases, a number of high bandwidth groups may be combined into one group in order to address the tradeoff between heterogeneity and overhead.

The material proposed in this paper is closely related to the following articles. In [14] and the follow-on work of [15], Jiang et al. explore the issue of improving inter-receiver fairness in multicast ATM sessions with an Available Bit Rate (ABR). In order to determine the optimal partitioning and allocation of the group rates, the authors formulate a max-min fairness optimization problem subject to the maximum loss tolerance of a set of receivers. The authors apply their formulation to replicated media systems in the context of DSG protocol. They also provide a set of heuristic rules for solving the formulated problem. Their three proposed heuristic rules are consistent with our practical discussion of the previous paragraph and are intended for ensuring (1) dissimilar receivers are not grouped together, i.e., a set of receivers are increasingly ordered and partitioned in terms of their isolated rates, (2) receivers of similar performance levels are grouped together, i.e., the normalized standard deviations of the isolated rates of the receivers in each partition are relatively small, and (3) a group of receivers can only be split into two groups if the difference between the resulting group rates is larger than the smaller group rate. In [16], the same group of authors apply their work to Internet-driven applications with the considerations of TCP-friendliness. In [29], Yang et al. provide a dynamic programming algorithm to simultaneously solve the problems of optimal partitioning and rate allocation for layered media systems.

The main objective of the current research work is to provide an analytical framework for the partitioning strategy and rate allocation of both layered and replicated 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 [26], [19], [28], [22], and [27]. In addition, our work of [30] proposes a framework of flow control for layered and replicated media streams. The main contributions of this paper are in three areas. First, the paper introduces an analytical approach in which a noncontinuously differentiable max-min fairness function is extrapolated by a class of mathematically well-behaved continuously differentiable functions. The extrapolated functions satisfy the conditions required for applicability of traditional optimization 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.

Specifically, we formulate a two-phase optimization problem of partitioning and rate allocation after extrapolating the so-called max-min fairness metric with a mathematically well-behaved function. In the first phase, we analytically solve the optimal rate allocation problem for individual layers of the media session assuming the number of layers is given. 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 the extrapolated inter-receiver fairness metric. In the second phase, we provide an 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 optimization strategy and is guaranteed to reach an ɛ-neighborhood of a local optimal 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 solutions 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 focus on the general rate allocation and partitioning problem of the layered and replicated media sessions. The problem aims at transmitting a stream of digital media to a set of receivers with different bandwidth capabilities such that
each receiver can create a reconstruction of the stream with a quality proportional to its own bandwidth capability. We formulate the problem in a manner similar to that of [16], [15], and [29] with an extra constraint on the overall available bandwidth to the session. The previous problems 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. Recall that for a media session with $N$ receivers and $K$ groups, a set $P = \{G_1, \ldots, G_K\}$ is called a partitioning of the receiver set $R = \{1, \ldots, N\}$ 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 aggregate 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 $g_1, g_2, \ldots, g_K$ such that $g_1 \leq g_2 \leq \cdots \leq g_K$, the layer rates of a layered media session are calculated in the form of

$$g_1 + g_2 - g_1 + g_3 - g_2 + \cdots + g_K - g_{K-1}.$$  

(1)

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

Interpretation of our formulation in the case of replicated media streams is also straightforward. For an ordered partitioning of the receivers into $K$ groups $G_1, G_2, \ldots, G_K$ with ordered group rates of $g_1, g_2, \ldots, g_K$ such that $g_1 \leq g_2 \leq \cdots \leq g_K$, 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 some 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 multirate max-min fair rate of $r_i$ as described in both [14] and [23]. 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 a nondecreasing order, i.e., $r_1 \leq r_2 \leq \cdots \leq r_N$. 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 groups $G_1, \ldots, G_K$:

$$g_k \leq \frac{r_i}{1 - L_i} \quad \forall i \in G_k \quad k = 1, \ldots, K.$$  

(2)

In [14], a class of fairness utilities $\mathcal{F}(r_i, g_k)$ are defined for receiver $i$ of group $G_k$ by means of satisfying the following conditions:

- $\mathcal{F}(r_i, g_k) \in [0, 1]$.
- $\mathcal{F}(r_i, r_i) = 1$.
- $\mathcal{F}(r_i, g_k) < 1$ if $r_i \neq g_k$.
- $\mathcal{F}(r_i, g_k)$ is nondecreasing in the range $[0, r_i]$.
- $\mathcal{F}(r_i, g_k)$ is nonincreasing in the range $(r_i, \infty)$.

In this paper, we work with the most widely accepted example of such utility functions, the so-called max-min fairness utility function defined as

$$F(r_i, g_k) = \min \left( \frac{g_k}{r_i}, 1 \right) = \begin{cases} \frac{g_k}{r_i} & : g_k \leq r_i \\ \frac{r_i}{g_k} & : g_k \geq r_i \end{cases}.$$  

(3)

The group utility for the group $G_k$ with a group rate $g_k$ is defined as

$$\text{IRF}_k = \sum_{i \in G_k} F(r_i, g_k) = \sum_{i \in G_k} \frac{\min(r_i, g_k)}{\max(r_i, g_k)}.$$  

(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 $\alpha_i$ with the following characteristics:

$$\sum_{i=1}^{N} \alpha_i = 1$$

$$0 \leq \alpha_i \leq 1, \quad \text{for } i = 1, \ldots, N$$

$$\alpha_i = 0, \quad \text{for } i \notin G_k.$$  

(5)

The choice of parameters $\alpha_i$ is a design decision allowing for unequal contribution of the receivers to a group utility according to their importance. The parameters may be statically assigned or dynamically vary over time. Generally speaking, the choice of parameters $\alpha_i$ does not have any significant impact in our study. The session utility of the partitioning $P = \{G_1, \ldots, G_K\}$ is defined as

$$\text{IRF}_{\text{Total}} = \sum_{k=1}^{K} \sum_{i \in G_k} F(r_i, g_k)$$

$$= \sum_{k=1}^{K} \frac{\min(r_i, g_k)}{\max(r_i, g_k)}.$$  

(6)

The objective of both heuristics given in [15] and the dynamic programming algorithm given in [29] 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

$$\max_{g_1, \ldots, g_K} \text{IRF}_{\text{Total}} = \max_{g_1, \ldots, g_K} \sum_{k=1}^{K} \text{IRF}_k$$

$$= \max_{g_1, \ldots, g_K} \sum_{k=1}^{K} \sum_{i \in G_k} \frac{\min(r_i, g_k)}{\max(r_i, g_k)}$$  

(7)

subject to: $g_k \leq \frac{r_i}{1 - L_i}$ $\quad$ $i \in G_k \quad k = 1, \ldots, K$  

(8)

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

In Theorem 1 of [29] 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 nondecreasing in the interval $[0, g]$ and nonincreasing in the interval $[g, \infty)$ for a fixed $r$.
\( F(r, g) \) is nondecreasing in the interval \([0, r]\) and nonincreasing in the interval \([r, \infty)\) for a fixed \( g \).

We now introduce an extrapolation technique to replace the noncontinuously 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. Such an extrapolation technique provides us with the opportunity to introduce a more effective solution to the problem of rate allocation and partitioning in terms of time and space complexity. By mathematically well-behaved, we mean that our so-called extrapolated function \( E(r_i, g_k) \) is continuously differentiable and has no poles over the real numbers axis.

We select a rational function \( E(r_i, g_k) \) in the form of

\[
E(r_i, g_k) = \frac{(2 + \alpha)r_i g_k}{g_k^2 + a r_i g_k + r_i^2}
\]

and note that not only \( E(r_i, g_k) \) is well behaved for parameter \( \alpha \) satisfying the boundary condition \(-2 < \alpha < 2\), but it satisfies the boundary and maximum conditions of function \( F(r_i, g_k) \). The matter is best explained by a graphical illustration. Fig. 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 find the optimum value of the parameter \( q \) within the interval of interest \([0, (r_i^2)/(1 - L_i)]\) considering the constraint function of (8) and as shown below:

\[
\min_{\alpha} \left[ \alpha \right] \left[ \int_0^{r_i} \left( \frac{(2 + \alpha)r_i g_k}{g_k^2 + a r_i g_k + r_i^2} - \frac{g_k}{r_i} \right)^2 dg_k + \int_{r_i}^{r_i^2/(1 - L_i)} \left( \frac{(2 + \alpha)r_i g_k}{g_k^2 + a r_i g_k + r_i^2} - \frac{r_i}{g_k} \right)^2 dg_k \right].
\]

Solving (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 \( \alpha \). In our calculations, we perform a table look up operation to extract the optimal value of parameter \( \alpha \). 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

\[
\max_{g_1, \ldots, g_K} \text{IRF}_{\text{Total}} \equiv \max_{g_1, \ldots, g_K} \sum_{k=1}^{K} \text{IRF}_k
\]

\[
= \max_{g_1, \ldots, g_K} \sum_{k=1}^{K} \sum_{i \in G_k} \frac{(2 + \alpha)r_i g_k}{g_k^2 + a r_i g_k + r_i^2}
\]

subject to:

\[
g_k \leq \text{BWL}_k \quad k = 1, \ldots, K
\]

\[
g_k \leq \text{BWF}_k \quad k = 1, \ldots, K
\]

(11)

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

\[
\text{IRF}_k \equiv \sum_{i \in G_k} E(r_i, g_k) = \sum_{i \in G_k} \frac{(2 + \alpha)r_i g_k}{g_k^2 + a r_i g_k + r_i^2}.
\]

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

\[
\max_{g_1, \ldots, g_K} \text{IRF}_{\text{Total}} = \max_{g_1, \ldots, g_K} \sum_{k=1}^{K} \text{IRF}_k
\]

\[
= \max_{g_1, \ldots, g_K} \sum_{k=1}^{K} \sum_{i \in G_k} \frac{(2 + \alpha)r_i g_k}{g_k^2 + a r_i g_k + r_i^2}
\]

subject to:

\[
g_k \leq \text{BWA}_k \quad k = 1, \ldots, K.
\]

(16)

We note the difference between the loss tolerance constraints \( \text{BWL}_k \) and the group bandwidth upper bounds \( \text{BWF}_k \). While the former reflects the receiver’s 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 (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. The general problem of (15) and Constraint (16) can be converted to an optimization problem without constraints by defining a Lagrangian function in the form of

\[
\text{L}_{\text{IRF}} = \text{IRF}_{\text{Total}} + \sum_{k=1}^{K} \mu_k (g_k - \text{BWA}_k)
\]

(17)
where the parameters $\mu_k$ for $k = 1, \ldots, K$ are the Lagrange multipliers in the Lagrangian equation (17). The solution to the unconstrained problem can then be obtained by solving $\nabla I_{G, k}$ with $g_k = 0$. However considering the specific form of the function $I_{F, k}$ and the constraint set of (16), the most straightforward way of solving for the optimal solution is to decompose the system of $2K$ equations and $2K$ unknowns obtained from $\nabla I_{F, k}(g_k) = 0$ 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 $\nabla I_{F, k}(g_k) = 0$ and then investigating 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.

$$\max_{g_k} I_{F, k} = \max_{g_k} \sum_{i \in G_k} \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2}$$

subject to: $g_k \leq BWA_k$ (18)

where $k = 1, \ldots, K$. The set of optimization problems of (18) can be solved by finding the roots of the following equations:

$$\frac{\partial I_{F, k}}{\partial g_k} = \sum_{i \in G_k} \frac{(2 + a)r_i (r_i^2 - g_k^2)}{(g_k^2 + ar_i g_k + r_i^2)^2} = 0$$

(19)

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

$$\frac{\partial^2 I_{F, k}}{\partial g_k^2} = \sum_{i \in G_k} \frac{2(2 + a)r_i (g_k^3 - 3r_i^2 g_k - ar_i^2)}{(g_k^2 + ar_i g_k + r_i^2)^3} \leq 0.$$  

(20)

Prior to proceeding with the solution to individual optimization problems, we review the mathematical characteristics of the function $I_{F, k}$. We first note that the function $I_{F, k}$ is nondecreasing in the interval $[0, r_{\text{min}}]$ and nonincreasing in the interval $[r_{\text{max}}, \infty]$ where $r_{\text{min}}$ indicates the minimum isolated rate and $r_{\text{max}}$ indicates the maximum isolated rate of the receivers belonging to group $G_k$. This is true because the function $I_{F, k}$ consists of a sum of a number of the receiver utility functions $E(r_i, g_k)$ which are all nondecreasing in the interval $[0, r_{\text{min}}]$ and nonincreasing in the interval $[r_{\text{max}}, \infty]$. Consequently, (19) has no roots in the intervals $[0, r_{\text{min}}]$ and $[r_{\text{max}}, \infty]$. We also remind that any acceptable optimal point has to satisfy Constraint (16). Combining the above conditions, we can argue that for $r_{\text{min}} \geq BWA_k$ the optimal solution equals to $BWA_k$ and for $r_{\text{min}} < BWA_k$ any acceptable maximum point falls into the interval

$$[r_{\text{min}}, BWA_k].$$  

(21)

For $r_{\text{min}} < BWA_k$, Constraint (16) has no impact on the optimal solution.

Generally speaking, the function $I_{F, k}$ 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 $I_{F, k}$ is hence equivalent to applying a root finding algorithm on (19) and extracting the global maximum from the set of optimal points satisfying Inequality (20) and Constraint (16).

In our extensive set of simulations, we have consistently observed that the function $I_{F, k}$ includes a single global maximum point if the individual receiver utilities are distributed in such a way that every two consecutive isolated rates $r_i$ and $r_{i+1}$ satisfy the relationship $r_{i+1} \leq 2r_i$. The latter is a practical assumption for a set of receivers with similar bandwidth capabilities. Fig. 2 shows a typical $I_{F, k}$ function. Finding the global maximum of the function $I_{F, k}$ in such a case is hence equivalent to applying a single root finding algorithm such as bisection or Newton algorithms to (19). These algorithms can identify the single root of (19) with a time complexity of $O(N \log N)$. We argue that if a media session can choose the number of groups such that our heuristic rule of $r_{i+1} \leq 2r_i$ is satisfied, all of the corresponding $I_{F, k}$ 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 $I_{F, k}$ 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, 10 Mb/s LAN, and 100 Mb/s 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 map 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 $I_{F, k}$ function with multiple optimum points. Additionally, even in the case of observing multiple maximum points for the function $I_{F, k}$, our numerical results have only shown one maximum for any subset of receivers with isolated rates satisfying $r_{i+1} \leq 2r_i$. Fig. 3 shows an example of such an $I_{F, k}$ function. In order to prevent a significant quality degradation at a receiver, we assume that the maximum acceptable loss tolerance of a receiver does not exceed 50%. This implicitly means that $BWA_k$ defined as $\min(BWL_k, BWF_k)$ with $BWL_k$ defined as $\min_{i \in G_k} (r_i / (1 - L_i))$ will typically not exceed $2r_{\text{min}}$ where $r_{\text{min}}$ indicates the minimum isolated rate of the receivers be-
longing 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
$\text{IRFA}_k$. Applying the interval of (21), Newton, bisection or a
similar numerical technique can be employed to find the first
positive real maximum of $\text{IRFA}_k$ function.

As an important special case and by substituting $\text{BW}_k$ with
$\text{BW}_L_k$, the general formulation of our problem reduces to the
no flow constraint problem formulated in [29] and [15]. 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 [29].
In practice, the time complexity of solving for the optimum
point of equation set (19) over all of the existing groups is
$O(KN \log N)$. The search for the root of (19) determines the
overall time complexity of the solution considering the fact that
the rest of calculations are in the complexity order of $O(N)$. The
time complexity of the algorithm is by far better than $O(N^2)$
the complexity of the dynamic programming algorithm
offered by [29]. This is aside from the fact that a dynamic pro-
gramming approach in general does not provide an analytical
solution to an optimization problem and the algorithm of [29]
needs minor modifications to be able to solve the formulation of
the general problem of (15) considering the impact of enforcing
a flow control algorithm.

Before we proceed to phase 2 of our solution, it is also rele-
vant to investigate the impacts of facing some of the source and
receiver limitation scenarios when solving LMMC optimization
problem. First, we consider a source limitation scenario that ap-
ppears 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 gran-
ularity. In practice, layered encoding techniques may limit the
source to some pre-determined quantized discrete group rates.\footnote{Examples of standard layered encoding techniques with pre-determined quantized discrete rates include MPEG-2 [8], H.263 [10], and new-generation MPEG-4 [9]. AVC/H.264/ISO 14.496-10 [11]. We note that the family of MPEG standards originally supported successively refinable video in the range of several Mb/s and eventually covered lower rates. The H.263 standard and the follow-on standards were originally designed to support successively refinable video at a wider range of rates starting at tens of Kbs.}

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 dis-
crete rate. Although the approximated solution is sub-optimal in
this case, it reduces the complexity of the problem to a great ex-
tent and yields acceptable results so long as the discrete achiev-
able rates of the underlying encoder are not very far from each
other. The latter is a reasonable assumption for many of the cur-
rently available encoders. Considering distribution of the dis-
crete group rates, we choose the second approach as the prac-
tical 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 tol-
erance scenario with $L_i = 0$ for $i = 1, \ldots, N$ in our opti-
mization algorithm is to change the definition of $\text{BW}_L_k$ from
$\text{BW}_L_k \equiv \min_{\text{g} \in G_k}(r_i/(1 - L_i))$ to $\text{BW}_L_k \equiv \min_{\text{g} \in G_k} r_i$ for
$k = 1, \ldots, K$. Since the previous constraint qualifications hold
for $0 \leq L_i < 1$ with $i = 1, \ldots, N$, we do not foresee any
changes on the method of obtaining our optimal solution. How-
ever, we make note that this scenario greatly simplifies the re-
Sults considering the fact that the function of (15) would have
no zero slope point satisfying Constraint (16) for $N_k > 1$. In-
putively, we anticipate that the optimal rate of each group is al-
ways less than or equal to the lowest isolated rate of the group.

IV. PHASE 2: LMMC NEAR-OPTIMAL PARTITIONING
STRATEGY

In Section I, we briefly described the heuristic partitioning
rules of [15]. We note that the heuristic rules are well cate-
gorized under probabilistic classification and clustering methods
for nonconvex optimization problems. In [31], we provide a
formal classification method that is closely related to the par-
titioning 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 deduc-
tion techniques on the properties of mathematical concepts and
closely related computational algorithms concepts rather than
being useful for approximate or exact solutions to the optimiza-
tion problems. Nevertheless, these techniques come handy in
the case of solving optimization problems and in the absence of
a formal solution.

In addition, the dynamic programming algorithm of [29] pro-
vides an optimal receiver partitioning strategy for a media ses-
sion while computing the optimal layer rates. The main disad-
veniences of utilizing a dynamic programming approach to solve
an optimization problem are (1) the lack of providing an ana-
lytical answer, and (2) a relatively high degree of complexity.
However, we make note of the fact that dynamic programming
is one of the best tools and in many cases the only available tool

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

}$
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 $O(N)$ 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 (15) and for a set of given group rates $\{g_1, \ldots, g_K\}$, it is imperative that a receiver with isolated rate $r_i$ is assigned to the group with rate $g_k$ 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 with rate $g_k$ such that

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

We now translate the latter observation to a simple group assignment mechanism. Let us first consider the fairness function of (9) with parameter $g_k$ and variable $r_i$. We note that in Sections II and III, the function of (9) with parameter $r_i$ and variable $g_k$ was considered instead. Given the group rates $\{g_1, \ldots, g_K\}$, we first plot the family of functions $E(r_i, g_k)$ versus $r_i$ for different parameter values of $g_k$ where $k = 1, \ldots, K$. Fig. 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 set of equations for variables $r_i$ and parameters $g_k$ and $g_{k+1}$ where $k = 1, \ldots, K$:

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

Solving (23) yields

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

Although in the general form of (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

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

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

**Theorem 4.1:** The value of the receiver utility as defined in (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 $\{g_1, \ldots, g_K\}$ if $\sqrt{g_k - g_k g_{k+1}} < r_i \leq \sqrt{g_k g_{k+1}}$. The receiver utility is maximized for the choice of the group rate $g_1$ if $r_i \leq \sqrt{g_1}$ and for the choice of the group rate $g_K$ if $r_i > \sqrt{g_K - g_{k-1} g_K}$.

**Proof:** As graphically observed in Fig. 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 the case of three groups. The graphical proof remains the same by expanding partitioning thresholds from $\sqrt{g_1 g_2}$ to $\sqrt{g_K - g_{k-1} g_K}$ 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 the case of solving the constrained problem of (16). Considering the definitions of (16) and (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:

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

where $\mu$ and $\sigma$ 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:

$$\frac{r_i}{1 - \frac{1}{L_k}} \geq g_k^* \quad (27)$$

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.

The LMMC near-optimal partitioning algorithm 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 $O(KN)$. As an alternative
and to achieve a 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 op-
timization problem, this version of the algorithm introduces a
time complexity order of $O(KN \log N)$. 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 a media session and assuming the group
rates $\{g_1, \ldots, g_K\}$ are given

\[
\text{for } (k = 2 \text{ to } K) \{
\]

- Calculate the partitioning threshold $\sqrt{g_{k-1}g_k}$.
- Repartition groups $k - 1$ and $k$. For every receiver be-
  longing 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 $k$ ac-
  cording to the new partitioning.

\[
\] for $(k = 2 \text{ to } K) \}

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

\[
\min(r_i, g_k) = \max(r_i, g_{k+1}).
\]

We conclude that the general algorithm of this section can be
used in conjunction with any rate allocation algorithm by proper-
ly identifying partitioning thresholds. In specific, the algo-
rithm 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 solu-
tion 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 solving the optimal rate allocation
problem again with the alternative partitioning strategy. The op-
timal layer rates of the first phase can then be applied to the
near-optimal partitioning strategy of the second phase to par-

tition 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.

Recall that for a media session with $N$ receivers, $K$ groups,
and the group rate set $g = \{g_1, \ldots, g_K\}$, a set

\[
P = \{G_1, \ldots, G_K\}
\]

is called a partitioning of the receiver set $R = \{1, \ldots, N\}$ 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
rate set $g^*$ such that

\[
\text{IRFA}_{\text{Total}}(P^*, g^* \geq \text{IRFA}_{\text{Total}}(P, g)
\]

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

\[
\text{IRFA}_{\text{Total}}(P_{\text{fixed}}, g^{*}_f) \geq \text{IRFA}_{\text{Total}}(P_{\text{fixed}}, g)
\]

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

\[
\text{IRFA}_{\text{Total}}(P^*, g^{*}_s) \geq \text{IRFA}_{\text{Total}}(P, g^{*}_s)
\]

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
  $j_{\text{max}}$. 
- Step 2: Calculate the optimal group rates $g^{*} =
  \{g^{*}_1, \ldots, g^{*}_K\}$ and the resulting session utility $\text{IRFA}_{\text{Total}}$
  by numerically solving the system of (19) while satisfy-
  ing conditions (20) and (16). Save the previously
  calculated $\text{IRFA}_{\text{Total}}$ in variable $q_1$ and the currently
  calculated $\text{IRFA}_{\text{Total}}$ in variable $q_2$.
- Step 3: If $(q_1 - q_2)/q_1 < \delta$ or $j > j_{\text{max}}$ STOP.
- Step 4: for $(k = 2 \text{ to } K) \{
  - Calculate the partitioning threshold $\sqrt{g_{k-1}g_k}$.
  - Repartition groups $k - 1$ and $k$. For every receiver be-
    longing 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 $k$ ac-
    cording to the new partitioning.

\[
\] for $(k = 2 \text{ 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 allo-
cation 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 solu-
tion to the second phase near-optimal partitioning approach
while adjusting the optimal rate of the corresponding group ac-
cording to the new partitioning. We note that the time com-
plexity of our iterative algorithm is $O(INK \log N)$ where $I$
indicates the number of iterations. Comparing the overall complexity of LMMC algorithm with that of the dynamic programming algorithm of [29] $O(N^3)$, LMMC algorithm achieves a 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 (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 (15) has both a lower bound and an upper bound. Next, we observe that the session utility function of (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 nondecreasing sequence with an upper bound equal to the number of fairness functions. We also note that any nondecreasing 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.

Intuitively, LMMC algorithm is employing steepest descent optimization 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 the case of solving the unconstrained problem. The claim is accurate considering the fact that the sequence of session utility functions of (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 the 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 [31] may be deployed as the partitioning strategy. The use of our proposed algorithms yields fast converging results in most cases as shown by our simulation 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 the dynamic programming algorithm of [29]. 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 $\{128$ Kbps, $1$ Mbps, $10$ Mbps, $100$ Mbps, $1$ Gbps$\}$. We note that the choice of modal distributions represents the distribution of bandwidths associated with 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 $1.8$ GHz Pentium $4$ CPU, $512$ MB of physical memory and $1$ GB of virtual memory. Further, we rely on the Gnu Scientific Language (GSL) optimization toolbox to provide a balance between the speed and robustness of program execution.

We recall that the time complexity of the iterative optimized LMMC algorithm is $O(\hat{K}N\log N)$ where $\hat{I}$ indicates the number of iterations and the time complexity of the Dynamic Programming (DP) algorithm of [29] is $O(N^3)$. In addition, the space complexity of the LMMC algorithm in our implementation is $O(N)$ where as the space complexity of DP algorithm proposed in [29] is $O(N^2)$. 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.

Figs. 5–10 compare the sample results of LMMC algorithm with those of DP algorithm of [29]. In each experiment, we have considered the same loss tolerance for all of the receivers of the session. Different figures have been obtained for 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. In order to do a fair comparison, we have used the fairness function of (9) for LMMC and the fairness function of (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 (9) is an approximation of the fairness function of (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.

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 the case of DP algorithm than the case of LMMC algorithm.

Finally, we would like to review the impacts of using criteria set (26) and criteria set (27) in controlling successive groups repartitioning. Generally speaking, we have observed that the proper choice of coefficients C₁ in criteria set (26) and C₂ in criteria set (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 set (27) has yielded better results than using criteria set (26). We have experimentally observed that for a loss tolerance of 10% a value of C₂ = 0.790 best controls the repartitioning process while for a loss tolerance of 20% a value of C₂ = 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 criteria set (26) and criteria set (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 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 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 [29], 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 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 64 Kbps indicate dial-up users, receivers with isolated rates distributed in the range of 1 Mbps indicate Cable/DSL users, and receivers with isolated rates distributed in the range of 100 Mbps 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 interfaces 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 problem of optimal partitioning and rate allocation for layered and replicated media systems over multicast IP networks. We formulated such a problem as a two-phase optimization problem. By means of extrapolating max-min fairness utilities of individual receivers, we proposed our Layered Media Multicast Control (LMMC) solution to the problem. In the first phase, we analytically calculated the optimal rates allocated to the individual layers of a media session. In the second phase, we obtained the best partitioning strategy of the receivers based on the optimal allocated rates of the first phase. 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 one phase to another. Finally, we evaluated the performance of LMMC solution and illustrated its effectiveness and scalability in realistic network topologies through the use of simulations.

We are currently working on integrating the rate allocation and receiver partitioning aspect of LMMC with its end-to-end error control aspect.

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 (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 Fig. 1. We select a rational function of \( g_k \) and \( r_i \) in the form of

\[
E(r_i, g_k) = \frac{N(r_i, g_k)}{D(r_i, g_k)}
\]

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 \( \text{Deg}(N(r_i, g_k)) = \text{Deg}(N(g_k)) \leq M \) 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 \( \text{Deg}(N(g_k)) + 1 = \text{Deg}(D(g_k)) = 2 \) in the form of

\[
E(g_k) = \frac{N(g_k)}{D(g_k)} = \frac{b g_k}{a_2 g_k^2 + a_1 g_k + a_0},
\]

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 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 \( F(r_i, g_k) \).

\[
b > 0, \quad a_0 > 0, \quad a_1 \geq 0, \quad a_2 > 0
\]

\[
E(0) = 0 \Rightarrow a_0 \neq 0
\]

\[
E(\infty) = 0 \Rightarrow \text{Deg}(N(r)) < \text{Deg}(D(r))
\]

\[
E(r_i) = 1 \Rightarrow a_2 r_i^2 + (a_1 - b)r_i + a_0 = 0
\]

\[
E'(r_i) = 0 \Rightarrow a_2 r_i^2 + a_1 r_i + a_0 - r_i(2a_2 r_i + a_1) = -a_2 r_i^2 + a_0 = 0
\]

\[
a_0 = a_2 r_i^2
\]

\[
\Delta D(r_i) < 0 \Rightarrow a_1^2 - 4a_0 a_2 < 0
\]

\[
|a_1| < 2\sqrt{a_2 a_0}
\]
Without loss of generality, we assume that \( a_2 = 1 \) and \( a_1 = ar_i \). Applying the conditions of (34) to the general form of (33) introduces the specific form of

\[
E(r_i, g_k) = \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2}
\]

(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 overall fairness function \( F(r_i, g_k) \) defined in (3). Applying least square estimation technique in the interval of interest \([0, (r_i)/(1 - L_i)]\) while considering the constraint function of (8) yields the optimum value of parameter \( a \) in terms of parameters \( r_i \) and \( L_i \).

\[
\min_a [LSE(a, r_i, L_i)] = \min_a \left[ \int_0^{r_i} \left( \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{g_k}{r_i} \right)^2 \, dgk \, dr_i + \int_{r_i}^{r_i^2} \left( \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{r_i}{g_k} \right)^2 \, dgk \, dr_i \right].
\]

(36)

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

\[
LSE(a, r_i, L_i) = r_i \left( L_i + \frac{1}{3} \right) + \frac{r_i}{a - 2} \left[ (4 - a) - \frac{(1 - L_i)(a^2 + a(1 - L_i) - 2)}{L_i^2 - (a + 2)L_i + (a + 2)} \right] + a(a - 2) \log(a + 2) - r_i \frac{a + 2}{a - 2} \times \frac{3(a^2 - 2a + 6)}{4(4 - a^2)} \arctan \left( \frac{2}{2 + a} \right) - \frac{4(a - 1)}{\sqrt{(4 - a^2)}} \arctan \left( \frac{-L_i}{\sqrt{L_i^2 - (2 - a)(2 + a)}} \right). \tag{37}
\]

Alternatively, a single nonparametric optimal value for parameter \( a \) is the one minimizing the integral of (36) for a fixed value of loss tolerance \( L_i \) and calculated over a continuous range of isolated rates from 0 to \( r_{\text{max}} \) where \( r_{\text{max}} \) indicates the maximum feasible value of the receivers isolated rates. Considering the available bandwidth ranges, a feasible value for \( r_{\text{max}} \) is 1 Gb/s.

\[
\min_a [LSE(a)] = \min_a \left[ \int_0^{r_{\text{max}}} \int_0^{r_i} \left( \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{g_k}{r_i} \right)^2 \, dgk \, dr_i + \int_0^{r_{\text{max}}} \int_{r_i}^{r_i^2} \left( \frac{(2 + a)r_i g_k}{g_k^2 + ar_i g_k + r_i^2} - \frac{r_i}{g_k} \right)^2 \, dgk \, dr_i \right].
\]

(38)

In solving the problem, 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. Fig. 11 plots the optimal value of parameter \( a \) versus the loss tolerance percentage \( L_i \). Reviewing the results of 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 \([-1.0012, -1.5153]\). In our calculations, we extract the optimum \( a \) by performing a simple table look up operation.

We also note that with the proper choices of parameters in the general form of (32) with \( M = 2 \), one can potentially model any function belonging to the class of fairness utilities satisfying the conditions defined in [14].

### 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 (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 straightforward. First, we note that the constraint set of (13) is reduced to a single constraint in the form of

\[
g_K \leq \text{BWF}_K \tag{39}
\]

considering the fact that the group rate \( g_K \) is the aggregate rate of layers \( 1, \ldots, K \) according to (1). The problem of (15) and (16) can then be solved the same way as described in Section III by simply substituting \( \text{BWA}_k \rightarrow \text{BWF}_k \) for \( k = 1, \ldots, K - 1 \). In the case of replicated media sessions, the constraint set of (13) is reduced to a single constraint in the form of

\[
\sum_{k=1}^{K} g_k \leq \text{BWF} \tag{40}
\]

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 (15) as

\[ L_{\text{IRF}} = \sum_{k=1}^{K} \mu_k (g_k - BWF_k) + \lambda \left( \sum_{k=1}^{K} g_k - BWF \right) \]

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

\[ \max_{g_1, \ldots, g_K} \mathcal{L}_{\text{IRF}} \]

subject to

\[ \sum_{k=1}^{K} \mu_k (g_k - BWF_k) + \lambda \left( \sum_{k=1}^{K} g_k - BWF \right) \]

for \( A(g^*) = \{ k \mid g_k^* - BWF_k = 0 \} \) and \( \lambda \leq 0 \)

\[ \max_{g_1, \ldots, g_K} \mathcal{L}_{\text{IRF}} \]

subject to

\[ \sum_{k=1}^{K} \mu_k (g_k - BWF_k) + \lambda \left( \sum_{k=1}^{K} g_k - BWF \right) \]

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

Considering the fact that the Lagrangian function \( L_{\text{IRF}} \) satisfies all of the conditions mentioned above, finding the optimal solution is equivalent to finding the solutions of (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 \( k = 1, \ldots, K \) 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 (12) and (40):

\[ \frac{\partial L_{\text{IRF}}}{\partial g_k} \bigg|_{g_k^*} = \left( \sum_{i \in G_k} \frac{r_i(2 + \alpha)r_i g_k}{g_k^2 + \alpha r_i g_k + r_i^2} + \mu_k + \lambda \right) g_k^* \]

where \( k = 1, \ldots, K \). The solution to the nonlinear system of \( 2K + 1 \) equations and \( 2K + 1 \) unknowns can be obtained by finding the positive real root of (47) such that \( r_{k_{\text{max}}} \leq g_k^* \leq r_{k_{\text{min}}} \) where \( r_{k_{\text{max}}} \) and \( r_{k_{\text{min}}} \) indicate the 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 (47) holds. The time complexity of solving for the root of this equation over all of the existing groups is \( O(KN \log N) \) and determines the overall complexity of the solution considering the fact that the rest of calculations are in the time complexity order of \( O(N) \). 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).

**REFERENCES**


The function \( f : \mathbb{C} \to \mathbb{R} \) defined over the convex set \( C \subset \mathbb{R} \) is called concave if \( v_{1,2} \in C \) 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.


Hamid Jafarkhani received the B.S. degree in electronics from Tehran University, Iran, in 1989 and the M.S. and Ph.D. degrees both in electrical engineering from the University of Maryland at College Park in 1994 and 1997, respectively.

From June 1996 to September 1996, he was a summer intern at Lucent Technologies (Bell Labs). He joined AT&T Labs-Research as a Senior Technical Staff Member in August 1997. Later, he was promoted to a Principle Technical Staff Member. He was with Broadcom Corporation as a Senior Staff Scientist from July 2000 to September 2001. Currently, he is an Associate Professor in the Department of Electrical Engineering and Computer Science at the University of California, Irvine, where he is also the Deputy Director of Center for Pervasive Communications and Computing. Dr. Jafarkhani ranked first in the nationwide entrance examination of Iranian universities in 1984. He was a co-recipient of the American Division Award of the 1995 Texas Instruments DSP Solutions Challenge. He received the best paper award of ISWC in 2002 and an NSF Career Award. He is an associate editor for the IEEE COMMUNICATIONS LETTERS and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Amir Habibi received the B.S. degree in computer science from Sharif University of Technology in 1994.

Since then, he has been working in different areas of software engineering with a focus on distributed systems and applications. Some of his large scale software projects include implementation of NetBEUI and CIFS protocols for Linux operating system as well as object-oriented scalable distributed message processing engine for interactive web applications. Recently, he has worked on algorithms for optimized routing of ADA trips in transportation industry. His research interests include distributed processing, operating systems, and computer networks.