A Generalized Satisfaction Oriented Bandwidth Management

Guowang Miao, Zhisheng Niu
State Key Lab on Microwave and Digital Communications
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R. China,
Telephone: +86-10-62794328 Fax: +86-10-62773634

Abstract—While there are other criterias to measure QoS, bandwidth is of interest to the largest number of applications. The bandwidth allocation problem becomes more difficult in future multi-service networks for QoS oriented services. Hence efficient bandwidth allocation is among the most important open issues. There are two prevailing classes of bandwidth allocation: static allocation and dynamic allocation. Since static bandwidth allocation policies lack adaptive mechanism to combat the dynamics in network, many studies have already focused on dynamic bandwidth allocation schemes. Lesiak et al. [3] has given a description of the threshold-based algorithm, where a bandwidth modification request is issued wherever a threshold is crossed. [1] provides a harmonic proportional allocation scheme to guarantee proportional sharing in terms of streaming bit rate between classes with different priorities. [2] consider that applications can tolerate a certain degree of QoS degradation. With QoS metric: degradation ratio and upgrade/degrade frequency, [2] exploits an adaptive bandwidth allocation for QoS provisioning with multilevel degradable quality. Based on differential pricing, [4] analyzes the fairness and truthfulness properties from a game-theoretic perspective and gets the allocation result from Nash equilibrium to satisfy the max-min fairness.

These algorithms work well sometimes, and give allocation in different QoS forms. However, how to deal with both heavy and light load situations together for QoS differentiation is still an open issue. Besides, the characteristics of incoming stream which might greatly injure the bandwidth exploitation are rarely considered. In this paper, we would like to introduce a user oriented QoS model and a generalized bandwidth allocation scheme which actively adapts to the dynamics of the network. The QoS policy for multimedia traffic has great flexibility that effective QoS differentiation can be guaranteed in both heavy and light load conditions. For heavy load cases, the allocation results will satisfy all users as much as possible. For light load cases, all redundant bandwidth will be fully exploited while distinct QoS differentiation is kept. Based on our scheme, a soft system capacity similar to the one in CDMA networks will be provided.

I. INTRODUCTION

The information running through the networks is increasing dramatically. Meanwhile, the information flow strength is also varying sharply from time to time. Though with the development of the communication equipments, transmission conditions are getting better and better. Due to network fluctuations or occasional unavailability of resources, some connections that could otherwise have been accepted if the traffic load were better balanced are instead rejected. Hence, the sharp variation of traffic flow still urges us to develop effective traffic control mechanisms to adapt these ongoing changes, which is getting more and more important in both the academics and the telecommunication industry. With sophistication of communication technology, services provided to customers are divided more and more delicately, and service providers will charge correspondingly. To help this development, effective Quality of Service(QoS) guarantee policies are needed. There are two basic types of service differentiation schemes([5]). One is absolute service differentiation in which users receive absolute share of resource usages. The second is relative service differentiation in which QoS of higher priorities is better or at least no worse than lower priorities.

While there are other criteria to measure QoS like delay or jitter, bandwidth is of interest to the largest number of applications. The bandwidth allocation problem becomes more difficult in future multi-service networks for QoS oriented services. Hence efficient bandwidth allocation is among the most important open issues. There are two prevailing classes of bandwidth allocation: static allocation and dynamic allocation.

1This paper has been partially supported by the National Nature Science Foundation of P.R.China (60272021)
multicast traffic flows, each of which will be received by a user or a group of users. The reallocation of the bandwidth will be executed when some existing flow has finished its transmission or a new connection needs to be established.

A. Network Dynamic Characteristics

Thanks to the dynamic characteristics of network, for each running flow, the variation of the incoming stream, which is decided by the end-to-end condition between remote ISP and current node, must be considered for efficient data transmission. The condition we need to obtain here is end-to-end bandwidth. Owing to the varying load of both network links and nodes, the end-to-end bandwidth varies between different connections or even changes rapidly during one connection time. Hence the varying data rate of incoming stream may be lower than the data rate request the flow carries and lower than the bandwidth allocated by the current node, which causes the inefficient utilization, since redundantly assigned bandwidth may have been used to carry other user’s data.

Measuring end-to-end bandwidth is an tough problem frequently discussed recently, and many policies and techniques have been proposed [6], [7], [8], [9]. The varying bandwidth of each incoming flow is measured, calculated and predicted, and to be provided to help make sure that the bandwidth allocated to each flow no more than the available bandwidth from ISP to the current node: \( R_i \leq B_i \), where \( R_i \) is the bandwidth assigned to flow \( i \), and \( B_i \) is the available end to end bandwidth from the ISP to the current node for flow \( i \).

B. User Request

Each flow corresponds to a user for unicast traffic or a group of users for multicast traffic. The requests have been carried inside the flow. Flow \( i \) is marked with \( P_i \) to represent the priority, where \( P_i \) is set to \( K (K > 1) \) levels. Flows with higher priority will receive better services. Each flow gives a basic data rate request: \( R_i \geq C_i \), where \( R_i \) is the data rate flow \( i \) hopes to obtain, and \( C_i \) is the minimum data rate request. Due to the dynamic characteristics of both network conditions and available bandwidth of current node, data rate obtained by each flow varies from time to time. And it is possible that the system cannot provide the required data rate to the user, which means sometimes certain user cannot be satisfied.

C. User Satisfaction factor

Users cannot always be satisfied because of the limited available bandwidth of the current node and the varying network conditions. Hence in order to supply differentiated services to users with different priorities, it is important to give some yards to measure the services every customer acquires. The user satisfaction factor is defined for this purpose: \( \gamma_i = f_i(S_i) \) where \( S_i \) is the service user \( i \) receives, \( f_i(\bullet) \), with different forms, is the satisfying function for calculating contentment of user \( i \), and \( \gamma_i \) is user satisfaction factor representing how much user \( i \) is content with the obtained service. Usually, \( \gamma_i \) is scaled from 0 to 1, and the bigger \( \gamma_i \) is, the more satisfied will user \( i \) be.

Different satisfying functions should be designed for different applications. We design satisfying functions for two typical applications. Denote \( R \) for the bandwidth allocated, \( C \) for the data rate requirement given by user, and \( \gamma_0 \) for the basic satisfaction factor when the user is assigned with a bandwidth of \( C \). Then the first satisfying function, designed for critical real-time applications like VOIP, video conferences, etc, is shown in formula (1). The second one, designed for rate-adaptive applications which can adjust the transmission rate in response to network congestion, is shown in formula (2).

\[
\gamma = f(R) = \begin{cases} 
0 & 0 < R \leq \gamma_0 C \\
\gamma_0 - \gamma_0 \sqrt{1 - \frac{(R - \gamma_0 C)^2}{(1 - \gamma_0)^2 C^2}} & \gamma_0 C < R \leq C \\
\gamma_0 (1 - \gamma_0) \frac{1}{1 - \frac{(R - (2 - \gamma_0) C)^2}{(1 - \gamma_0)^2 C^2}} & C < R \leq 2 - \gamma_0 C \\
1 & R > 2 - \gamma_0 C
\end{cases}
\]

(1)

\[
\gamma = f(R) = \begin{cases} 
\gamma_0 & 0 < R \leq C \\
\frac{\gamma_0 R^2}{(1 - \gamma_0)^2 C^2} & 1 - \frac{(R - \gamma_0 C)^2}{(1 - \gamma_0)^2 C^2} \geq 0 \\
1 & R > C
\end{cases}
\]

(2)

These two satisfying functions have been depicted in the subfigure (a) and (b) of Fig.(1) respectively. (a) shows a critical real-time application satisfying function in which \( C = 1.024Mbps \) and \( \gamma_0 = 0.9 \), and (b) shows a rate-adaptive application where \( C = 1.024Mbps \) and \( \gamma_0 = 0.5 \).

![Fig. 1. A Typical Relationship between Allocated Bandwidth and User Satisfaction](image)

D. Satisfaction Oriented Bandwidth Allocation Model

To satisfy all users’ requirement as good as possible is our purpose. Those with different priorities should receive different qualities of services, measured by how they are satisfied, and users with higher priority will be more satisfied. This can be expressed with: \( E(\gamma_i) = g(P_i) \), where \( E(\gamma_i) \) is the mean value of the satisfaction factor for user \( i \), and \( g(\bullet) \) is the priority-satisfaction function which reveals the relationship between priority and satisfaction factor. The following linear priority-satisfaction function \( g(\bullet) \) is assumed in our policy:

\[
E(\gamma_i) = g(P_i) = \gamma_0 + (P - 1) \frac{\gamma_m - \gamma_0}{K - 1}
\]

(3)

where \( \gamma_0 \) is basic satisfaction factor of users with priority 1, and \( \gamma_m = \max\{f_i(S), i = 1, \ldots, N\} \), where \( S \) is the maximum capability of system. \( N \) is the number of active flows. Given the satisfaction factor, the corresponding priority is: \( P_i = g^{-1}(\gamma_i) = (K - 1) \frac{\gamma - \gamma_0}{\gamma_m - \gamma_0} + 1 \). Hence, with allocation of \( S_i \), the actual priority user \( i \) obtains can be calculated.

Our objective is twofold: first, we want to increase satisfaction of all users; second, we want to assure fairness among different users. Besides, thanks to the coexistence of unicast and multicast flows, different satisfaction weight should be
considered, since the service provided for a multicast flow will be perceived by many users. However, we could not simply add all the users satisfaction. Because in that case, a multicast flow with lots of receivers will probably occupy all the bandwidth of the current node to achieve a large total satisfaction, which does not make too much sense and would be unfair to other flows.

In order to take both the difference and fairness into consideration, the flow satisfaction factor for each flow is defined: 

\[ \alpha_i = \left(1 - \frac{M}{(U_i)^E}\right) \frac{P_i^E}{P_i}, \]

where \( i \) is the flow ID, \( P_i \) is the expected priority, \( P_i^E \) the actual priority provided by the current node, and \( U_i \) is the number of users that will receive this flow while \( U_i = 1 \) for unicast flows, and \( U_i > 1 \) for multicast flows. \( M, E \) are constant given by: \( M = 1 - \frac{1}{E}, E = \log_2 \frac{M}{1 - M} \), where \( K \) is the maximum priority. \( \alpha_i \) is the flow satisfaction factor. Our purpose is to maximize the satisfaction of all flows, defined as: \( \text{TS} = \sum_{i=1}^{N} \alpha_i \), where \( N \) is the number of active flows.

At the reallocation time, the bandwidth that has been allocated to the flows of critical real-time applications should not be reallocated. Since users with these applications would be even unhappier if their applications are suddenly interrupted for the bandwidth reallocation. Assume that the total bandwidth of the current node is \( R \), the bandwidth that has been occupied by critical real-time applications is \( R_c \). The remaining bandwidth will be reallocated by the other flows or new connections. And totally, assume that \( N \) flows either unicast or multicast will share the available bandwidth. Considering the varying data rate of incoming streams, our optimization objective is:

\[ \text{TS}_1 = \max \left\{ \sum_{i=1}^{N} \left[1 - \frac{M}{(U_i)^E}\right] \frac{g^{-1}(\gamma_i(R_i))}{P_i} \right\}, \quad \text{TS}_2 = \max \left\{ \sum_{i=1}^{N} \left[1 - \frac{M}{(U_i)^E}\right] \frac{g^{-1}(\gamma_i(R_i))}{P_i} \right\}, \]

where \( B_i \) is the data rate of incoming stream of flow \( i \) and without considering the varying data rate of incoming streams, the optimization objective will be:

\[ \text{TS} = \sum_{i=1}^{N} \alpha_i, \]

III. SATISFACTION ORIENTED BANDWIDTH ALLOCATION METHOD SUPPLYING QOS DIFFERENTIATION

The bandwidth allocation according to our policy is nonlinear programming which contains nonlinear objective function and linear constraints. Many techniques have been proposed for solving constrained optimization problems([10], [11]). Considering the specific form of our optimization objective and its constraints, our bandwidth allocation algorithms will be based on Rosen’s gradient projection method([12]), which is based on successive projections on the subspace tangent to the active constraints, and has inexpensive operations when the constraints are simple. The Kuhn-Tucker conditions will be used as a termination criteria.

The objective function and its constraints in formulas (4) and (5) can be written in the following way:

\[ -\text{TS}_1 = G(R) = \min \left\{ -\sum_{i=1}^{N} \frac{g^{-1}(\gamma_i(R_i))}{P_i} \right\}, \]

\[ s.t. \quad A \cdot R \geq B \]

for \( I = 1, 2 \). \( A \) can be \( A_1 \) or \( A_2 \) for \( \text{TS}_1 \) or \( \text{TS}_2 \) respectively. \( A_1 \) and \( A_2 \) are \((2N+1)\times N\) and \((N+1)\times N\) matrixes respectively:

\[ A_1 = \begin{pmatrix} -1 & \cdots & -1 \\ \vdots & \ddots & \vdots \\ -1 & \cdots & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -1 & \cdots & -1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \]

\( R \) is \( N \times 1 \) vector; \( R = [R_1, B_2 \cdots R_N]^T \); and \( B \) is \((2N+1)\times 1\) vector for \( TS_1 \) optimization; \( B = [-B_1 - B_2 \cdots - B_N - R + R_{DV}B 0 \cdots 0]^T \); and \((N+1)\times 1\) vector for \( TS_2 \) optimization: \( B = [-R + R_{DV}B 0 \cdots 0]^T \). \([\bullet]^T\) means the transpose of vector or matrix \([\bullet]\).

At each iteration point \( R^{(k)} \), the next better feasible point \( R^{(k+1)} \) is found to improve the objective value along a feasible direction. If the current feasible point exists inside the feasible region, the feasible direction will be the minus gradient of the objective function. Otherwise, if the current feasible point is on the working surface(some active constraints exists), the direction generated by projecting the minus gradient to the null space of the effective active constraint matrix will be the feasible direction for searching better objective values.

The gradient of the objective functions at \( R^{(k)} \) is:

\[ \nabla G(R^{(k)}) = \frac{\partial G(R^{(k)})}{\partial R_1}, \frac{\partial G(R^{(k)})}{\partial R_2}, \ldots, \frac{\partial G(R^{(k)})}{\partial R_N} \]

\[ (7) \]

where:

\[ \frac{\partial G(R^{(k)})}{\partial R_i} = -\frac{1}{P_i} \frac{\partial g(\gamma_i(R^{(k)}))}{\partial (\gamma_i(R^{(k)}))} \frac{\partial (\gamma_i(R^{(k)}))}{\partial (P_i)} \]

In which \( P_i' \) can be obtained with parameter \( \gamma_i(R^{(k)}). \)

At feasible point \( R^{(k)} \), reorganize the constraints \( A_R \geq B_1 \) into active constraints: \( A_1 R^{(k)} = B_1 \), and inactive constraints: \( A_2 R^{(k)} > B_2 \), where: \( A = [A_1 A_2]^T \), \( B = [B_1 B_2]^T \). Then, the projection matrix of \( A_1 \) is: \( P = I - A_1^T (A_1 A_1^T)^{-1} A_1 \) (note \( A_i \{P \gamma \} = 0 \) for all \( \gamma \), so \( P \gamma \) is the null space of \( A_1 \)) if the current feasible point \( R^{(k)} \) is inside the feasible region, \( A_1 \) would be a null matrix and let \( P \) be a unit matrix.

Project the objective function’s gradient to the null space of active constraint matrix, and if the projection satisfies \( P \nabla G(R^{(k)}) \neq 0 \), then the new feasible direction would be:

\[ d^{(k)} = -P \nabla G(R^{(k)}). \]

\[ (8) \]
Otherwise, let \((A_1 A_1^T)^{-1} A_1 \nabla G(R^{(k)}) = U\). If \(U \geq 0\), \(R^{(k)}\) is the Kuhn-Tucker point, and the optimal bandwidth allocation \(R = R^{(k)}\) is found. For \(U \geq 0\), find \(u_i = \min_i \{u_i\} < 0\), where \(u_i\) is the \(i\)th element of vector \(U\). Then remove the \(i\)th row of \(A_1\) corresponding to \(u_i\) and get the new effective constraint matrix \(A_i\). The new projection matrix will be: \(P = I - (A_1^T)^T (A_i^2 A_1^T)^{-1} A_1^T\) and the new feasible direction is:

\[
d^{(k)} = -P \nabla G(R^{(k)}),
\]

Hence, the new feasible point \(R^{(k+1)}\) for next iteration is:

\[
R^{(k+1)} = R^{(k)} + \lambda^{(k)} d^{(k)}
\]

in which \(\lambda^{(k)}\) satisfies:

\[
G(R^{(k)} + \lambda^{(k)} d^{(k)}) = \min \{G(R^{(k)} + \lambda^{(k)} d^{(k)}) \mid 0 \leq \lambda^{(k)} \leq \lambda_{max}\},
\]

where \(\lambda_{max}\) is:

\[
\lambda_{max} = \min_i \{a_i \frac{R^{(k)} - 1}{a_i d^{(k)}} \mid a_i d^{(k)} < 0\},
\]

in which \(a_i\) is the \(i\)th row item of constraint matrix \(A\), and \(b_i\) is the \(i\)th element of vector \(B\).

The objective value improves each iteration until Kuhn-Tucker condition satisfies, where the optimal allocation is achieved. The complete satisfaction oriented bandwidth allocation method(SOBAM) is described in Table I(for detailed theoretical analysis, please refer to [12]).

### Table I: SOBAM

<table>
<thead>
<tr>
<th>Step</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>Initialization.</strong></td>
<td>Set the initial feasible point (R^{(0)}, k = 0)</td>
</tr>
<tr>
<td>2. <strong>Projection.</strong></td>
<td>Terminate if (R^{(k)}) is KT point; otherwise, get (d^{(k)}) according to (8) or (9)</td>
</tr>
<tr>
<td>3. <strong>Line search</strong></td>
<td>Get (R^{(k+1)}) according to (10), in which (\lambda^{(k)}) is determined by line search</td>
</tr>
<tr>
<td>4. <strong>Updating</strong></td>
<td>Set (k \leftarrow k + 1), and go to 2.</td>
</tr>
</tbody>
</table>

## IV. Simulation Performance and Capacity Analysis

Apply our algorithm to a router. Assume that the arrivals of new connections are Poisson processes with gaussian distributed data rate requirement, and the lifetime of each connection is exponentially distributed. Data flows are classified into 4 priorities. The basic satisfaction factor is 0.2 for unicast data flows, and 0.9 for multicast data flows.

First, consider the coexistence of unicast and multicast flows. A multicast flow containing DTV program and three unicast flows generated from FTP content downloading will compete for the available 10Mbps router bandwidth. The multicast flow has priority 2, and has 20 receivers. Thanks to the compression of original videos and the traffic smoothing techniques used in the MPEG-2 encoder, the bandwidth required for transmitting DTV programs varies slightly from time to time. Hence, the multicast flow is assumed to be near CBR stream having Gaussian distributed bandwidth requests with mean 6Mbps and variance 100Kbps. The three unicast flows, marked with priorities 4,3 and 1, have basic data rates requests 256Kbps, 384Kbps and 1.024Mbps respectively. The Internet link bandwidth between the router and the ISP servers for the three unicast flows are assumed to be gaussian distributed with mean 2.3Mbps, 1.3Mbps, 1.9Mbps and variance 0.3Mbps, 0.2Mbps and 0.5Mbps respectively. We collect the simulation results for a random 100 reallocation times when these four flows coexists. Fig.2 shows how these flows are satisfied during each bandwidth reallocation for illustration, in which the four curves are marked by “M_P2”, “U_P4”, “U_P3”, “U_P1”, corresponding to the flows above. The simulation results prove that the SOBAM algorithm can effectively supply services to users with different satisfaction according to their QoS level. The mean satisfaction of users receiving the four flows are 0.94, 0.87, 0.75 and 0.54 respectively.

Fig.3 presents the simulation results examining the influences of incoming stream characteristics to the utilization efficiency of current router bandwidth. Two unicast streams A and B with priorities 1 and 3 are assumed to compete for the remaining 4Mbps router bandwidth with data rate requirement 1.536Mbps and 2.048Mbps respectively. Meanwhile The link between the router and remote ISP for stream B is supposed to be very well, and is gaussian distributed with mean 3.072Mbps and variance 500Kbps. As shown in Fig4, the average satisfaction of users receiving the four flows are 0.96, 0.87, 0.75 and 0.54 respectively. The mean satisfaction of users receiving the four flows are 0.94, 0.87, 0.75 and 0.54 respectively. As shown in Fig4, the average satisfaction of users receiving the four flows are 0.94, 0.87, 0.75 and 0.54 respectively. The mean satisfaction of users receiving the four flows are 0.94, 0.87, 0.75 and 0.54 respectively.

Based on SOBAM, there’s no upper bound of the number of flows that can be connected simultaneously. Hence SOBAM has soft capacity similar to the one in CDMA systems. Any new flow can compete with the other existing flows for connection any time, which will cause the bandwidth reallocation and make users receiving other flows less satisfied. Fig4 gives the simulation result, and shows the relationship between average user satisfaction and system capacity(measured by user number) under different settings(basic satisfaction factor \(\gamma_0\)). We assume that all flows are unicast flows with the same priority, and are requesting ftp services with basic data rate 384Kbps. Seven curves are plotted to illustrate the simulation results, and are marked by \(\gamma = 0, 0.1, 0.2, \ldots, 0.6\) to show the capacity-satisfaction relationship when \(\gamma_0\) is configured to be \(0, 0.1, \ldots, 0.6\) respectively. As shown in Fig4, the average satisfaction decreases in a quasi linear way with the increase of flows. However, users cannot be too disappointed. The router may choose to reject new comers when the average user satisfaction has been as low as a certain degree, e.g. \(\gamma_0\), rather than provide worse services to all users. Thus a basic service level can be guaranteed.
V. CONCLUSIONS AND FUTURE PLAN

With new QoS concepts: user/flow satisfaction, this paper has introduced a user oriented QoS model and a generalized bandwidth allocation scheme which actively adapts to the dynamics of network. The kind of mechanisms developed is a generalized scheme suitable for deployment at various routing points in the communication traffic which may consist of both unicast and multicast traffics. Our approach allows a service provider to differentiate the services between different types of customers based on their priority and service type. The QoS policy for multimedia traffic has great flexibility that effective QoS differentiation can be guaranteed in both heavy and light load conditions. For heavy load cases, the valuable bandwidth will be allocated to satisfy the users as much as possible. And for light load cases, all the redundant bandwidth will be fully exploited while distinct QoS differentiation is kept. The allocation scheme SOBAM contains two allocation methods: the first one considers the incoming stream characteristics and can be used for bad Internet conditions to make efficient use of the current node bandwidth, while the second one doesn’t consider the incoming stream characteristics and can be used for good Internet conditions for execution simplicity. The simulation results powerfully show that SOBAM give good QoS differentiation to different users and make plentiful utilization of all available bandwidth. With SOBAM, the system can provide a soft system capacity which is similar to the one in CDMA networks.

REFERENCES


