Journal of Science and Technology, Vol. 32 No. 2 (2012), pp 94-103 94

http://dx.doi.org/10.4314/just.v32i2.11

RESEARCH PAPER

IMPLEMENTATION OF INTERNET PROTOCOL NETWORK ARCHITECTURE FOR EFFECTIVEBANDWIDTH ALLOCATION IN A MULTIPARTY, MULTIMEDIA CONFERENCING

M. Asante
Department of Computer Science, College of Science, KNUST, Kumasi

ABSTRACT

Advances in multimedia technologies and development of overlay networks foster the opportunity for creating new value-added services over the current Internet. In this paper, a new service network architecture that supports multiparty multimedia conferencing applications, characteristics of which include multi-channel, high bandwidth and low delay tolerance has been proposed. The new service network architecture is built on an array of service nodes called Multiparty Processing Centers (MPCs) which constitute a service overlay network, serving as the infrastructure for multiparty conferencing, and are responsible for conferencing setup, media delivery and the provision of Quality of Service. In this paper, the main focus is on the bandwidth allocation management over the proposed service network. The analysis will determine the bandwidth demand for virtual links among the MPCs. Multimedia traffic is modeled as M/G/∞ input processes and divided into several classes, with the constraint that the aggregate effective bandwidth is within the link capacity times a prescribed utilization threshold.

Keywords: *Overlay Networks, Multiparty Multimedia Conferencing, Bandwidth Allocation, Quality of Service (QoS), Multiparty Processing Centers (MPC)*

INTRODUCTION

Advanced multiparty conferencing with multi-channel media capture and playback, such as those promised in MPEG-4 related applications, provides a setting that allows participants to attend a meeting over the Internet in real time with the goal of creating an immersive experience for the participants similar to that of an actual meeting in our daily life. In multiparty conferencing, in terms of meeting setup, participants may join and leave the conferencing session at any time. In terms of content, multiparty conferencing applications involve transmission and reception of multimedia streams including audio, video and, of course, text. These media data streams may be acquired and organized for transmission in a multi-channel fashion whereby multiple inputs and multiple outputs (MIMO), e.g., multiple microphones, loudspeakers, cameras and displays, are present for each type of media. Several evident technological advances have emerged and the availability of more and more network resources that contribute to multimedia conferencing is more likely to be the telecom service of the future. With the ubiquity of the Internet, it is natural to conduct multiparty multimedia conferencing over the Internet. However, due to its original design, purpose and the limitation of its current protocols and architecture in operation, the current Internet cannot effectively support multiparty multimedia conferencing applications. These limitations can be viewed from three aspects: routing, multicasting, and QoS.

The absence of efficient support of these functionalities poses challenges for multimedia conferencing applications to be successful in the current Internet.

Routing in the Internet is not optimal due to the nature of mechanisms employed in the border gateway protocol (BGP) (Rekhter and Li, 1995). Unfavorable events such as link failure, router failure, and network congestion may take a long time to reflect in routing decisions; therefore they are likely to cause excessive delay, serious packet loss, and even disruptions of ongoing communications. As a result, multimedia streams transmitted over the Internet may not reach the receivers for signal reconstruction with reasonable quality. In multiparty conferencing applications, multimedia streams originating from one participating site are distributed to many other participants (Weiss, 1995). Efficient group distribution of the media stream can be achieved with a multicasting protocol (Deering and Cheriton, 1990*).* In reality, multicasting is not available in the Internet in an end -to-end manner. Quality of service (QoS) (Braden *et al.*, 1994; Blake *et al.*,1998) is imperative in multiparty conferencing applications, which entails stringent delay constraints, high bandwidth and unimpeded interactivity.

QoS is not supported in an end-to-end manner either, due to lack of incentive for ISPs to employ it as many still rely on the traditional best effort service model offering their subscribers only Internet browsing and email as services. Recognizing the difficulties in the deployment of IP multicasting and QoS, a new service network architecture that has the potential of providing an effective venue for multiparty conferencing and collaboration but does not require a dramatic change in the current mode of IP network operation is proposed.

There are also a discussion on algorithms and approaches that address various issues under the new network architecture as they are potentially, key to the success of multiparty conferencing and other services. The proposed architecture is built upon an overlay network in which dedicated servers function as service nodes. This architecture allows a service provider to provide multiparty multimedia conferencing service to their subscribers.

OVERVIEW OF THE PROPOSED SERVICE OVERLAY NETWORK ARCHITECTURE

In view of the rigidity of the operating mode of the current networks, researchers have attempted to develop alternative networks for the Internet. One of several promising solutions that have attracted attention is the overlay network. Built on top of existing networks, called the native networks, an overlay network can provide new services and functionalities that allow fast deployment without the need for modification of current networks in which case it addresses various applications, including multicasting (Chawathe *et al.*, 2000*;* Subramania *et al.*, 2004*),* content distributions and peer-to-peer file sharing (Stoica, 2001*).* Overlay-Networks can be implemented by two approaches: the end-host approach (Chu and Zhang, 2002*)* and the proxy-based approach (Chawathe *et al.*, 2000*;* Jannotti *et al*, 2000*).* In the first approach, an array of end hosts form a self-organized topology, and each host can send or relay data. In this approach, data are routed along end hosts to the destinations. This feature may incur unduly long data latency, which is not desirable in multimedia stream delivery (Chuand Zhang, 2002*).* Proxy-based overlay networks employ powerful proxy servers that constitute the core of the provision of data distribution and allocate network resources more efficiently with the benefit of multiplexing thus maintaining QoS (Kashihara *et al.*, 2009).

To support multimedia and multiparty conferencing applications, proxy-based approach is used to propose a Multiparty Processing Center (MPC)-based overlay network as the basis for the approach in order to allow a third party service provider to provide value-added services for its customers. In this architecture, as illustrated in Fig. 1, where solid lines represent

physical connection and dashed lines denote logical connection, MPCs are located at some point of a network domain and act as clients from the perspective of that network domain (Zhou and Xu, 2004). The links among MPCs in the overlay network are virtual links which actually go through the Internet via different Internet Service Providers (ISPs) and users communicate with MPCs to obtain multiparty conferencing service via regular unicast packet delivery. MPCs perform the role of a bridge to bring new services to the subscribers (users). The service provider acquires bandwidth for MPCs from (infrastructure) ISPs. Under stringent delay constraints and interactivity requirement of conferencing applications, all the delivery for streams between participants and MPCs and among MPC uses the Real time Protocol and the User Datagram Protocol (RTP/UDP) (Yao and Chen, 2009*;* Schulzrinne *et al.,* 1996).

MPC structure

Usually configured as powerful servers with high access bandwidths and processing capabilities, MPCs are mainly responsible for building an overlay network to support multicasting, maintaining membership (both subscription and session) information in communications, providing user authentication and billing information, allocating resources to sessions to guarantee QoS, and performing media processing when applicable. Each user is mapped to an optimal MPC with regard to the conditions of the overlay network for possible network load balancing (with factors such as delay profile, available bandwidth between MPCs and the user, etc.). An MPC may serve a group of participants, each of which finds its optimal MPC and registers with the MPC before communicating with the others. For a specific conferencing session, the MPC through which the presenter is connected to the overlay network is termed entry MPC, while the MPC serving other participants is termed exit MPC. MPCs maintain the overlay network topology collectively by exchanging virtual link information.

Operation of a conference session

Participants get access to the Internet by connecting to the access networks (part of the native networks). The communication between a participant and its optimal MPC is conducted with best-effort unicasting as the usual access networks would provide. The streams from the presenter are delivered to the optimal MPC (entry MPC) via uni-cast path. Other participants inform their MPCs (exit MPCs) to get the streams for the conferencing session they want to join or to stop sending streams when they leave the current conferencing session. Fig. 2 illustrates how a conferencing session is conducted over the MPC-based overlay network.

Bandwidth allocation in the overlay network

Multimedia traffic is bandwidth consuming and the overlay network actually is carried over the access links at MPCs therefore it is critical to have enough bandwidth for MPCs to support multiparty conferencing with reasonable quality.(Gaofeng and Biing-Huang, 2005). The first multimedia traffic was formulated with M/G/∞ input processes. Traffic demand was divided into several classes of traffic with the consideration of bandwidth requirement and QoS. For each class of traffic, effective bandwidth is calculated. According to the theory of Large Deviation Principle (LDP) (Rabby and Ravindran, 2007) and the observation by Hui (1996), it was realized that the sum of the effective bandwidth (Kelly, 1996*)* of traffic on a link is within a certain utilization threshold portion of the link capacity and quality-of-service (QoS) is assured for the traffic as the packet buffer accumulated by experiencing low delay and low packets loss (Hossain and Bhargava, 2001). Therefore the bandwidth demand was modeled as a stochastic network problem with the constraint that the sum of the effective bandwidth for traffic on a link does not exceed the link capacity times the utilization threshold. The bandwidth allocation problem was further formulated into the optimization of revenue that a service provider can earn under certain blocking rate constraint.

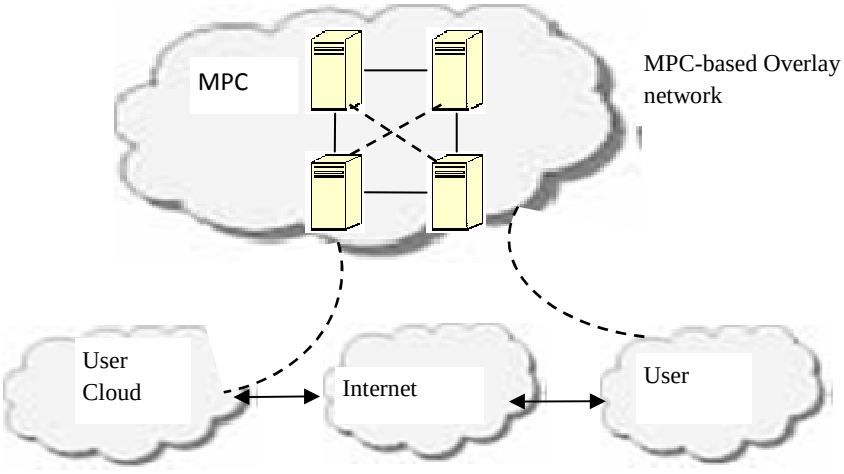

Fig 1: Overlay network architecture

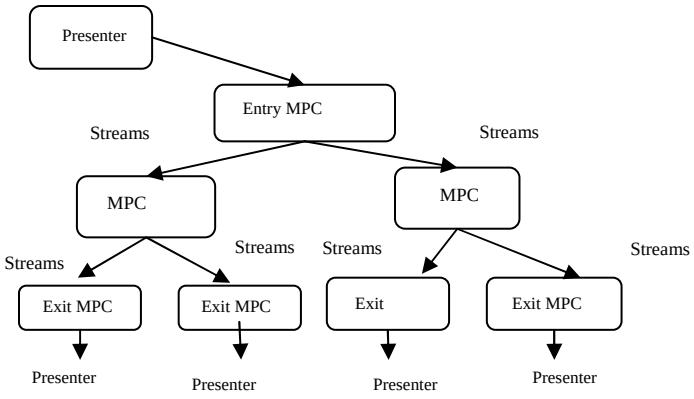

Fig 2: A conferencing session

Modeling a multimedia stream

The Markovian and Long-Range Dependence (LRD) models have been used for video traffic. In an audio-visual conferencing scenario, it was assumed that multimedia traffic is dominated by video. Therefore, a video traffic bandwidth requirement was primarily taken as the traffic demand of a multimedia stream. In Krunzz and Murkowski (1998), the authors advocated for the use of M/G/∞ processes which were shown to better characterize the video traffic. In accordance to M/G/∞, a multimedia traffic is described as a discrete-time system with an infinite number of servers. In the framework, the time was partitioned into time slots, the numbers of busy servers at the beginning of each slot were counted and the size of data in busy servers was regarded as the traffic for that slot.

However, under standard conditions, a stationary and ergodic version of such a process can be obtained, which is denoted by;

{vi, i = 0, 1, . . .}

This stationary process has special properties which makes it suitable for modeling video traffic. These properties are:

1) For each i=0, 1,. . . , *vi* is a Poisson random variable with parameter $\lambda E[\sigma]$, where 'λ' is the arrival rate in the M/G/∞ process, and'σ'is the service time.

2) $\lim_{n\to\infty} \sum_{i=0}^{n} vi = \lambda E[\sigma]$

3) The covariance of {*vi*, i=0,1,. . . }is given by

$Cov(vi, vi+k) = \lambda E[\sigma(\sigma-k)^{+}], \quad i,k$ =0,1

Modeling traffic arrival from customers

Conferencing session was categorized into three classes based on the consideration of bandwidth and QoS requirement so as to simplify the problem formulation. For traditional conferencing setting, participants schedule the conferencing start time normally at certain hour points (e.g. 9:00am, 9:30am, etc.) during business hours. Most participants join the conferencing around the start time. The distribution of conferencing start time has peaks around those hour points. Therefore, the process of overall conferencing session arrival is generally not a Poisson process. The overlay service network proposed in this paper is designed to have the capability of providing conferencing service at any time to a large amount of customers anywhere. It is reasonable to assume that each class traffic arrives according to independent Poisson processes with rates λi, i=1,. . . ,*C*. Each class' traffic (C) stays in the network for a time that is exponentially distributed with rate μi, i=1,. . . , *C*.

Where 'μj' =traffic arrival time

Traffic are assumed to be independent. When a traffic terminates (e.g. in this case, a conference ends), the bandwidth used by that traffic on those links becomes free simultaneously.

Traffic demand calculation

For its operation across the overlay network, each conferencing session requires enough bandwidth from each of the virtual links it goes across. However, the multimedia traffic rate varies with time. The effective bandwidth of the conferencing traffic was used as the traffic demand for the overlay network. In the case of no transcoding and no rate adaptation, the traffic demand is the same for all the links involved in a conferencing session. The effective bandwidth is computed as follows: consider a process {*a* (*t*), $t \geq 0$} as conferencing session traffic.

$$A(t) = \int_0^t aj(x)\,dx,\ \Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E e^{\theta A(t)},\ a^{*}(\theta) = \Lambda(\theta)/\theta$$

Where A is the absolute bandwidth at time t, is

$\Lambda(\theta)/\theta$ The effective bandwidth of

{*a* (*t*), $t \geq 0$} *a* is the bandwidth at time t and θ is the rate of transmission. Since Λ (θ) is convex, the effective bandwidth '*a**(θ)' is increasing (or non-decreasing) in θ. As $\theta \rightarrow 0$, '*a** (θ)' approaches its average rate, while $\theta \rightarrow \infty$, *a**($\theta$) approaches its peak rate. Therefore, variation of θ can change the bandwidth demand in the calculation and change QoS.

METHODOLOGY

The overlay network consists of a set '*L*' of virtual links, each of which has capacity *bi*, for $i \in L$. Let '*aj*' *be* the effective bandwidth of class '*j*' *traffic*, which can be calculated using the method used in the calculation of the effective bandwidth. The traffic is transmitted along the virtual links in a multicast tree to reach users thus any conferencing session occupies multiple links in the overlay network. Each confer-

encing session with traffic class '*j*' requires '*aij*' bandwidth each of virtual link *i* in the set '*L*'. For links serving the conferencing class ' *j* ' traffic, '*aij*'s are equal to '*aj*', and for other links '*ai*'s are 0. In order to achieve a standard Quality of Service (QoS), the utilization of each link *'i'* is kept below the prescribed threshold '*ρi*'. Since effective bandwidth is used as traffic demand, the sum of effective bandwidth of traffic across link 'i' should be less than '*ρibi*' with '*ρi*' being the prescribed threshold and *bi* the network capacity.

Consider the stochastic process n = {n_1, n_2,..... *nC*} in which '*nj*' denotes the number of conferencing sessions with class *'j'* traffic. This process is a constrained multi-dimensional birth-death process. The constraint is that the utilization of each link '*i*' *should* not exceed '*ρi*', the prescribed threshold. Each component '*nj*' of the process '*n*' is a reversible birth-death process with the same constraint and operates independently. It was assumed that the process *'n'* is still a reversible process defined on the state space 'S' in which case:

$$S = \left\{ n: 0 \le \sum_{j=1}^{c} aijnj \le \rho ibi, \epsilon L \right\}$$

The reasoning behind this claim lies in two theorems: one theorem states that a reversible Markov process restricted on a subset of its state space is still a reversible Markov process, which implies that each component '*nj*' is a reversible birth-death process with a constraint. Another theorem states that a process consisting of multiple independent reversible Markov processes is also a reversible Markov process, from which it was deduced that '*n*' is a reversible process which makes it easier (Serfozo, 2004),to obtain the stationary distribution of this process on the state space '*S*'. Actually, the constrained process has the same form of stationary distribution as the original unconstrained process, and the only difference is the normalization constant. The stationary distribution is:

$$\pi(n) = k \prod_{j=1}^{c} \left(\frac{\lambda j}{\mu j}\right)^{nj}$$

Where '*k*' is the normalization constant such that

$$\sum_{n \in S} \pi(n) = 1$$

and 'λ', 'μ' are the arrival rate and the effective rate respectively.

If for any link 'i', $\sum_{j=1}^{c} aijnj \ge \rho ibi - aij$

the class '*j*' traffic will be blocked due to the lack of bandwidth in link '*i*' such that '*aj*' is the effective bandwidth for class '*j*' traffic, '*nj*' is the number of conferencing sessions and '*bi*' is the network capacity.

From the stationary distribution, we can easily get the blocking rate for each class of traffic and total blocking rate for all traffic can be obtained. Denoting the blocking rate for each class as '*Pj*', the total blocking rate

$$P.Sj = \{n: 0 \le \sum_{j=1}^{c} aijnj \le \rho ibi - aiji, i\epsilon L\} \qquad (1)$$

is the state within which the network can admit class '*j*' traffic in a virtual link L.

$$Pj = 1 \quad \sum_{n \epsilon Sj} \pi(n) \qquad (2)$$

and

$$P = \sum_{j=1}^{c} \frac{\lambda j}{\lambda} Pj$$

where $\lambda = \sum_{j=1}^{c} \lambda j$

For a service provider to deploy this service overlay network, it definitely depends on returns on investment to maintain the business.

The service provider pays ISPs to purchase bandwidth at the cost '*ηi*' per unit of bandwidth

for link '*i*' and charges customers using the conferencing services at the rate of '*γj*' per unit of bandwidth for class '*j*' *traffic*. Sufficient bandwidth should be allocated to achieve a low blocking rate to attract customers in order to get maximum revenue while maintaining the blocking probability less than a threshold '*ε*'. The bandwidth allocation problem was formulated as an optimization problem. The objective function revenue is

$$R = \sum_{n \in S1} \pi(n) \sum_{i=1}^{c} \eta j \alpha j \gamma j - \sum_{i \in L} bi \eta i \qquad (3)$$

The optimization problem is to find optimal '*bi*'s to maximize *R* with the constraint

$$\boldsymbol{P} \leq \varepsilon. \qquad (4)$$

The objective function and the constraint are both nonlinear with respect to '*bi*'s. Sufficient bandwidth should be allocated to allow the traffic generated from a source in a conferencing activity. In order to simulate the effect of the bandwidth demand caused by interactivity, '*βj*' was introduced to denote the portion of time that participants become presenters for class 'j' traffic. The total bandwidth consumed by class 'j' traffic is determined by '*λj*' and '*β'j*. It was assumed that the extra traffic is represented by a new Poisson process with rate '*λjβj*' with the reasoning that the number of listeners switching to presenters is proportional to the number of participants. Now the compound traffic arrival process for traffic class '*j*' has rate,

$$\lambda' j = \lambda j + \lambda j \beta j$$

The same approach in the last subsection can be applied to obtain the bandwidth demand for each link with a specified blocking rate.

Optimization tools were utilized to obtain some numerical results from the solution given in equation 3. The overlay network topology considered is illustrated in Fig.3, which consists of eight virtual links. Two scenarios were examined.

One scenario is that there are only two classes of traffic sharing all the links (i.e. each link carries the exact same two classes of traffic). In the other scenario, there are three classes of traffic; any two classes could use any of the links. In the example, class 1 traffic traversed links 1, 4, 2, 6, class 2 traffic also traversed links 2, 6, 3, 8, and class 3 traffic traversed links 1, 5, 3, 7. The virtual link bandwidth '*bi*'s which are chosen to optimize revenue defined in Equation (3) with the constraint (4) are listed in Table 1 along with the specific traffic arrival parameters '*λj*'s, '*μj*'s, and the cost '*γj*'s. In the calculation, all '*ηi*'s were assumed to be 1 and link utilization '*ρi*'s was assumed to be 0.7. It must be noted that each link was considered in the computation of cost, while users were charged based on conferencing sessions and not on the number of links. This contributed to the fact that income rate 'γj' is much higher than the number of links used '*ηi*' since one conferencing session normally occupies multiple links.

RESULTS AND DISCUSSION

Based on the bandwidth allocations obtained in Table 1, several observations can be made.

In a two-class traffic case, all links were assigned the same bandwidth. This is obvious since all links were assumed to carry the same traffic. The blocking rate threshold affected the bandwidth allocation significantly in which case the bandwidth for achieving blocking rate 0.01 in a two-class traffic (i.e. 48,48,48,48,48,48,48,48) were much higher than that for 0.05 blocking rate which were (16,16,16,16,16,16,16,16). The revenue accrued was 25.4 and 20.8 respectively. Again for a blocking rate threshold of 0.01, the bandwidths were (35,31,46,15,20,31,26,20) for a three-class traffic which were much higher than that of a corresponding blocking rate of 0.05 (i.e. 21,20,25,11.10,30,13,13). The revenue accrued were 17 and 33.5 respectively. It was also observed that the lower the blocking rate, the lower the revenue, since higher bandwidth allocated for lower blocking rate costs more

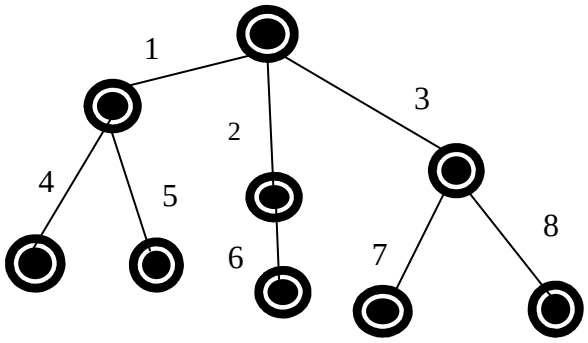

Fig. 3: Network Topology Overlay

Table 1: Numerical results for bandwidth allocation

Class	Bandwidth ‘bj’								Reve-nue R	Effective band-width rate ‘aj’	Traffic rate ‘λj’	Traffic arrival time ‘μj’	Cost ‘γj’	Block-ing rate thresh-old θ
2	10	10	10	10	10	10	10	10	3	0.5,1	2,3	3,4	30,30	0.05
2	16	16	16	16	16	16	16	16	20.8	0.5,1.0	2.5,3.5	3,4	25,30	0.05
2	48	48	48	48	48	48	48	48	25.4	0.5,1.0	2.8,3.7	3,4	25,30	0.01
2	23	23	23	23	23	23	23	23	58.6	0.5,1.0	2.8,3.7	3,4	25,30	0.05
3	21	20	25	11	10	30	13	13	33.5	0.5,0.7,1.0	2.5,3.5,4.5	3,4,5	15,20,25	0.05
3	35	31	46	15	20	31	26	20	17	0.5,0.7,1.0	2.5,3.5,4.5	3,4,5	15,20,25	0.01

and few users will utilize the link resulting in lower revenue. Provision of adequate bandwidth is a key objective when supporting many types of media applications, so setting up the initial threshold provided sufficient bandwidth for high-definition media capable endpoints. Additionally, ensuring adequate bandwidth and port interfaces provided adequate buffering capacity to handle the ‘burst’ in media applications, especially video-oriented media application.

It was also observed that an increase in the traffic load or rate at a lower threshold in a link increased the assigned bandwidth accordingly. However, larger traffic load can generate higher revenue, which implies that the network is more utilized. In the three different class traffic considered, most links have different traffic loads, with an accompanying bandwidth links, 2 and 6 in which case bandwidth resources dedicated to strict-priority queuing were all limited in order to prevent starvation of non-priority and yet business critical applications. In any case at higher bandwidth, the blocking rate will generally be low which implies that the network will be more utilized by customers resulting in higher revenue earnings. Sufficient

bandwidth should be allocated to achieve a low blocking rate to attract customers in order to get maximum revenue. Adequate bandwidth allocation can contribute in meeting the latency targets for video propagation delay which will account for over 95% of the network latency budget to achieve a low blocking rate to attract customers in order to get maximum revenue. Therefore the analysis can give the service provider an insight about the best way to manage bandwidth allocated to clients in a cost-effective way.

CONCLUSION

In this paper, service overlay network architecture was proposed as the platform for providing multiparty multimedia conferencing. MPCs were employed at strategic locations in the existing Internet transmission to combat the weaknesses of the Internet transmission to efficiently enable multiparty multimedia conferencing service. Key components in this architecture have been discussed in detail. In effect bandwidth allocation problem has been investigated and formulated and numerical results have been presented. The future work will focus on developing the related algorithms and protocols in this overlay service network to improve quality of service.

REFERENCES

Blake, S. D., Black, M., Carlson, M., Davies, E., Wang, Z. and Weiss, W. (1998). An architecture for differentiated services. IETFRFC 2475. Dec. 1998.

Braden, R., Clark, D. and Shenker, S. (1994). Integrated services in the Internet architecture: An overview. Internet RFC 1633. 1994.

Chawathe, Y., McCane, S. and Brewer, E. A. (2000). RMX: Reliable Multicast for Heterogeneous Networks. Proc. INFOCOM 2000:795–804.

Chu, S. R. Y. and Zhang, H. (2002). A case for end-system multicast, *IEEE Journal on selected areas in communications, 20(8): 1456 –1471.*

Deering S. and Cheriton, D. (1990). Multicast Routing in Datagram Internetworks and Extended LANs. *ACM Transactions on Computer Systems (TOCS), 8(2): 85–110.*

Gaofeng, Y. and Biing-Huang, J. (2005). An Overlay IP Network Architecture and Bandwidth Allocation for Multiparty Multimedia Conferencing. Proceedings of the 2005 International Conference on Internet Computing. ICOMP 2005, Las Vegas, Nevada, USA, June 27-30. 45-51.

Hossain, E. and Bhargava, V. K. (2001). Harmonic proportional bandwidth allocation and scheduling for service differentiation on streaming servers. *IEEE Journal on Selected Areas in Communications, 19(11): 2201 – 2214.*

Hui, J. (1988). Resource Allocation for Broadband Networks, *IEEE Journal on Selected Areas in Communications,6(9): 1598–1608.*

Jannotti, J., Gifford, D., Johnson K. L., Kaashoek, M. F. and Otoole, J. J. W. (2000). Overcast: Reliable Multicast with an Overlay Network, in Proc. 4th Symp. Operating System Design and Implementation (OSDI), October 2000.

Kashihara, S., Miyazawa, M., Ogaki, K. and Otani, T. (2009). Proposal of the Architecture of a QoS Assured Network by Cooperating between IP Flow Control and MPLS Diff-Serv-TE, IEEE International Conference on Communications, 2009. ICC '09: 1 – 6.

Kelly, F. (1996). Stochastic Networks: Theory and Applications, 1996. Notes on effective bandwidth.

Krunzz, M. M. and Makowski, A. M. (1998). Video Traffic Using M/G/∞ Input Processes: A compromise Between Markovian and LRD Models, *IEEE Journal on Selected Areas in*

Communications, 16(5): 733–748.

Rabby, M. and Ravindran, K. (2007). Dynamics of End-to-End Bandwidth Allocations in QoS-adaptive Data Connections, Local & Metropolitan Area Networks.LANMAN2007. 15th IEEE Workshop on Performance Evaluation of Bandwidth Allocation in 802.16j Mobile Multi-Hop Relay Networks.

Rekhter, Y. and Li, T. (1995). A Border Gateway Protocol 4 (BGP-4). March 1995, Internet RFC 1771.

Schulzrinne, H., Casner, S., Frederick, R. and Jacobson, V. (1996). RTP: A Transport Protocol for Real-Time Applications. IETFRFC 1889.

Serfozo, R. (2004). Notes on Markov Processes, Martingales, Brownian Motion and Stochastic Ordering. Springer.

Stoica, I., Morris, R., Karger, D., Kaashoek, M. and Balakrishnan, H. (2001). Chord: A scalable peer-to-peer look-up service for Internet applications. Proc. ACM SIGCOMM.

Subramanian, L., Stoica, I., Balakrishnam, H. and Katz, R. (2004). Over QoS offering Internet QoS using overlays. *ACM SIGCOMM Computer Communication, 33(1): 11 –16*

Weiss, A. (1995). An Introduction to Large Deviations for Communications Networks. *IEEE Journal on Selected Areas in Communications, 13(6): 938–952.*

Yao, L. and Xuan. C. (2009). Network Design and Architectures for Highly Dynamic Next-Generation IP-Over-Optical Long Distance Networks, Business video ready enterprise IP network architecture. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2009. BMSB '09.

Zhou, X., Xu, C. Z. (2004). Harmonic proportional bandwidth allocation and scheduling for service differentiation on streaming servers. *IEEE Transactions on Parallel and Distributed Systems, 2004, 15(9): 835 – 848.*