INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2000; **13**:79–96

Traffic analysis for multiparty videoconferencing in virtual path-based ATM networks

Gang Feng* and Tak-Shing Peter Yum

Department of Information Engineering, The Chinese University of Hong Kong, N.T., Shatin, Hong Kong

SUMMARY

With effective bandwidth concept encapsulating cell-level behaviour, asynchronous transfer mode (ATM) network design and analysis at the call-level may be formulated in the framework of circuit-switched loss networks. In this paper, we develop an analytical framework for a kind of multiparty videoconferencing in the VP-based ATM network at call-level. For this kind of conference, only the video of the current speaker is broadcast to other conferees. We first address several conference management issues in the VP-based ATM network, including the bandwidth allocation strategies, routing rule, call admission policy and speaker change management. Next, we formulate a traffic model for the conferences. Since an exact analysis of such a multiparty conference network is mathematically intractable, an approximate analysis for such conferences in a fully connected VP network is performed. The key of our method is to make use of the reduced-load approximation and open Jackson network model to derive the traffic loads from new conferences as well as that from the speaker change of the on-going conferences. Our study shows that the proposed analysis can give accurate predictions of the blocking probabilities for the new conference calls as well as video freeze probabilities for the on-going conferences. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: teleconferencing; traffic model; performance evaluation

1. INTRODUCTION

ATM is rapidly becoming the dominant transport technology for broadband services and real-time multimedia applications such as multiparty videoconferencing will undoubtedly be one of the most important applications. Traditional conferencing systems have been designed for deployment on the narrowband leased-line and circuit-switched networks. The advent of low-cost codecs has even brought conferencing to the desktop over telephone-networks. But ATM networks is a much better platform for videoconferencing.

Previous research on videoconferencing in ATM networks includes system architecture, system design,[1] traffic descriptors for VBR video,[2] call admission control,[3] and routing algorithms etc.[4–7] However, to the best of our knowledge, there is no paper on the analysis of multiparty videoconferencing in ATM network on the call level appeared in the literature.

Different kinds of videoconferences have different configurations, user-interactions, quality of service (QOS) requirements,[8–11] and network resource requirements. Selectable media, common media,[10] and virtual space conferences[9] are some examples of the proposed multiparty videoconferencing types. For these kinds of conferences, each conference has a central server called

* Correspondence to: Gang Feng, Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong.

CCC 1074–5351/2000/010079–18$17.50

Received May 1998
Accepted April 1999

conference bridge that performs the functions of collecting the video and audio signals from the conference sites, mixing the audio signals, composing the specific video signals and distributing the resulting video and audio signals to the individual conference sites.

In this paper we focus our study on a special type of conference called speaker-video conference. Speaker-video conference only requires the video of the current speaker be broadcast to all other conferees. The audio signals, on the other hand, are handled in the usual way, i.e., mixed and then sent back. Speaker-video conference has the advantage of demanding the least amount of equipment and bandwidth compared to those conference methods requiring a central server.

In speaker-video conferences, since the conference network under consideration allows statistical multiplexing of different conference traffic streams to share a link at the call level, the bandwidth on a link may not always be available when a video source(speaker) turns active. When that occurs, video freeze will be experienced by certain conferees. In this paper, The call blocking probability and the video freeze probability are chosen as the QOS measures for the speaker-video conferences at the call level.

The video bandwidth allocation problems of the speaker-video conferences were first studied in Reference 12. Traffic engineering aspects for this kind of conference were also studied in References 13 and 14 where the video freeze probability was obtained using a closed queuing network model. Based on this model, the capacity space of a conferencing network was derived and the call blocking probability was computed. However, this model is mathematically intractable for any reasonable size conferencing networks.

The analysis of multiparty videoconferencing in a general network is very difficult and no effective analytical technique is available. Because a conference typically requires the simultaneous possession of the bandwidth from several links, even without the alternate routing or call waiting (which we do not consider in this paper), the analysis is quite complicated. Kelly[15] investigated the blocking probability in the circtuit switched networks where each customer requests a fixed number of channels from a set of links. However, Kelly's model requires the enumeration of all the possible paths in the networks. Therefore, for any non-trivial network, this model cannot lead to numerical solutions.

Whitt[16] analysed a mathematical model of blocking system with a simultaneous resources possession. In his model, there are several multiserver service facilities at which several classes of customers arrive in independent Poisson processes. Each customer requests service from one server in each facility in a subset of the service facilities, with the subset depending on the customer class. Whitt concluded that although exact blocking probability expression for each customer class is available, it is very complicated to be useful. He henceforth proposed two kinds of upper bounds and a computing scheme based on an improved reduced-load approximation.

In this paper, we construct a traffic model for speaker-video conferences in networks which takes into account the dynamics of the conference arrival, conference departure, and the change of speakers in a conference. If we regard each group of links required by a conference call as a facility and the conference making such request as a customer, the queuing model can be formulated as follows. Let there be several classes of customers arriving in independent Poisson processes to several multiserver facilities. Each customer requests service from one server in each facility in a subset of the service facilities, with the subset depending on the customer class. If service can be provided immediately upon arrival at all the required facilities, then service begins and all servers assigned to the customer start together. Otherwise, the arrival is blocked and lost. When the service finishes, the customer requests service from one server in each facility in *another*

subset of the service facilities with a certain probability. The problem is to determine the blocking probability for each customer class. For this model, exact analysis is even more complicated than Whitt's model. On the other hand, if speaker change is not allowed, our model reduces to Whitt's model.

In Section 2, we first describe the VP network carrying the speaker-video conferences and then present the conference management issues including VP bandwidth reservation strategies, routing and call admission control policy. Section 3 presents the traffic model of speaker-video conference network. In our model, the VP-based ATM network carrying the speaker-video conferences is assumed to be an N-node fully connected network, as commonly assumed in the homogeneous VP-based ATM networks. This is the same assumption used in References 17 and 9 and was regarded as a reasonable assumption because the connectivity of the backbone ATM network is usually very high. In Section 4, we derive the offered traffic load of new and on-going conference calls, and make use of the 'reduced-load approximation' to derive the fixed-point equations. Section 5 gives the exact analysis for the speaker-video conferencing network. An exact evaluation of the link congestion probability is available, but it is shown to be complicated. In Section 6, we derive the conference level call blocking probability and video freeze probability. In Section 7, we study a 5-node network example whereby analytical and simulation results are provided and compared. Finally, Section 8 concludes this paper.

2. SPEAKER-VIDEO CONFERENCE IN VP-BASED ATM NETWORKS

2.1. VP network

An important feature of ATM networks is the virtual path (VP).[18] A VP is a direct logical connection between the two nodes with some assigned capacity. Typically, each node, while not physically connected to all the nodes in the network, can have an assigned VP to all the other nodes in the network. Each VP contains a bundle of virtual circuits (VCs) that are routed together as a unit. There are many advantages of using VP. For example, it can reduce call set-up delays, simplify node structure, segregate different types of traffic with different QOS requirements, reduce nodal processing and simplify routing. The price to pay is the decrease of statistical multiplexing gains among traffic sources from different VPs in the same physical link.

A logical VP network can be defined by viewing the VP endpoints as nodes and VPs as links. For instance, consider the physical network shown in Figure 1(a), which consists of four switches (nodes). If a VP is assigned to each node pair, as shown in Figure 1(b), an N-node fully connected to VP network is formed consisting of $N(N-1)$ unidirectional links (VPs).

We assume that the topology of the VP subnetwork remains fixed for purposes of routing VCs, as the time scale of interest is significantly smaller than those involved with the dynamics of VP subnetwork.[19] From now on, we refer to the VP subnetwork as a *conference network* and a VP as a *link*. Then the network under consideration can be viewed as a graph with ATM switches as nodes and VPs as links.

2.2. Bandwidth allocation

In an ATM network, traffic can be considered at the VP level, the call level, the burst level or the cell level.[20] In videoconferencing network, various multicast connections may be established by either VCs or VPs on an ATM network. In this paper, we assume that the speaker-video

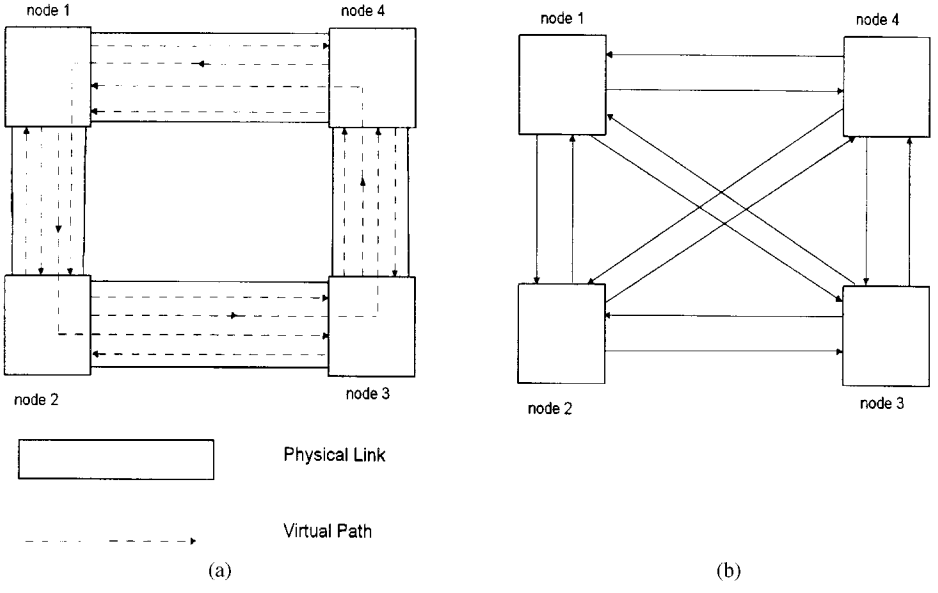

Figure 1. A 4-node network and its corresponding VP network. (a) Physical network. (b) Virtual path network

conferences are established through setting up VPs. A route-configuration for a conference can thus be realized by setting up a VC multicast tree on the VP network. To construct a VP based conference network, an essential problem is the bandwidth allocation of VPs. Generally, there are two VP bandwidth reservation strategies: *deterministic strategy*[6] and *statistical strategy*.[18]

The deterministic strategy[6] reserves separate link capacity for each VP passing through the link. A certain amount of buffer space at the source node and bandwidth at each physical link on this VP are dedicated to this VP to ensure the QOS of videoconferencing at the cell level, such as cell loss rate, maximum and mean end-to-end cell transfer delay, cell delay variation, etc. Since there is no statistical multiplexing among the VPs at the link-level, the sum of the reserved VP bandwidth allocations on a link is not permitted to exceed the total capacity of this link.

The statistical strategy[18] allows statistical multiplexing of cells from different VPs onto a physical link. In this approach, instead of explicitly reserving link bandwidth for each VP, each VP is allocated a dedicated number of virtual circuits, each with a guaranteed QOS. Then the call setup processing can still be done so long as the number of existing virtual circuits within the VP is less than this number. The actual reservation is made as if each VP has accepted as many as the dedicated number of virtual circuits into the network. Statistical strategy can provide higher cell multiplexing gain than the deterministic strategy. However, this advantage is offset by the fact that when some connections are routed over multiple VPs a more stringent QOS guarantee needs to be provided by each component VP.

Deterministic or statistical multiplexing can both be used for capacity reservation at all the levels of ATM networks. For example, consider a network that does not use any virtual path connection (VPC); then all the traffic can be statistically multiplexed on virtual circuit connections (VCC) sharing common network links. Although this may result in a higher bandwidth utilization, the VCC call establishment cost could be significant because each VCC would need to

negotiate a connection request at each intermediate node between the source and destination. On the other hand, if peak rates are allocated at each level, the network resources would be poorly utilized.

When capacity is reserved on each VP, it may be desirable to dynamically adjust the allocation to improve link bandwidth utilization as well as to adapt to dynamic changes in network traffic flows. In this paper, we argue that the re-allocation of VP capacity should be done periodically on a much longer time scale than the inter-arrival time of successive calls. Furthermore, we assume that the time interval between the two VP capacity re-allocations is significantly larger than a typical call duration. That is, we assume that the capacity of each VP in the videoconferencing network is constant. Engineering on the VP network level can hide the stochastic behavior of the cell level and this is essential for the tractability of the design problem. The VP network can now be seen as offering conferencing service as a circuit-switched loss network.

2.3. Routing

Let there be a *conference bridge* which performs the functions of routing, admission control, and the management of the change of active nodes. When a new conference is initiated or when there is a change of active node in an on-going conference, a conference management process is created. The conference bridge collects information such as the number of conferees, their location, and their busy/idle status, etc., and tries to set up a VC route-configuration in the VP network.

To make the routing and admission control algorithm simple, only routes (VCs) consisting of a single VP (direct) are tried. In other words, shortest path routing is used. For example, in the VP network shown in Figure 1(a), if we let node 1, 2 and 3 be the conference nodes and node 1 has the current speaker attached. Then, the obtained route-configuration under the shortest path routing rule is shown in Figure 2(a). When a conferee at node 2 becomes the next speaker, the route-configuration is changed to that in Figure 2(b).

Of course, if a direct link for a connection request is not available, an alternate path consisting of multiple VPs can be used, just like that in the least-loaded type of routings.[7,17] As the performance evaluation under dynamic routing rules is beyond the scope of this paper, we restrict our analysis to the fixed routing case.

2.4. Call admission

When a new call arrives, the conference bridge should carry out a call admission policy. For the ease of network management and control, we propose the following admission policy:

Identify a route-configuration in the VP network according to the shortest path routing rule and check the route-configuration whether a VC with guaranteed QOS can be established on all the involved VPs. If 'yes' accept the call; otherwise, reject the call.

2.5. Speaker change management

In the speaker-video conference network, the network resources should be dynamically allocated and released in response to change of speakers throughout the conference session. The conference bridge needs to perform a management process when there is a change of speaker. Specifically, if the next speaker is attached to the same node, the conference bridge only needs to switch the video source from the former speaker to the next one and keep the existing

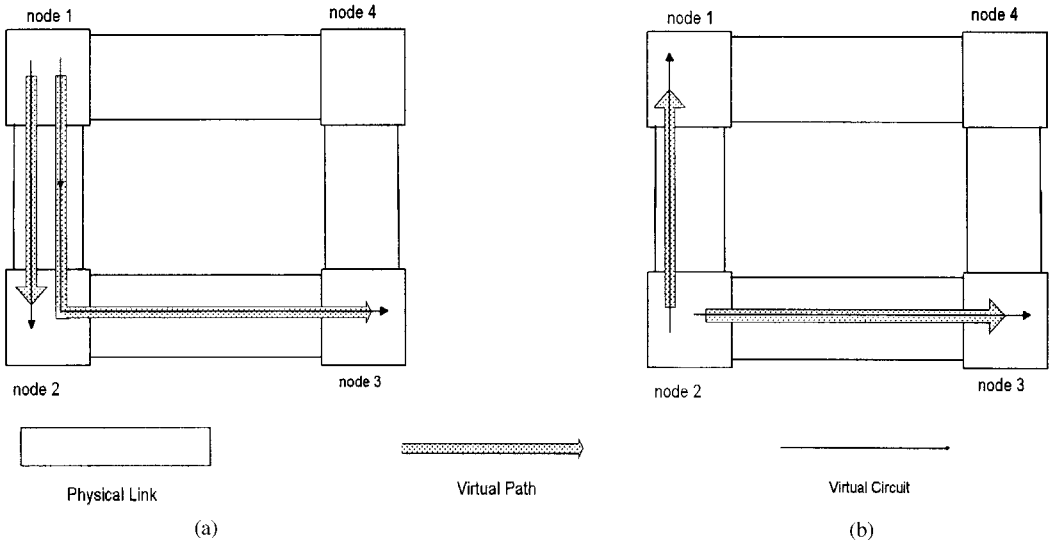

Figure 2. Route-configuration for a conference with 3-conference nodes in VP-based conference network. (a) Node 1 transmitting. (b) Node 2 transmitting

route-configuration in the VP network. Otherwise, a new route-configuration of VCs with guaranteed QOS on all the involved VPs is identified and established and the VCs in the former route-configuration are released. If there is a temporary shortage of bandwidth, the conference will experience a temporary video freeze. For the mathematical tractability, we will not consider the case with dynamic joining and withdrawing of conferees during a conference session in the analysis.

3. CONFERENCE TRAFFIC MODEL IN VP NETWORK

In the last section, we introduced a framework for network design and management for the videoconferencing in an ATM-based backbone network based on VP. Next, we need to relate the network capacity to the traffic and the quality of service (QOS) quantitatively. For simplicity, we do it on an N-node fully connected network under the shortest path routing.

Let the capacity of links be characterized in terms of the number of video channels it can support and let c_i denote the capacity of link ℓ_i. Moreover, let $\ell_{(i,j)}$ denote the specific link from node i to node j and let $E = N(N-1)$ be the total number of links in the fully connected network.

We assume that all the videos are transmitted in the same format and the voice traffic is unrestricted. Hence in the case of a video freeze event, conference activities are not affected, and only the conference quality is affected by the delayed onset of video when there is a change of speaker. Nodes that have at least one conferee attached are called *conference nodes*. Among the conference nodes, we define the *active node* as the node which has the current speaker attached. A conference spanning K conference nodes has K *modes of operation* where each mode corresponds to one of the conference nodes being active. The mode determines the route-configuration. Links currently used by an on-going conference are called active links.

Let S_0 be the maximum number of conferees allowed in a conference. Then, the conference calls can be classified into $S_0 - 1$ types, where a type s ($s = 2, 3, \ldots, S_0$) call is a conference call with s conferees.

Let b_i be the total numbers of conference subscribers at node i and let all conferees have equal community interest on all the others. Then the probability q_i that a conferee is located at node i is $q_i = b_i / \sum_{j=1}^{N} b_j$. Let the arrivals of each type of conference calls be a Poisson process and let γ_s ($s = 2, 3, \ldots, S_0$) be the arrive rate of the s-party calls. Speech duration of any speaker is assumed to be exponentially distributed with mean $1/\mu$. When a conferee finishes speaking, the conference ends and leaves the network immediately with probability p_e and continues with probability $1 - p_e$. If the conference continues, the new speaker is equally likely to be any one of the other conferees. It was shown in Reference 12 that the duration of a conference call in the network is a geometric sum of the independent and identically distributed exponential random variables and is therefore exponentially distributed with mean $(\mu p_e)^{-1}$.

4. ANALYSIS

In this section, we make use of the link independence assumption and the reduced-load approximation[21] to derive the blocking probability and video freeze probability. The offered traffic to each VP is assumed to be Poisson, with the rate reduced suitably to account for blocking. The approximation leads to a set of fixed-point equations which can be solved recursively. We start with the derivation of the offered loads from the new and on-going calls.

4.1. A new call traffic

A conference requires a channel on each of the active links simultaneously. Therefore, we can decompose an s-party call into a set of channel requests on the set of active links. This decomposition is allowed because of the two properties of Poisson processes: (1) independent random splitting of a Poisson process produces independent Poisson processes, and (2) the superposition of independent Poisson processes is a Poisson process.

For an s-party conference, let $\mathbf{K}_s \equiv (K_1, K_2, \ldots, K_N)$ be the conferee distribution with K_i being the number of conferees located at node i and $K_1 + K_2 + \cdots + K_N = s$. Let $\mathbf{k} = (k_1, k_2, \ldots, k_N)$ where the k_i's are non-negative integers. Under the assumption that all the conferees have equal community interest to all the others, we have

$$\text{Prob}[\mathbf{K}_s = \mathbf{k}] = \begin{cases} \binom{s}{k_1, k_2, \ldots, k_N} \prod_{i=1}^{N} q_i^{k_i} & \text{for all } \mathbf{k} \text{ with } \sum_{j=1}^{N} k_j = s \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Let r be the link number of $\ell_{(i,j)}$ and let $\beta_s(\ell_r | \mathbf{k})$ be the probability that an s-party call with the conferee distribution $\mathbf{K}_s = \mathbf{k}$ *is admissible and requires a channel on link* ℓ_r. Moreover, let B_r be the congestion probability on link ℓ_r. Under the shortest path routing rule, a channel is required on each of the links from the active node to all the other conference nodes. Let C_i be the set of links used by the conference under consideration when the current speaker is located at node i. Then,

with the link independence assumption,[22] we have

$$\begin{aligned}\beta_s(\ell_{(i,j)}|\mathbf{k}) &= \text{Prob}\,[\text{node } i \text{ is the source node}]\cdot \text{Prob}\,[\text{node } j \text{ is a conference node}]\\ &\quad\times \text{Prob}\,[\text{links in set } C_i\backslash\ell_r \text{ are all non-blocking}]\\ &= \frac{k_i}{s}u(k_j)\left[\prod_{m\in C_i\backslash r}(1-B_m)\right] \quad \forall i,\, j\,;\, i\neq j \end{aligned} \tag{2}$$

where the term $[\cdot]$ indicates the 'load reduction' factor and $u(\cdot)$ is the unit step function.

By removing the conditioning on the conferee distribution $\mathbf{K}_s$, the probability that an s-party call can be admitted and it requests a channel on link ℓ_r, denoted as $\beta_s(\ell_r)$, is given by

$$\beta_s(\ell_r) = \sum_{\mathbf{k}\in\Omega_s}\beta_s(\ell_{(i,j)}|\mathbf{k})\,\text{Prob}\,[\mathbf{K}_s=\mathbf{k}], \quad r=1,2,\ldots,E \tag{3}$$

where

$$\Omega_s = \left\{\mathbf{k}\,\middle|\,\sum_{i=1}^{N}k_i = s\right\}$$

The '*reduced*' traffic load contribution to link ℓ_r, denoted by $\lambda_s(\ell_r)$ is therefore,

$$\begin{aligned}\lambda_s(\ell_r) &= \gamma_s\beta_s(\ell_r)\\ &= \gamma_s\sum_{\mathbf{k}\in\Omega_s}\left[\frac{k_i}{s}u(k_j)\left(\prod_{m\in C_i\backslash\ell_r}(1-B_m)\right)\text{Prob}\,[\mathbf{K}_s=\mathbf{k}]\right], \quad r=1,2,\ldots E\end{aligned} \tag{4}$$

Figure 3 illustrates this decomposition.

4.2. *Traffic from on-going conferences*

Under the assumption that the next speaker is equally likely to be any one of the other conferees, when the current speaker of a conference located at node i finishes speaking, the next speaker will be the conferee at the *same* node with probability $((K_i-1)/(s-1))$ and the conferee at the other nodes with the remaining probability. The channel holding time Y of link ij is thus a geometric sum of the exponential random variables and can be shown to be exponentially distributed with the mean $(s-1)/(s-K_i)\,\mu$.[12]

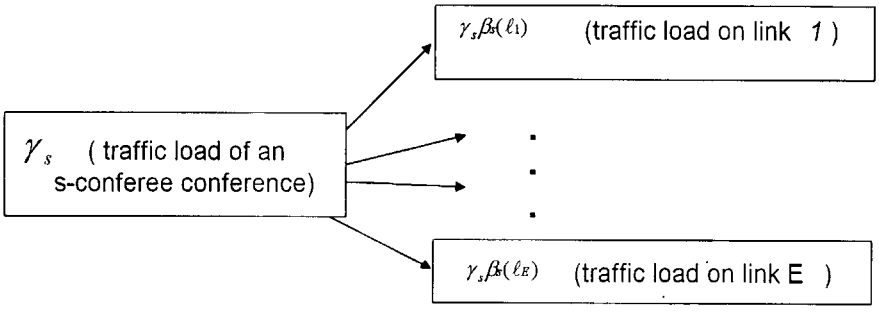

Figure 3. Decomposition of traffic load of an s-party conference on each link in the network

Let h_{ij} be the probability that a conferee at node i finishes speaking and the next speaker is located at node j. under $\mathbf{K}_s = \mathbf{k}$, we have

$$h_{ij}(\mathbf{k}) = \text{Prob}\,[\text{conference continues}]\,\text{Prob}\,[\text{the next speaker is located at node } j\,]$$

$$= \begin{cases} (1-p_e)\dfrac{k_j}{s-1}, & i \neq j;\ k_i \geqslant 1; k_j \geqslant 1 \\ 0 & k_i = 0 \text{ or } k_j = 0 \end{cases} \tag{5}$$

If the next speaker is located at the same node, we have

$$h_{ij}(\mathbf{k}) = \text{Prob}\,[\text{conference continues}]\,\text{Prob}\,[\text{the next speaker is located at node } j\,]$$

$$= \begin{cases} (1-p_e)\dfrac{k_i}{s-1}, & k_i \geqslant 2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

For a specific on-going conference with conferee distribution $\mathbf{k}$, let C_i and C_m be the link sets used by this conference in operation mode i and m, respectively. When the operation mode changes from i to m, the traffic load on the links in C_i is transferred to the links in C_m. Figure 4 shows an example of a conference with $\mathbf{K}_4 = (2, 1, 1)$ in a three-node network changing from modes 1 and 3.

Without loss of generality, let us consider the transfer of traffic load from link ℓ_1 which is the link from node i to node j ($\ell_{(i,j)}$) to link ℓ_2 which is the link from node m to node n ($\ell_{(m,n)}$). Let $T_s(\ell_2|\mathbf{k}, \ell_1,)$ be the event that *under conferee distribution* $\mathbf{k}$ *an s-party conference using link* ℓ_1 (*among others*) *uses link* ℓ_2 (*also among others*) *after a mode change*. Under the shortest path routing rule, for different operation modes, the corresponding sets of links being used are exclusive to each other. Thus we have

$$\text{Prob}\,[\,T_s(\ell_2|\ell_1, \mathbf{k})] = h_{im}(\mathbf{k})$$

Removing the conditioning on $\mathbf{k}$, we have

$$\text{Prob}\,[T_s(\ell_2|\ell_1)] = \sum_{\mathbf{k} \in \Omega_s'} \text{Prob}\,[T_s(\ell_2|\ell_1, \mathbf{k})]\,\text{Prob}\,[\mathbf{K}_s = \mathbf{k}|\ell_1]$$

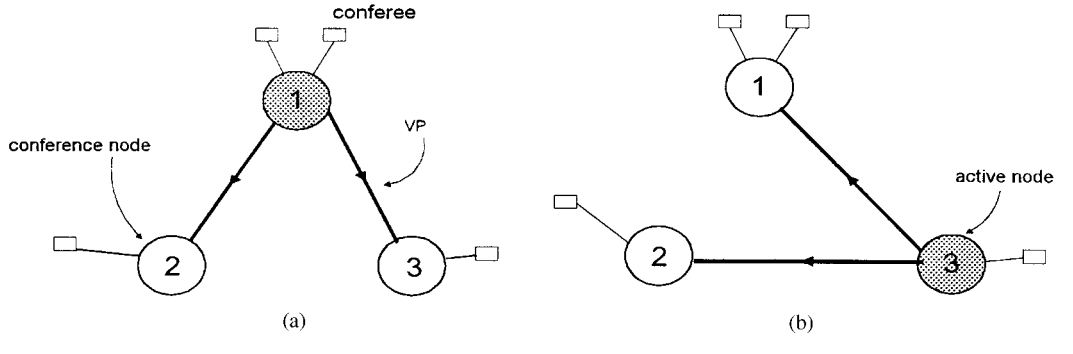

Figure 4. Link set used by a 3-node conference. (a) Mode 1: $C_1 = \{\ell_{(1,2)}, \ell_{(1,3)}\}$. (b) Mode 3: $C_3 = \{\ell_{(3,1)}, \ell_{(3,2)}\}$

where

$$\Omega'_s = \left\{ \mathbf{k} \,\middle|\, \sum_{p=1}^{N} k_p = s \quad \text{and} \quad k_i \geqslant 1;\, k_j \geqslant 1 \right\}$$

and

$$\begin{aligned} &\text{Prob}[\mathbf{K}_s = \mathbf{k} | \ell_1] \\ &\quad = \text{Prob}[\mathbf{K}_s = \mathbf{k} | \ell_{(i,j)} \text{ is used before the mode change}] \\ &\quad = \binom{s-2}{k_1, k_2, \ldots, k_i - 1, \ldots, k_j - 1, \ldots, k_N} (q_i)^{k_i - 1} (q_j)^{k_j - 1} \prod_{\substack{p=1 \\ p \neq i,j}}^{N} (q_p)^{k_p} \end{aligned}$$

As ℓ_2 can receive contributions from the links other than ℓ_1, the relative portion that comes from ℓ_1, which we denote as the traffic transition probability $f_s(\ell_2 | \ell_1$, conference continues) is simply

$$f_s(\ell_2 | \ell_1, \text{ conference continues}) = \frac{\text{Prob}[T_s(\ell_2 | \ell_1)]}{\sum_{r=1}^{E} \text{Prob}[T_s(\ell_r | \ell_1)]}$$

But the conference can end with probability p_e. If so, $f_s(\ell_2 | \ell_1$, conference ends) = 0. Therefore,

$$f_s(\ell_2 | \ell_1) = (1 - p_e) \frac{\text{Prob}[T_s(\ell_2 | \ell_1)]}{\sum_{r=1}^{E} \text{Prob}[T_s(\ell_r | \ell_1)]} \tag{7}$$

As a check, $\sum_{i=1}^{E} f_s(\ell_i | \ell_1) + p_e = 1$ as it should.

Continue the example in Figure 4, to derive the transition probability from $\ell_{(1,2)}$ to $\ell_{(2,3)}$, the set of conferee distributions to be considered is $\Omega'_4 = \{(2, 1, 1), (1, 2, 1), (1, 1, 2)\}$. Similarly, for transition from $\ell_{(1,2)}$ $\ell_{(1,2)}$, $\Omega'_4 = \{(2, 1, 1), (2, 2, 0), (3, 1, 0)\}$.

Let $\Lambda_s(\ell_i)$ denote the total traffic rate (including the new and the on-going conferences) from all type s conferences on link ℓ_i. Then the transferred traffic load from ℓ_1 to ℓ_2 has the rate $\Lambda_s(\ell_1) f_s(\ell_2 | \ell_1)$.

By adding the external and internal traffic contributions to the specific link ℓ_i, we have

$$\Lambda_s(\ell_i) = \lambda_s(\ell_i) + \sum_{j=1}^{E} \Lambda_s(\ell_i) f_s(\ell_i | \ell_j), \quad i = 1, 2, \ldots, E \tag{8}$$

where $\lambda_s(\ell_i)$ is given by (4). This set of equations can be solved by the Guassian elimination or Guass–Sidel iteration.

The combine 'new call arrival plus mode change' process on link ℓ_i has the rate $\Lambda(\ell_i)$ given by

$$\Lambda(\ell) = \sum_{s=2}^{S_0} \Lambda_s(\ell) \tag{9}$$

Let L_i be the occupancy at link i and let $p(\mathbf{n}) = \text{Prob}[L_1 = n_1, L_2 = n_2, \ldots, L_E = n_E]$. Under the link independence assumption, the above becomes a Jackson open queuing network and its steady-state solution [7] is

$$p(\mathbf{n}) = \prod_{i=1}^{E} \left(\frac{\Lambda(\ell_i)}{\mu} \right)^{n_i} \frac{p(\mathbf{0})}{n_i!} \tag{10}$$

where

$$p(\mathbf{0}) = \left[\sum_{\mathbf{n} \in \Theta} \left(\prod_{i=1}^{E} \frac{(\Lambda(\ell_i)/\mu)^{n_i}}{n_i!} \right) \right]^{-1} \quad \text{and} \quad \Theta \equiv \{\mathbf{n} | 0 \leqslant n_i \leqslant c_i, \quad i = 1, 2, \ldots, E\}$$

The link congestion probability can be obtained from (10) by appropriately summing over the states of all the other links:

$$B_i = \sum_{\mathbf{n} \in \Theta_i} p(\mathbf{n}), \quad i = 1, 2, \ldots, E \tag{11}$$

where $\Theta_i \equiv \{\mathbf{n} | n_i = c_i\}$.

4.3. Fixed-point equations

Equations (4) and (11) are of the forms $\tilde{\Lambda} = \Psi(\tilde{B})$ and $\tilde{B} = \Phi(\tilde{\Lambda})$, respectively, where $\tilde{\Lambda} = \{\Lambda_i\}_{i=1,2,\ldots,E}$ and $\tilde{B} = \{B_i\}_{i=1,2,\ldots,E}$. The method of successive approximation is typically effective for their solutions. This method starts with a certain initial values $\tilde{\Lambda}^{(0)}$ and iterates,

$$\begin{aligned} \tilde{B}^{(k+1)} &= \Phi(\tilde{\Lambda}^{(k)}) \\ \tilde{\Lambda}^{(k+1)} &= \Psi(\tilde{B}^{(k+1)}) \end{aligned} \tag{12}$$

until a convergence criterion is satisfied.

4.4. Numerical solution

In the above analytical framework, we need to compute the traffic rates for new calls and the ongoing conferences for each of the s types of conferences. $N(N-1)$ traffic rates on all the links are computed by iteratively solving the fixed-point equations of (12). In each iteration, the running time is dominated by the computation of equation (4) which includes a summation of many terms. For a pair of specific s and N values, the number of terms is the same as that listed in Table I. We can see that if the maximum conference size $S_0 \leqslant 7$ and $N \leqslant 9$, the number of summation terms is limited to a few thousands. In addition, an efficient recursive algorithm[23] can be employed to compute the marginal probabilities in equation (11). The number of iterations needed to satisfy the convergence criterion of the fixed-point equations is generally between 10 and 20. Compared to the exact analysis for speaker-video conferences in References 13 and 14, the present method can solve a conference network of much larger size.

Table I. Number of Jackson queuing networks to be solved

N	4	5	6	7	8	9	10
$S_0 = 3$	20	35	56	84	120	165	220
$S_0 = 5$	56	126	252	462	792	1287	2002
$S_0 = 7$	120	330	792	1716	3432	6435	11 440
$S_0 = 9$	220	715	2002	5005	11 440	24 310	48 620

5. THE EXACT ANALYSIS

In this subsection, we generalize Whitt's model so as to capture the speaker change dynamics in the videoconferences and present the approach to compute the exact link congestion probability. As before, let L_j represent the occupancy of link j. Using theorem 4 in Reference 16, we can show that the distribution of $(L_1, L_2, \ldots, L_E\}$ can be described in terms of the random vector $(L_1^4, L_2^4, \ldots, L_E^4\}$, where L_j^∞ represents the occupancy of link j when all the links have infinitely many channels. In the infinite-sever model the steady-state distribution is easy to derive because there is no blocking, so there is no interaction among the conferences. Therefore, we shall assume all links to have infinite capacity in the following.

In the conferencing network, we regard each group of links being used simultaneously under a specific operation mode as a 'facility' and the conference making such request as a 'customer'. Then, along with the changes of modes, the customer changes the facilities it uses accordingly. For a specific conference with conferee distribution $\mathbf{k}$, the conference network can be formulated as a Jackson queuing network. Because the conferences are truly independent (as contrast to 'assumed independent' in the last section), we can derive the offered load to a particular link of the network by adding the load from all the conferences using it. Referring to the example in Figure 4, we can find the corresponding transition diagram in Figure 5. Here, state i refers to the situation that the group i links are used. When a customer received his service at facility i (group i), he moves on to use facility j with probability $t_{ij}(\mathbf{k})$ $(i, j = 1, 2, 3)$ where $t_{ij}(\mathbf{k})$ is the traffic transition probability under the conferee distribution $\mathbf{k}$ which can be obtained using (5) and (6).

Now consider the general case. Let $y(\mathbf{k}) = \sum_{j=1}^{N} u(k_i)$ be the number of conference node for s-party conference with a conferee distribution $\mathbf{k}$. Then there are y link groups in our model. Let $\Lambda_s(i|\mathbf{k})$ denote the total traffic rate to each link in group i from this conference. For this Jackson queuing network, the flow balance equations says

$$\Lambda_s(i|\mathbf{k}) = \lambda_s(i|\mathbf{k}) + \sum_{j=1}^{y(\mathbf{k})} t_{ji}(\mathbf{k})\Lambda_s(j|\mathbf{k}), \quad i = 1, 2, \ldots, y \tag{13}$$

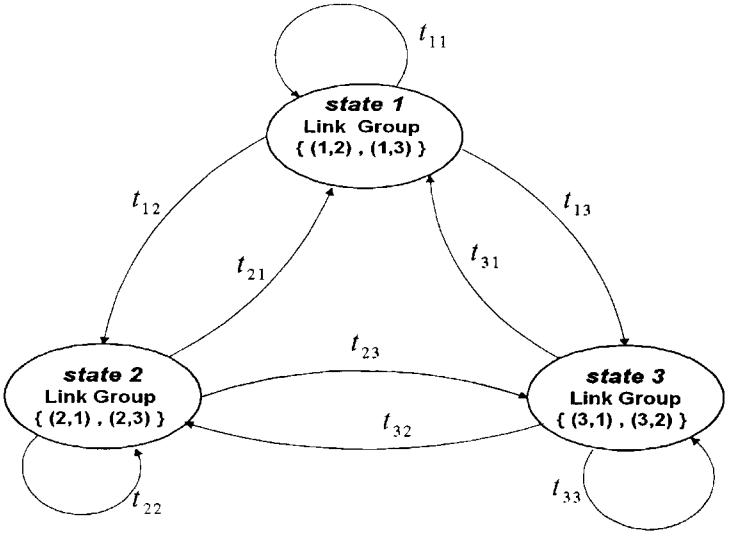

Figure 5. State transition diagram

where $\lambda_s(i|\mathbf{k})$ is the arrival rate of 'external' s-party calls which is similar to that in (4) *except* that there is no 'load reduction' factor in (2). This set of equations can be solved in the usual manner.

To determine the traffic rate from each conference type to a specific link, we need to consider all the possible conferee distributions of this type. Let $X(\ell_r) \subset \Omega_s$ be the set of all conferee distributions using link ℓ_r. Then the total arrival rate to link ℓ_r from the s-party conferences, denoted as $\Lambda_s(\ell_r)$, can be obtained as

$$\Lambda_s(\ell_r) = \sum_{\mathbf{k} \in X(\ell)} \Lambda_s(i|\mathbf{k}) \text{ Prob}[\mathbf{K}_s = \mathbf{k}], \quad r = 1,2, \ldots, E$$

where ℓ_r belongs to the link group i of the respective conferee distribution $\mathbf{k}$ in each of the summation term.

We then can obtain the total traffic rate on link j by summing up the contributions from all types of calls: $\Lambda(j) = \sum_{s=2}^{S_0} \Lambda_s(j)$. Thus, the steady-state distribution is given by

$$\text{Prob}[L_j^{\infty} = n_j, 1 \leqslant j \leqslant E] = \prod_{j=1}^{E} \left(\frac{\Lambda(j)}{\mu}\right)^{n_j} \frac{1}{n_j!}$$

Using Theorem 4 and Corollary 4.2 in Reference 16, we can obtain $\text{Prob}[L_j = n_j, 1 \leqslant j \leqslant E]$, in terms of $\text{Prob}[L_j^{\infty} = n_j, 1 \leqslant j \leqslant E]$ and hence the congestion probability of each link.

The complexity of the exact blocking probability for Whitt's model is high.[16] With the inclusion of the dynamics of speaker change in our model, the complexity increases as here we need to solve the open Jackson queuing network many times, once for each conferee distribution. Table I shows the number of Jackson queuing networks to be solved for some typical N (number of network nodes) and S_0 (maximum conference size) values.

6. CONFERENCE BLOCKING PROBABILITY AND VIDEO FREEZE PROBABILITY

In this section, we compute the conference level call blocking probability and video freeze probability.

6.1. *Call blocking probability*

When a conference call arrives, the conference bridge performs the admission control function. According to the admission policy, when any link in the route-configuration is in congestion, this conference call is blocked. Let $B(s|m, \mathbf{k})$ be the call blocking probability of an s-party conference, given that the conference has conferee distribution $\mathbf{K}_s = \mathbf{k}$ and the speaker is located at node m.

With the link independence assumption, $B(s|m, \mathbf{k})$ can be computed as

$$B(s|m, \mathbf{k}) = 1 - \prod_{i \in C_m} [1 - B_i]$$

where B_i is the probability that link ℓ_i is in congestion which is given by (11), and C_m is the set of links used by the conference in mode m.

By removing the conditioning on $\mathbf{K}_s = \mathbf{k}$ and m, the probability that the call blocking probability for an s-party conference, denoted as $B(s)$, is simply given by

$$B(s) = \sum_{\mathbf{k} \in \Omega_s} \left[\sum_{j \in I(\mathbf{k})} \frac{k_j}{s} B(s|j, \mathbf{k}) \right] \text{Prob}[\mathbf{K}(s) = \mathbf{k}]$$

Where $I(\mathbf{k})$ is the set of conference nodes, i.e. $I(\mathbf{k}) = \{i | k_i > 0\}$; $\text{Prob}[\mathbf{K}(s) = \mathbf{k}]$ is given by (1) and Ω_s is given by (3).

6.2. *Video freeze probability*

For an on-going conference, when there is a change of operation mode, a new route-configuration should be established. When a link in the route-configuration is in congestion, the conferee(s) attached to the affected node will experience a period of video freeze. We now define two video freeze measures as follows:

1. $F(s)$: Video freeze probability when there is a change of speaker:
   Let the s-party conference has a conferee distribution $\mathbf{k}$ at mode m. If the next speaker is located at the same node as the current speaker, the conference will not experience video freeze. But if the next speaker is located elsewhere, we have

$$F(s|m, \mathbf{k}) = \text{Prob}\{\text{next speaker at node } i\}\, \text{Prob}\{\text{blocking}|i, \mathbf{k}\}$$

$$= \sum_{\substack{i=1 \\ i \neq m}}^{N} \frac{k_i}{s-1} B(s|i, \mathbf{k})$$

   For a particular conference with conferee distribution $\mathbf{K}_s = \mathbf{k}$, the probability that the current speaker is located at node m is k_m/s. Therefore, unconditioning on $\mathbf{K}_s = \mathbf{k}$ and m, the video freeze probability of an *s-party* conference when there is a change of speaker, denoted as $F(s)$, can be obtained as

$$F(s) = \sum_{\mathbf{k} \in \Omega_s} \left[\sum_{j=1}^{N} \frac{k_j}{s} F(s|j, \mathbf{k}) \right] \text{Prob}[\mathbf{K}(s) = \mathbf{k}]$$

   where $\text{Prob}[\mathbf{K}(s) = \mathbf{k}]$ is given by (1) and Ω_s is given by (3).
2. $F_c(s)$: Probability of video freeze during a conference: The probability that *an s-party conference experiences video freeze during the entire conference session* $F_c(s)$ is another QOS measure of interest for the speaker-video conferencing service. It is given by

$$F_c(s) = \sum_{i=1}^{\infty} p_e (1 - p_e)^i (1 - [1 - F(s)]^i)$$

   where p_e is the exit probability of a conference and $p_e(1 - p_e)^i$ gives the probability that there are total i speech sessions in the conference.

7. NUMERICAL EXPERIMENT

We check the accuracy and range of application of the proposed analytical framework via a hypothetical 5-node fully connected VP network as shown in Figure 6. Each VP is assigned a dedicated bandwidth in terms of the number of conferencing video channels, indicated besides each link. For simplicity, the network has $c_{ij} = c_{ji}$ in this example. However, it should be noted that our analysis framework is valid for the network with $c_{ij} \neq c_{ji}$.

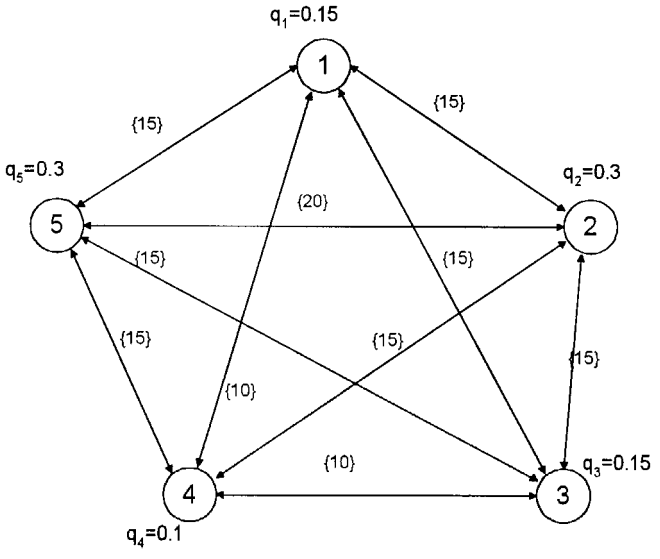

Figure 6. Five-node VP network carrying speaker-video conferences

Let there be four types of conference calls, corresponding to 2-, 3-, 4- and 5-party conferences. We let the *base traffic intensity* be $(\gamma_2, \gamma_3, \gamma_4, \gamma_5) = (2{\cdot}5, 5{\cdot}0, 5{\cdot}0, 2{\cdot}5)$ and we assume that conference exit probability $p_e = 0{\cdot}3$. The mean speech duration time is normalized to one time unit, i.e. $1/\mu = 1$.

The call blocking probabilities and the video freeze probabilities for each type of conference are computed and checked by a computer simulation. The convergence criterion adopted for the fixed-point equations in (12) is $((|\Lambda_i^{(k+1)} - \Lambda_i^{(k)}|)/\Lambda_i^{(k+1)}) < 10^{-7}$ for all i. For all the simulation results shown in the following figures, enough number of simulation runs were performed to make the 95% confidence intervals smaller than the size of the markers.

Figure 7 shows the blocking probability as a function of load increase over that of the base load. We can see that: (1) Under all traffic conditions, the blocking probability obtained by simulation is upper bounded by the analytical results. (2) There is no significant difference between the results at $B \leqslant 0{\cdot}02$.

Figure 8 shows the same for the video freeze probability. Again, we see a very tight upper bound provided by the analytical result.

For the above example of a 5-node network, the compiled C program running on a DEC Alpha workstation requires about 1 s to compute one probability value. As a comparison, computer simulation on the same machine takes about 8 min for each point to meet the confidence interval requirement. We also test the analysis method on a larger size networks. The computing times are about 2, 5, 14, 37 and 98 s for 6, 7, 8, 9 and 10 node networks, respectively.

8. CONCLUSION

The exact performance evaluation for multiparty videoconferencing network is mathematically intractable. In this paper, we have developed an analytical framework for the speaker-video

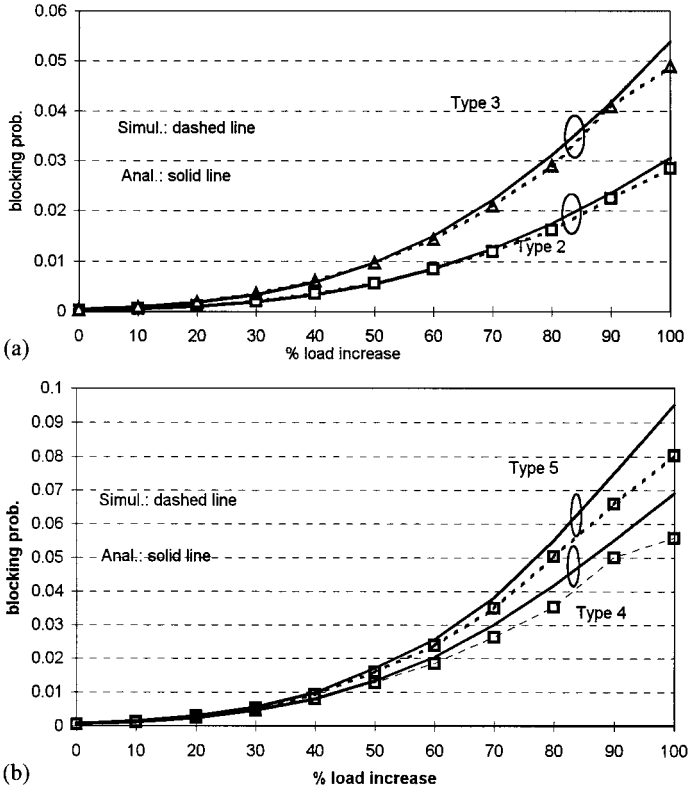

Figure 7. Call blocking versus load increase. (a) Types 2 and 3 conference call. (b) Types 4 and 5 conference call

videoconferencing in the VP based ATM network. The key of our approach is to make use of the reduced-load approximation and open Jackson network model to derive the traffic loads from the new conferences as well as that from the speaker change of the on-going conferences. Further, a set of fixed-point equations is derived for computing the call blocking probability and video freeze probability. The numerical results show that this analytical framework is accurate and can be used to evaluate the performance of a large size conference network.

Kelly[15] and Whitt[16] have proved that the reduced-load system has an unique solution and the reduced-load approximation is asymptotically correct even in heavy traffic. Our numerical results show that these two properties hold also for our model. Further generalization to multirate case for conferences with varying levels of qualities appears to be possible. It will be however be a topic for further investigation.

In this paper, only the speaker-only videoconferencing has been studied. The traffic models for other types of conferencing such as common-media conferencing, selectable-media conferencing are quite different with that of the speaker-only conferencing, as the traffic flows among the conferencing network do not change in the way of the speaker-only conferencing. This could lead to more variations of queuing models. The analysis for these queuing models is an interesting topic for further study.

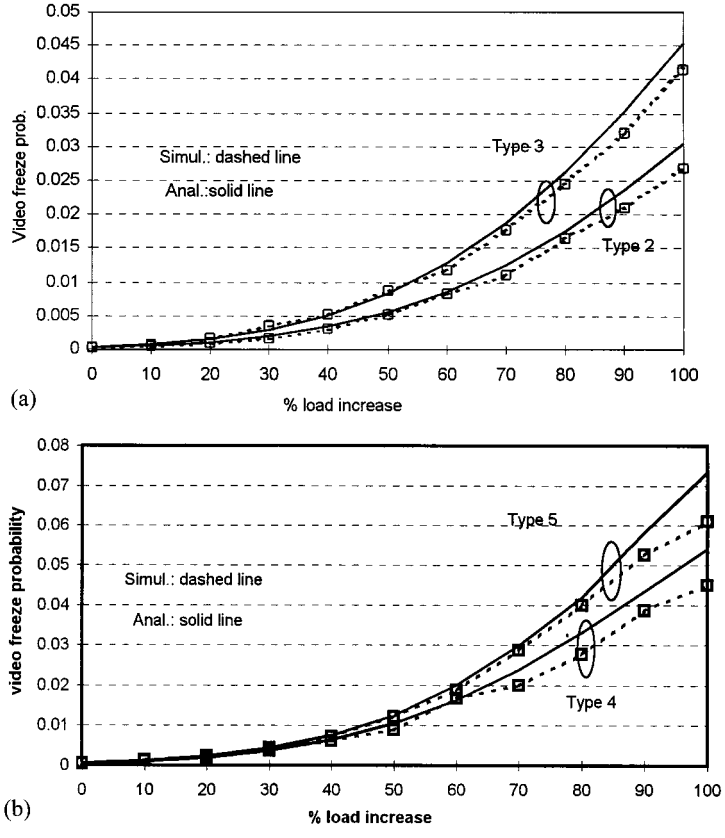

Figure 8. Video freeze probability versus load increase. (a) Types 2 and 3 conferences. (b) Types 4 and 5 conferences

REFERENCES

1. W. S. Choe, T. J. Geok *et al.*, 'ATM-based multi-party conferencing system', *IEEE ICC '95*, pp. 592–596.
2. A. R. Reibman and A. W. Berger, 'Traffic descriptors for VBR video teleconferencing over ATM networks', *IEEE/ACM Trans. Networking*, **3**, 329–339 (1995).
3. H. W. Chu, H. K. Tsang and T. Yang, 'Call admission control of teleconference VBR video traffic in ATM networks', *IEEE ICC '95*, pp. 847–851.
4. V. P. Kompella, J. C. Pasquale *et al.*, 'Multicast routing for multimedia communication', *IEEE/ACM Trans. Networking*, **1**(3), (1993).
5. M. H. Ammar, A. Y. Cheung and C. M. Scoglio, 'Routing multipoint connections using virtual paths in an ATM network', *INFOCOM '93*, pp. 98–105.
6. S. Gupta, K. Ross and M. E. Zarki, 'Routing in virtual path based ATM networks', *GLOBECOM '92*, pp. 571–575, 1992.
7. R. H. Hwang, J. F. Kurose and D. Towsley, 'MDP routing in ATM networks using virtual path concept', *INFOCOM '94*, pp. 1509–1517, 1994.
8. S. Sabri and B. Prasada, 'Video conferencing systems', *Proc. IEEE*, **73**(4), 671–688 (1985).
9. Haruo Noma, Yasuichi Kitamura *et al.*, 'Multi-point virtual space teleconferencing system', *IEICE Trans. Commun.*, **E78-B**(7), 970–979 (1995).
10. Y. W. Leung and Tak-shing Yum, 'Connection optimization for two types of videoconferences', *IEE Proc. Commun.*, **143**(3), 133–140 (1996).
11. T. H. Turletti and C. Journal, 'Videoconferencing on the internet', *IEEE/ACM Trans. Networking*, **4**, 340–351 (1996).

12. T. S. Yum, M. S. Chen and Y. W. Leung, 'Video bandwidth allocation for multimedia teleconferences', *IEEE Trans. Commun.*, 457–465 (1995).
13. Tat Keung Chan and Tak-Shing Peter Yum, 'Analysis of multipoint videoconferencing under basic route-configuration assignment', *IEEE GLOBECOM '96*, pp. 867–876.
14. Tat Keung Chan and Tak-Shing Peter Yum, 'Analysis of multipoint videoconferencing under basic reroutable route-configuration assignment', *IEEE GLOBECOM '96*, pp. 899–906.
15. F. P. Kelly, 'Blocking probabilities in large circuit-switched networks', *Adv. Probab.*, **18**, 473–505 (1986).
16. W. Whitt, 'Blocking when service is required from several facilities simultaneously', *AT&T Tech. J.*, **64**(8), 1807–1856 (1985).
17. Eric W. M. Wong, *et al.*, 'Bandwidth allocation and routing in virtual path based ATM networks', London, pp. 1715–1720, November 1996.
18. Y. Sato and K. Sato, 'Virtual path and link capacity design for ATM networks', *IEEE J. Selected Areas Commun.*, **9**, 104–111 (1991).
19. D. Mitra, A. Morrison and K. G. Ramakrishnan, 'ATM network design and optimization: a multirate loss network framework', *IEEE/ACM Trans. Networking*, **4**(4), 531–543 (1996).
20. J. Filipiak, 'M-architecture: a structural model of traffic management and control in broadband ISDN', *IEEE Commun. Mag.*, **27**(5), 25–31 (1989).
21. S. P. Chung and K. W. Ross, 'Reduced load approximations for multirate loss networks', *IEEE Trans. Commun.*, **41**, 1222–1231 (1993).
22. J. S. Kaufman, 'Blocking in a shared resource environment', *IEEE Trans. Commun.*, **COM-29**, 1474–1481 (1981).
23. Keith W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer, Berlin, pp. 23–25, 1995.
24. L. Kleinrock, *Queuing Systems*, Vol. 1: Theory, p. 152, 1975.

AUTHORS' BIOGRAPHIES

Gang Feng received the BEng degree and MEng degree in Electronic Engineering from the University of Electronic Science and Technology of China (UESTC), in 1986 and 1989, respectively, and the PhD degree in Information Engineering in 1998 from the Chinese University of Hong Kong. From 1989 to 1995, he was with the Research Institute of Information Systems, UESTC. There his research work included the modeling and equalization of the nonlinear digital channels. He is currently with the Department of Electronic Engineering, City University of Hong Kong, as a term staff. His current research interest includes routing and performance evaluation for ATM networks, Reliable Multicast in Internet and IP over ATM.

Peter T. S. Yum worked in Bell Telephone Laboratories, U.S.A for two and a half years and taught in the National Chiao Tung University, Taiwan, for two years before joining the Chinese University of Hong Kong in 1982. He has published original research on packet switched networks with contributions in routing algorithms, buffer management, deadlock detection algorithms, message resequencing analysis and multiaccess protocols. In recent years, he branched out to work on the design and analysis of cellular network, lightwave networks that can accommodate the needs of individual customers. Professor Yum's research benefits a lot from his graduate students. Six of them are now professors at the local institutions. His recent hobby is reading history books. He enjoys playing bridge and badminton.

