Performance Analysis of Path Rerouting Algorithms for Handoff Control in Mobile ATM Networks

Jun Li, Roy Yates, and Dipankar Raychaudhuri, *Fellow, IEEE*

***Abstract*—This paper studies the effects of user mobility and handoff path rerouting on the traffic distributions in a mobile network environment. In mobile ATM networks, extra traffic load may be added to network links due to user mobility and handoff path rerouting. This requires higher network link capacity and possible topology reengineering in order to support the same quality of service (QoS) for mobile services. To capture the dynamic variations in mobile ATM networks, we propose to use a flow model. The model represents the mobile-generated traffic as a set of stochastic flows over a set of origin–destination (OD) pairs. The user mobility is defined by transfer probabilities of the flows and the handoff path rerouting algorithm is modeled by a transformation between the routing functions for traffic flows. The analysis shows that user mobility may cause temporal variations as well as smoothing effects on the network traffic. Using the flow network model, typical handoff path rerouting algorithms are evaluated through both analytical and experimental approaches. The evaluation methodology can be used for either redesigning the network topology for a given path rerouting algorithm or selecting a path rerouting algorithm for a given network topology under a specific mobile service scenario.**

***Index Terms*—Handoff control, mobile ATM network, network traffic modeling.**

I. Introduction

In RECENT years, the demand for wireless and broad-band services has been growing rapidly. Wireless ATM [1] has been considered a candidate solution for providing broad-band wireless services. A wireless ATM network consists of an ATM radio access network and a "mobile ATM" core network. The "mobile ATM" network is a common network infrastructure that supports user mobility for wireless ATM as well as other mobile services, such as global system of mobile communications (GSM), wireless LAN, etc. [2]–[4].

The key mobility support function in mobile ATM networks is the path rerouting process that is required when a mobile terminal moves from one access point to another. Previous studies, conducted through both research [3]–[11] and standardization [12]–[15] activities, have mainly focused on protocol design. Performance evaluation of path rerouting algorithms has been conducted through simulation [6].

This paper intends to evaluate path rerouting algorithms for handoff control. In particular, we examine the effect of the handoff path rerouting on fixed network using performance metrics such as resource utilization. It is commonly argued that the fixed network should have enough bandwidth relative to the radio link that the performance evaluation for the fixed network is unnecessary. However, we believe that, in the future, a large proportion of users will become mobile users who will share the same fixed network with the existing fixed users. When a large fraction of the users are mobile, inefficient handoff control for these mobile users can have a significant effect on resource consumption in the fixed network. Thus, studying the impact of handoff path rerouting on these fixed networks is very important for the optimum design of the network. Although this work is motivated by the mobile ATM concept, the study of traffic distribution variation in a mobile environment is an important topic for IP based mobile networks as well. Traditionally, network traffic modeling is using a queueing model, as it has been used sucessfully for telephony systems. In such a queuing model, the traffic at burst level is constant bit rate (CBR) and the traffic at call level is assumed as a Poisson arrival process. Although an ATM network uses virtual circuits for each call, its traffic distribution is largely different from telephony system at both call level and burst level. The integrated services offered by an ATM network produce calls with various distributions of interarrival time and call lifetime, which result in diversified traffic statistics at the call level. At the burst level, the bit transmission rates of calls may include a variety of types, such as CBR, (variable bit rate (VBR), available bit rate (ABR), or unspecified bit rate (UBR) varying over a very large range.

In mobile ATM, two new features, multimedia and mobility, raise basic difficulties in traffic modeling and network performance analysis. Multimedia connections imply a large number of traffic classes with heterogeneous service rates and terminal mobility causes traffic flows to migrate from one network link to another.

In a traditional traffic modeling hierarchy [16], the most detailed model is the queuing model, which gives probability distribution over all states. The next level is a diffusion model that characterizes the traffic flows by mean and variance. The third level is fluid approximation which takes time-varying averages and analyzes the nonstationary behavior of the traffic. The flow model with long term averages of traffic flows, which is used for network design, is perhaps the most abstract. Based on the above hierarchy, multimedia traffic is often described by hybrid models which characterize traffic flows at different time scales with varying degrees of detail [17]. These include fluid flow models that uses a queuing model at call level and a fluid approximation at burst level [18]. Another example is Markov-modulated Poisson process (MMPP) [19], which uses a histogram approximation at burst level.

Manuscript received November 3, 1999.

J. Li and D. Raychaudhuri are with NEC USA, C&C Research Laboratories, Princeton, NJ 08540 USA.

R. Yates is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08855 USA.

Publisher Item Identifier S 0733-8716(00)01762-5.

Traffic modeling can be used for either offline network design or online traffic control. Two problems may be decoupled and use different traffic models. For network design, a static flow model is often used [20] to construct a multicommodity problem based on network topology and offered traffic. For the online traffic control in a broad-band network, traffic models with bursty level distributions are required to estimate the QoS over network links [17], [21]–[23].

In mobile ATM networks, we wish to use traffic models for both network design and traffic control. First, we want to know if mobility support results in higher network capacity requirements or, in general, different network topology design. This is a network design problem. Second, we want to control quality of service (QoS) guarantees for each connection in the network, especially to relieve the QoS degradation caused by handoff. On the one hand, a simple static flow model will not reflect the dynamic route assignments and correlations of traffic flows caused by user mobility. On the other hand, as we have observed, a multiclass queuing model is complicated and needs strong assumptions when mobility is incorporated. Therefore, we propose a network flow model that has moderately detailed statistics. In the model, the overall bandwidth of the aggregated traffic on each origin–destination (OD) pair is represented by a continuous random process (stochastic flow) and the user mobility is captured by the migration of the traffic flows on different OD pairs.

By providing integrated data services, the time scale for call level and burst level in an ATM network becomes vague because a user can request, at a short time, a burst of call setups with very small bandwidth requirements, or in a long time, a call of only one burst with very large bandwidth requirements.

In this case, if a traditional queuing model is used, many traffic classes should be specified and each class may have a specific mobility pattern associated with a type of mobile service. Using a queuing model, the performance analysis will become extremely complicated. Fortunately, our target for traffic modeling is not to know the number of calls in each traffic class as traditional call admission control (CAC) needs. Our target is to estimate overall bandwidth requirement of the integrated service under a mobile environment, in order to re-engineer the network topology or link capacity. Therefore, we consider a flow model representing the distribution of integrated traffic over the mobile network. A flow is a continuous function of continuous time, representing the bandwidth consumption of offered traffic. Without distinguishing the call level and the burst level, we claim that the statistical variation of a flow may be caused by both call arrivals, departures, handoffs and/or rate variation of each call at the same time scale. That is, in any time interval (Δt), the variations caused at call level and at the burst level are at the same order of Δt. We do not address the optimization problems in network topology or routing design. Instead, we evaluate the extra cost that a core access network must pay in order to support mobile services. In addition, we identify the relative costs of different path rerouting algorithms. The purpose of performance evaluation for handoff path rerouting algorithms is twofold. First, the results can be used for reengineering the network topology and link capacities of the ATM core network according to the user mobility and path rerouting algorithm. Second, for a given mobile ATM core access network, the results can be used for selecting a path rerouting algorithm or altering user mobility patterns for a specific mobile service. The paper is organized as follows. In Section II, we introduce the network flow model. In Section III, we give the analytical results from an example with simple traffic requirements and network topology. In Section IV, through simulation, experimental results are given for a generic PNNI network topology with two different user mobility patterns.

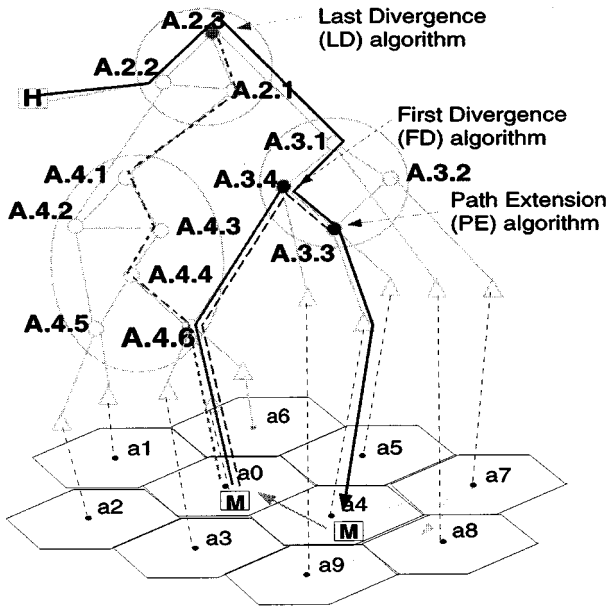

Fig. 1. A mobile ATM network and path rerouting.

II. Flow Model for Mobile ATM Networks

A. Definition of Mobile ATM Network

A mobile ATM network consists of a core ATM network and wireless access points. The core ATM network has ATM switches as nodes and fiber optic links between switches as edges. The network shown in Fig. 1 is a mobile ATM network. Some of the nodes (switches $A.4.5, A.4.6, A.3.4, A.3.3$) are radio access points (basestations). Each radio access point may serve several cells with multiple antennas. If there is a traffic flow from a mobile user M at the access point $A.3.3$ to a fixed host H, the flow may take a route as shown in the figure. When the mobile user moves from cell $a4$ to cell $a0$, it is necessary for the network to reroute the flow to a route which goes through the new access point $A.4.6$. This requires a handoff control process and path rerouting algorithm. Fig. 1 depicts three new routes. Each has a different anchor node (crossover switch or COS) to reroute the connection path, corresponding to a particular path rerouting algorithm. First is the last divergence (LD) algorithm with the COS at $A.2.3$. The COS is defined as the switch from where the new path and the old path merge, given that the new path is setup from the new BS to the CT. This algorithm aims to improve the efficiency of the network resource utilization by optimizing the path length after each handoff. This method is equivalent to the one with Loose Select CX (crossover switch), defined by Toh [6]. Second is the path

extension (PE) algorithm with the COS at $A.3.3$, which is the first switch from the mobile terminal. The PE algorithm aims for low complexity by just extending the original path from the old access point to the new access point. Third is the first divergence (FD) algorithm with the COS at $A.3.4$. The COS is defined as the switch from where the new path and the old path split, given that the new path is setup from the old BS to the new BS. FD offers a tradeoff method between complexity and performance by avoiding the duplicate paths generated by PE algorithm. FD algorithm is equivalent to Toh's backward tracking CX (crossover switch) [6]. In our previous work [11], we defined alternative path rerouting algorithms in a uniform way. These are three typical algorithms among them.

B. Network Topology and Mobility Pattern

A mobile ATM network has a *topology* which can be defined as a graph $G = [X, C]$, where $X = \{1, 2, \ldots, N\}$ are N nodes in the network and $C = \{c_{ij}, i, j \in X\}$ is a link capacity matrix. If $c_{ij} > 0$,, we say there is a link, or a C-edge (Capacity edge) from node i to node j. In addition, the network topology of a mobile ATM network has a mobility pattern, $G_w(t) = [Y, D(t)]$,, where $Y \subset X$ is a set of wireless access points and $D(t) = \{d_{ij}(t)\}$ is the user transition matrix of the network. At time t, a mobile user departing from access point i will enter node j with probability $d_{ij}(t)$. If $d_{ij}(t) > 0$, we say there is an M-edge (Mobility edge) from node i to j, i.e., node j is a neighbor of node i at time t. The user transition matrix $D(t)$ will depend on the users' call lifetimes and access point dwell times.

C. Traffic Flows and Handoff

For a given mobile ATM network with topology G, the network traffic can be specified as bandwidth requirements over the OD pairs. Each requirement can be represented by a random process (flow) in the unit of bandwidth per second. A flow consists of infinite numbers of infinitesimal calls over the OD pair. Each call can be viewed as a contributor of a mini flow, which may take a specific route in the network. The bandwidth of a traffic flow in a mobile ATM network is randomly changed because of call arrivals and departures, call handoffs, and call bursts. In general, a multimedia traffic flow may be quite complicated statistically. However, in this paper, we assume that the bandwidth requirements of traffic flows are memoryless, meaning the holding time is exponentially distributed. In this case, the traffic flow $\psi_k(t)$ on OD pair k is a Markov process which satisfies the following stochastic differential equations, for $k = 1, \ldots, K$

$$\Delta\psi_k(t) = -\mu_k(t)\psi_k(t)\Delta t + \sum_{l}^{M} p_{lk}(t)\mu_l(t)\psi_l(t)\Delta t + \Delta\phi_k(t). \quad (1)$$

In (1), $\phi_k(t)$ is the traffic demand representing new bandwidth requirement (in bits/s) over the time period Δt, $\mu_k(t)$ is the service rate of traffic flow $\psi_k(t)$ on OD pair k, and $p_{lk}(t)$ is the handoff transfer probability, representing the probability that a call in the traffic flow finished its service in an OD pair o_l and moves to a neighboring OD pair o_k. The probability that a departure call from o_l terminates is $p_{l0}(t)$. Clearly, $\sum_k p_{lk}(t) = 1$. With $\Psi(t) = [\psi_1(t), \ldots, \psi_K(t)]^T$ and $\Phi(t) = [\phi_1(t), \ldots, \phi_M(t)]^T$, (1) becomes in matrix form

$$\dot{\Psi}(t) = A(t)\Psi(t) + \dot{\Phi}(t) \quad (2)$$

where $A(t)$ is flow transition matrix with $A_{ij} = -\mu_i(t)$ for $i = j$ and $A_{ij} = \mu_i(t)p_{ij}(t)$ for $i \neq j$. The flow transition matrix $A(t)$ can be derived from the user transition matrix $D(t)$ given the call lifetime and user dwell times at the wireless access points, see Appendix A.

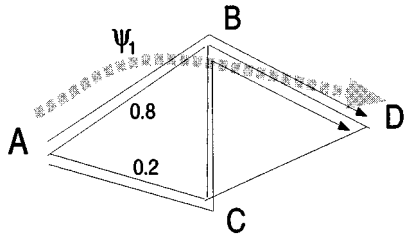

Fig. 2. Illustration of flow assignments by routing function.

D. Routing Functions and Path Rerouting Algorithms

In a mobile ATM network, each traffic flow $\psi_k(t)$ has a routing function defined by a vector $V_k(t) = [v_{k,1}(t), \ldots, v_{k,M}(t)]^T$, where $v_{k,u}(t)$ is the probability that the link u is being used by any call in the flow $\psi_k(t)$. For example, in Fig. 2, suppose ψ_1 is the flow for OD pair AD. One part of the flow takes the route ABD and rest of the flow takes the route $ACBD$. Corresponding to the order of the network links $[AB, AC, CB, CD, BD]$, one possible routing function is $V_1 = [0.8, 0.2, 0.2, 0, 1]^T$.

The routing function for each traffic flow may be static, as a result of network synthesis, or dynamic, as a result of dynamic routing process. In either case, the routing function may be varied by the path rerouting process for handoff control.

In particular, we assume that the traffic demand $\phi_k(t)$ is initially routed according to the routing function $Z_k(t) = [z_{k,1}(t), \ldots, z_{k,M}(t)]^T$, where $z_{k,u}(t)$ is the fraction of the traffic demand $\phi_k(t)$ on C-edge u. In addition, we define $w_{lk,u}(t)$ as the fraction of the handoff traffic on C-edge u based on a path rerouting algorithm. Suppose flow $\psi_k(t)$ on a network link u is $v_{k,u}(t)\psi_k(t)$. For all k and u, it must satisfy the traffic balance equation

$$\Delta(v_{k,u}(t)\psi_k(t)) = -\mu_k(t)v_{k,u}(t)\psi_k(t)\Delta t + \sum_{l}^{M} w_{lk,u}(t)p_{lk}(t)\mu_l(t)\psi_l(t)\Delta t + \Delta(z_{k,u}(t)\phi_k(t)). \quad (3)$$

To characterize $w_{lk,u}(t)$, we define

$$W_{lk}(t) = [w_{lk,1}(t), \ldots, w_{lk,M}(t)]^T = H_{lk}(t)V_l(t) \quad (4)$$

where the $M \times M$ matrix $H_{lk}(t)$ is called a path rerouting transform matrix, which reassigns the handoff flow from $\psi_l(t)$ to $\psi_k(t)$ to a new route according to given path rerouting algorithm.

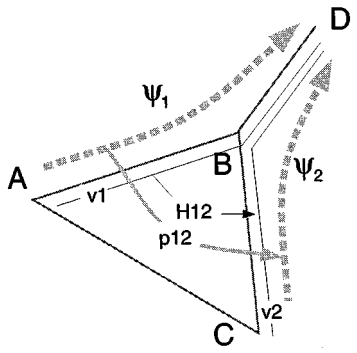

Fig. 3. Illustration of path rerouting algorithms.

Let us give an example in Fig. 3. The network has four nodes and two OD pairs with flows ψ_1 and ψ_2. The initial routing functions are $Z_1 = [10001]$ and $Z_2 = [00101]$, respectively, for links $[AB, AC, CB, CA, BD]$. A handoff path rerouting algorithm defined by $W_{12} = H_{12}V_1 = V_2$ and $W_{21} = H_{21}V_2 = V_1$ assigns handoff flows to the same routes as offered flow. In this case, the path rerouting transform matrix can be defined as

$$H_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$H_{21} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

In H_{12}, $[H_{12}]_{11} = 0$ means all handoff traffic from AB goes away while $[H_{12}]_{31} = 1$ means all handoff traffic from AB goes to CB. In general, the rerouting of the handoff flows can be either static and based only on fixed network topology or dynamic and based on both network topology and traffic loads. The $H_{lk}(t)$ may be deterministic functions of time or random processes. If $H_{lk}(t)$ is deterministic, for given path rerouting algorithm (given rules of rerouting), it is not difficult, though complicated for a large network, to obtain $\{H_{lk}(t)\}$. Otherwise, it is possible to estimate $\{H_{lk}(t)\}$ through statistical measurement at the time scale with any significant change of user mobility pattern. With the path rerouting transform matrix $\{H_{lk}(t)\}$, we can solve for the routing functions $\{V_k(t)\}$ by the following equation, which is based on (3):

$$\begin{aligned}\Delta(\psi_k(t)V_k(t)) = &-\mu_k(t)\psi_k(t)V_k(t)\Delta t \\ &+ \sum_{l}^{K} p_{lk}(t)\mu_l(t)\psi_l(t)W_{lk}(t)\delta t \\ &+ \Delta(\phi_k(t)Z_l(t)). \end{aligned} \tag{5}$$

We call the vector $\psi_k(t)V_k(t)$ a generic network flow of an OD pair o_k which carries its routing information. Solving the balance equations (2) and (5), we can evaluate the traffic loads over network links.

To summarize, a mobile ATM network is modeled by a flow network model with a fixed network topology G, an access network topology (mobility pattern) $G_w(t) = (G, D(t))$, network traffic flows $\{\psi_k(t)\}$, input traffic flow $\{\Phi(t)\}$, traffic service rate $\{\mu_k(t)\}$, routing functions $\{V_k(t)\}$, and path rerouting transformations $\{H_{lk}(t)\}$. The traffic distribution over the network depends on all factors; however, our analysis will focus on the effects of the user mobility $(D(t))$ and the handoff path rerouting $(\{H_{lk}(t)\})$.

III. Analysis of Traffic Distribution

We have given the flow balance equations at two levels. One is at OD pair level and the other is at the link level. In this section, the user mobility effects on traffic flows on both levels are analyzed.

A. Traffic Flows and User Mobility

1) Mean Values and Temporal Variation: Suppose traffic demands $\{\phi_k(t)\}$ are differentiable functions and let $d\Phi(t)/dt = \Omega(t)$, where $\Omega(t) = [\omega_1(t), \ldots, \omega_K(t)]^T$ represents the rates of traffic demand variation. If a steady state can be reached, the mean values of traffic flows can be solved as

$$m_\Psi(t) = -A^{-1}(t)m_\Omega(t). \tag{6}$$

We still denote the mean value as the function of time because of hour-by-hour variations of mobility pattern $A(t)$ can yield different steady-state operating points.

The solution above gives the properties of traffic flows as follows.

- The total traffic load, represented by $\sum_k m_{\psi_k}(t)$, is conserved regardless of user mobility.
- For a balanced OD pair o_l satisfying $\sum_k p_{lk}(t)\mu_l(t)\psi_l(t) = \sum_l p_{kl}(t)\mu_k(t)\psi_k(t)$, the traffic load is independent of user mobility. A balanced OD pair has the arrival rate of handoff flow equal to the departure rate of handoff flow. Equivalently, a balanced OD pair has the flows from new calls balanced with the flows from terminated calls, i.e., $\omega_k(t) = (1 - \sum_l p_{kl}(t))\mu_k(t)\psi_k(t)$.
- The traffic load on an unbalanced OD pair depends on user mobility. As user mobility increases, the traffic load may increase, if more flows handoff in than flows handoff out, or decrease, if more flows handoff out than flows handoff in.

These properties are quite intuitive and so the proofs are not given. The third property tells us that at a given time period, the traffic requirement over an OD pair may be larger or smaller because user mobility causes temporal variations on network traffic loads. Although the total traffic load is conserved at any given time, the network capacity requirements could be larger than those for nonmobile-generated traffic, since user mobility may cause traffic requirements to vary from time to time but the network topology and link capacity may not permit the corresponding changes.

Previous studies show that the dynamic routing schemes can help the network adapt to traffic load variations. Similarly, we will see the temporal variations can be relieved by dynamic path rerouting algorithms for handoff control.

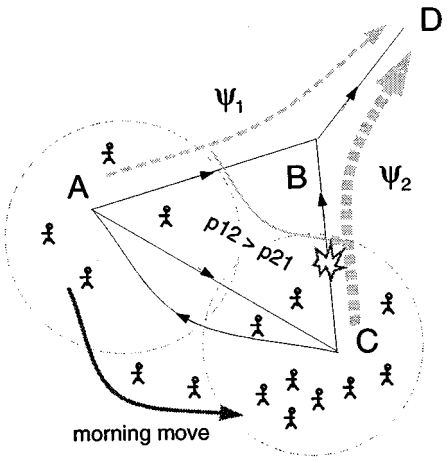

Fig. 4. Example of temporal variations.

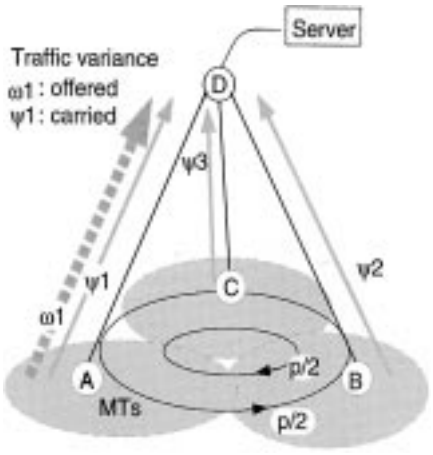

Fig. 5. Example for the smoothing effect.

Consider a simple example from Fig. 4. Suppose A is in a residential area while C is in a business area. D is attached with a newspaper server. Mobile users are reading newspapers while they are commuting from home to work or from work to home. In the morning $(t = t_m)$, more users move from A to C than from C to A. In the afternoon $(t = t_a)$, more users move from C to A than from A to C. They are reflected by $p_{12}(t_m) > p_{21}(t_m)$ and $p_{12}(t_a) < p_{21}(t_a)$, respectively. Suppose the new requests from A and C are symmetric and constant, i.e., $m_{\omega_1}(t) = m_{\omega_2}(t) = m_0$. The mean value of flows on OD pairs are

$$m_1(t) = \frac{(1+p_{21}(t))(1-p_{12}(t))}{(1-p_{12}(t)p_{21}(t))}\, Tm_0$$
$$m_2(t) = \frac{(1-p_{21}(t))(1+p_{12}(t))}{1-p_{12}(t)p_{21}(t)}\, Tm_0 \qquad (7)$$

where T is the average service time for traffic flow. When user mobility is symmetric, i.e., $p_{12}(t) = p_{21}(t)$, the traffic flows on both OD pairs are balanced, then the traffic loads are independent of user mobility, $m_1(t) = m_2(t) = Tm_0$. Otherwise, the mean values of flows $\psi_1(t)$ and $\psi_2(t)$ depend on the user mobility, as reflected by the handoff probability p_{ij}. For example, if in the morning $p_{21}(t_m) = 0.1, p_{12}(t_m) = 0.9$, then $m_2(t_m) = 1.8Tm_0$ and $m_1(t_m) = 0.2Tm_0$. Suppose traffic flow $\psi_2(t)$ always takes the route CBD, it is necessary to increase the capacity of CB to meet the traffic with 80% higher mean value. If in the afternoon, the user mobility is in the opposite direction, it requires the increment of the capacity of AB. As a result, to support user mobility, both AB and CB need more capacity.

2) Covariance and the Smoothing Effect: In a small network with few subscribers, the temporal variations caused by user mobility may be significant. However, if the number of subscribers is large, even in a short time period the traffic flow over an OD pair may be balanced by the movement of mobile users. However, even if the traffic flows are balanced, the variations in the traffic flows can still be observed through the second moments covariances.

If the rate of traffic demand $\Omega(t)$ is a white noise, or equivalently, the traffic demand $\Phi(t)$ is a Brownian motion process, the covariance of the traffic flows satisfies

$$A(t)R_\Psi(t) + R_\Psi(t)A^T(t) + \sigma_\Omega(t)\sigma_\Omega(t)^T = 0. \qquad (8)$$

For the proof, see Appendix B. The covariances of the traffic flows always depend on the user mobility regardless of the balance of traffic flows. However, when traffic flows are balanced, a smoothing effect can be observed, in that the variances of the traffic flows decrease as the user mobility increases.

Without loss of generality, we show this smoothing effect through an example in Fig. 5. The rate of traffic demands are independently identically distributed (i.i.d.) processes with mean m_ω and variance σ_ω, and user mobility is symmetric among three OD pairs with all $p_{kl} = p/2$. The traffic flows ψ_1, ψ_2, ψ_3 will be same on all three OD pairs because of the symmetry and will have mean and variance

$$m_\psi = \frac{1}{\mu(1-p)}\, m_\omega = Tm_\omega \qquad (9)$$

$$\sigma_\psi^2 = R_{kk} = -\frac{\sigma_\omega^2}{2\mu\left(-1+\dfrac{p^2}{2-p}\right)}$$
$$= T\sigma_\omega^2 \frac{2-p}{2(2+p)}. \qquad (10)$$

The correlation coefficient between traffic flows is

$$\rho = R_{kl}/R_{kk} = p/(2-p). \qquad (11)$$

As the user mobility increases with increasing p, the variances of traffic flows ψ_k decrease compared with the traffic demands $T\omega_k$, i.e., the traffic flows are being smoothed.

In Fig. 6, the variance vs.user mobility is shown, where user mobility $L = 1/(1-p)$ is the average number of cells a mobile user traverses during an average call lifetime. In addition to the variance, the correlation between traffic flows is shown. As the user mobility increases, the traffic flow correlation increases also. The simulation results (dashed lines) come from the model's simulation based on (1), which verifies the correctness of (8).

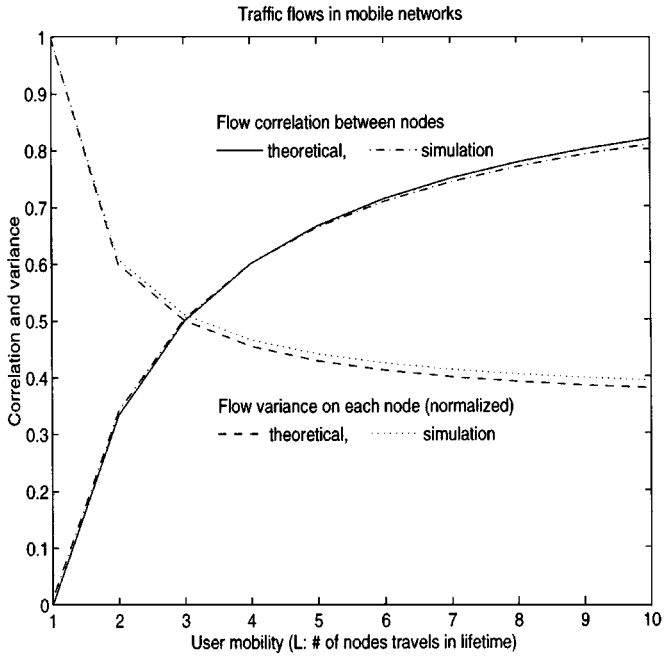

Fig. 6. Variance and correlation of traffic flows.

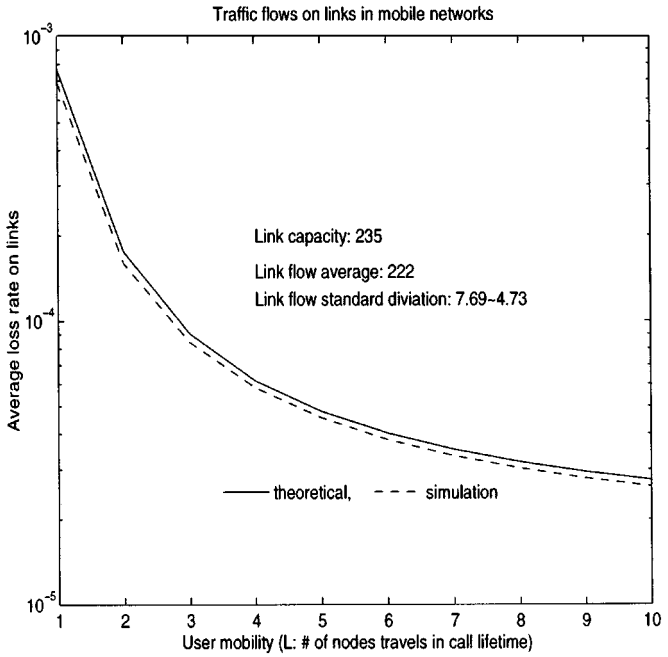

Fig. 7. Data loss rate reduced by smoothing effect.

Since the flow variances decrease as the user mobility increases, the traffic overflow (data loss rate) will decrease on those links with limited capacity. In Fig. 7, we show the loss rate versus user mobility by assuming traffic flows obey a Gaussian distribution. In general, the loss rate is determined by the tail distributions (or effective bandwidth) of the traffic flows. It is expected that the effective bandwidth of the traffic flows will decrease if the variances of the flows decrease. Thus the smoothing effect of user mobility should reduce the loss rate. The simulation result (dashed line) measure the average traffic overflow over a link with limited capacity.

From the analysis of the means and covariances of the traffic flows, we observed that user mobility causes temporal variations in the network traffic loads and has the smoothing effect on network traffic distribution. The results tell us that if in any period (short or long), the users' movement among cells is balanced, the traffic distribution over OD pairs is smoothed due to user mobility. In other words, the network service quality may be improved due to the user mobility. However, if the users' movements are not balanced among cells in any period, the traffic load on some OD pairs may increase while the others may decrease. As the result, the network service quality may be degraded due to the user mobility.

The smoothing effect seems contradictory to the analytical result of using traditional queuing model, in which user mobility always degrades system performance. The reason is that the higher is the user mobility, the higher the handoff frequency and the higher of call drop probability. In traditional queuing model, which is used for real-time service like telephony, the call drop rate is an important performance criterion. However, in our model, no call level QoS is specified, that is why we can see improvement of system performance in terms of data loss rate. We believe it is reasonable to assume a service scenario consistent with a session-oriented service. In this case, a call is a session which has certain QoS guarantees when it is active, but a session can go into a suspended mode if it fails to be admitted either initially or during handoff. A suspended session can become active when it is admitted later when resources are available. For real-time services, this would be unacceptable, but for mobile computing, this can be tolerated by many applications. When the majority of applications are session-oriented services, the smoothing effect can help to improve network performance.

B. Handoff Flows and Path Rerouting Algorithms

When a mobile user has a handover from one wireless access point to another wireless access point, all traffic flows associated with the user must migrate to routes linked to the new access point.

Equation (5) can be solved when routing functions $\{V_l(t)\}$ are independent of traffic flows $\{\psi_k(t)\}$, meaning the path rerouting transform matrices $\{H_{lk}(t)\}$ are deterministic and given by the path rerouting algorithm. In other words, alternative rerouting which dynamically adapts traffic distributions is not used. In this case, we can apply statistical averaging to (5), which yields, for $k = 1, \ldots, K$

$$\mu_k(t) m_{\psi_k}(t) V_k(t) = \sum_{l}^{K} p_{lk}(t) \mu_l(t) m_{\psi_l}(t) W_{lk}(t) + m_{\omega_k}(t) Z_k(t). \tag{12}$$

Substituting the solution of (6) into (12), we can have the solution of $V_l(t)$. With the solution of $\{V_l(t)\}$, we can obtain the traffic flow on each network link u

$$y_u(t) = \sum_{k}^{K} v_{k,u}(t) \psi_k(t) \qquad u = 1, \ldots, M. \tag{13}$$

Both means and covariances of $\{y_u(t)\}$ can be obtained based on the solutions for $\{\psi_k(t)\}$. Let us look at the example in Fig. 3 again, when we have different path rerouting algorithms. A call

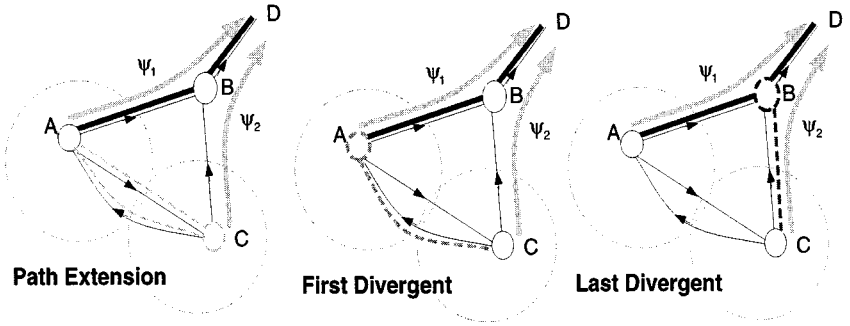

Fig. 8. Performance analysis of path rerouting algorithms.

in flow ψ_1 at A can rerouted to the flow ψ_2 at either B or C. If the anchor point is chosen as B, the link CB will take the handoff traffic flow from ψ_1. As we pointed out, the handoff traffic may not be balanced, for example, in the morning, there are more handoffs into C than handoffs out of C. The link capacity of CB must be increased to maintain the same QoS. However, it is possible to reroute the handoff traffic at other anchor nodes. For example, reroute the handoff traffic at A, i.e., the handoff flow from ψ_1 can take route $CABD$ instead of CBD. In this case, the link capacity of CA (or AC) needs to be increased to cover the handoff traffic flow while the link capacity of CB can remain the same. If the costs of links CA and AC are lower than the costs of links CB and AB, it is worthwhile to build CA (and/or AC) instead of increasing the capacity of AB and CB. This motivates the use of alternative path rerouting algorithm for handoff traffic control. A path rerouting algorithm for handoff control specifies where and how a connection path is rerouted. We have given three typical path rerouting algorithms in Section II, as illustrated by Fig. 1. Theoretically, path rerouting algorithm can be defined by the transform matrix $H_{lk}(t)$ with following features for each algorithm.

- *Last Divergence (LD) Algorithm:* Using the LD algorithm, statistically, all traffic flows for OD pair o_k will take the same set of routes $V_k(t)$, regardless if they are from new calls or handoff calls, since handoff calls also take the optimal route. This requires the path rerouting transform matrix to satisfy $H_{lk}(t)V_l(t) = V_k(t)$.
- *Path Extension (PE):* In the PE algorithm, the path rerouting transform matrix satisfies $W_{lk} = H_{lk}(t)V_l(t) = (I + \Delta_{lk}(t))V_l(t)$, where $\Delta_{lk}(t)$ represents the handoff flows transfered to the links on a extended path. For a mobile ATM network, if the extended paths are long, the performance of PE algorithm will be low. The $\Delta_{lk}(t)$ is always positive which makes the connection path longer and longer after each handoff.
- *First Divergence (FD):* the FD algorithm uses the same extended route as in PE for every handoff, however, the duplicated part is removed. Suppose the path rerouting algorithm satisfies $W_{lk}(t) = H_{lk}(t)V_l(t) = (I + \Delta_{lk}(t))V_l(t)$, where $\Delta_{lk}(t)$ represents the handoff flows transfer away from the links on the old path and to the links on the new path.

We will outline an analysis procedure to evaluate the performance of these three typical path rerouting algorithms through a simple example. Although the example is simple, as shown in Fig. 8, the methodology can be used in general situations. In Fig. 8, for the LD algorithm, the COS is always at B, so the handoff traffic flow from ψ_1 takes only the optimal route CBD after handoff. For the PE algorithm, the handoff traffic flow from ψ_1 will take an extended path CA, resulting in the route becoming $CABD$. In case a call is a handoff from A to C and then back from C to A, the handoff flow of the call will take an overall extended path ACA, which forms a loop. The route for the call becomes $ACABD$. For the FD algorithm, the COS is at either A or C and the handoff traffic flow from ψ_1 takes the extended path CA which makes the route become $CABD$. The FD algorithm takes the same extended path as the PE algorithm, however, it does not form a flow loop.

To identify the traffic load induced by path rerouting, we must:

- find the initial routing functions;
- find the path rerouting matrix;
- solve the balance equation on network links to obtain the routing functions;
- evaluate the incremental traffic load.

First, we determine the initial routing functions, i.e., the routing functions without user mobility. In this example with five network links AB, AC, CB, CA, BD, they are

$$Z_1 = [1,0,0,0,1]^T \quad Z_2 = [0,0,1,0,1]^T. \tag{14}$$

Second, we find the path rerouting transform matrices H_{12} and H_{21}. These will be different for three path rerouting algorithms.

- *LD Algorithm:* For the LD algorithm, by definition, the handoff flow in ψ_2 will take the route for flow ψ_1, i.e., $H_{21}V_2 = V_1$. Similarly, we have $H_{12}V_1 = V_2$. The path rerouting transform matrix can be represented as

$$H_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\Delta_{12} = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$H_{21} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\Delta_{21} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

where $\Delta_{lk} = H_{lk} - I$, represents the change in routing function due to handoff path rerouting. For example, $\delta_{lk,ij} = 1$ means the handoff flow from flow ψ_l on link i will be added to flow ψ_k onto link j.

The routing function $V_l(t)$ can be obtained by solving both balance equations for OD pairs and network links, (2) and (5). In this case, we have $V_l = Z_l$, meaning the routing function does not change due to handoff.

In the example, the mean and variance of traffic flows on network links are

$$m_Y = [m_1, 0, m_2, 0, m_1 + m_2]$$
$$\sigma_Y^2 = [\sigma_{\psi_1}^2, 0, \sigma_{\psi_1}^2, 0, \sigma_{\omega_1}^2 + \sigma_{\omega_2}^2].$$

Combined with the statistics we previously obtained for traffic flows, we can analyze the traffic distributions on network links. For example, if the mobility pattern is balanced, the only thing affected by the user mobility is the variances of the flows on

$$AB$$

and

$$CB$$

and they decrease as user mobility increases. If the mobility pattern is unbalanced, it is possible that $m_1 > m_2$, which requires AB to have higher capacity when user mobility exists.

- *PE Algorithm:* For the PE algorithm, the COS is always at the OldBS and every handoff extends the previous connection path by one more hop. This implies

$$H_{12} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\Delta_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$H_{21} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\Delta_{21} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Solving (5) based on the solution of (6), we find the routing functions for the PE algorithm

$$[V_1 V_2] = \begin{pmatrix} \frac{1}{1+p_{21}} & \frac{p_{12}}{1+p_{12}} & :AB \\ \frac{p_{12}p_{21}}{1-p_{12}p_{21}} & \frac{p_{12}(1+p_{21})}{(1+p_{12})(1-p_{12}p_{21})} & :AC \\ \frac{p_{21}}{1+p_{21}} & \frac{1}{1+p_{12}} & :CB \\ \frac{p_{21}(1+p_{12})}{(1+p_{21})(1-p_{12}p_{21})} & \frac{p_{12}p_{21}}{1-p_{12}p_{21}} & :CA \\ 1 & 1 & :BD \end{pmatrix}.$$

The routing functions depend on user mobility and they show how much of a given flow uses a specific network link. We can get the means and variances of link traffic flows $\{y_u(t)\}$ based on the solution of (2), for traffic flows on OD pairs. In our example, the mean traffic flow on AB is $m_{AB} = v_{1,1}m_1 + v_{2,1}m_2$. By substituting m_1 and m_2 from (8), we have $m_{AB} = Tm_0$, where T is the average call lifetime. It is independent of user mobility. While the mean traffic flow on AC is

$$m_{AC} = v_{1,2}m_1 + v_{2,2}m_2 = Tm_0\frac{p_{12}(1+p_{21})}{(1-p_{12}p_{21})} \qquad (15)$$

which will increase greatly as user mobility increases. The cost of the PE algorithm is the traffic flows on links AC and CA. To reduce the traffic flows on extended paths, we may use the FD algorithm, which removes possible duplication.

- *FD Algorithm:* For the FD algorithm, the COS is at the OldBS (or the NewBS) and a handoff may increase or decrease the path length by one. Thus

$$H_{12} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\Delta_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$H_{21} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\Delta_{21} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Given H_{12} and H_{21}, the solution of (5) and (6) yields

$$[V_1 V_2] = \begin{pmatrix} 1/(1+p_{21}) & p_{12}/(1+p_{12}) & :AB \\ p_{21}/(1+p_{21}) & 0 & :AC \\ p_{21}/(1+p_{21}) & 1/(1+p_{12}) & :CB \\ 0 & p_{12}/(1+p_{12}) & :CA \\ 1 & 1 & :BD \end{pmatrix}.$$

The routing functions depend on user mobility. However, unlike the PE algorithm, the traffic load on links AC and CA is not high since there is no duplicate usage of the links. The mean of traffic flow, $m_{AC} = Tm_0(v_{1,2}m_1 + v_{2,2}m_2)Tm_0p_{21}(1-P_{12})/(1-p_{12}p_{21})$, will not grow as fast as the traffic load for the PE algorithm. On the other hand, the total traffic loads on links AB and CB are same as that of the PE algorithm, which are independent of user mobility. So the FD algorithm gets the advantage of the PE algorithm with a much lower cost.

C. Performance Evaluation

The solution of the routing functions $V(t)$ can help us redesign the network topology or, if the network topology is fixed, evaluate which path rerouting algorithm should be chosen. As we have done in the example, we can analyze the mean and variance of the traffic flows on links. However, we may want to have a general idea how good or bad is a path rerouting algorithm in a given network. One criterion is the sum of traffic flows on all links, or network traffic volume, represented

$$|Y(t)| = \sum_{u}^{M} y_u(t). \tag{16}$$

A good path rerouting algorithm should have little increment in network traffic volume due to handoff control. In the example of Fig. 8, when handoff traffic flows are balanced, that is $p_{12} = p_{21} = p$, the network traffic volumes are $2(\psi_1+\psi_2)$, $(2+p/(1-p))(\psi_1+\psi_2)$ and $(2+p/(1+p))(\psi_1+\psi_2)$, for LD, PE and FD algorithms, respectively.

More generally, without the knowledge of traffic requirements $\psi_k(t)$, we may use link costs to evaluate a network topology. In particular, given g_u, the weight representing the cost of link u, we define the route lengths of flow l over OD pair o_l as

$$|V_l(t)|_g = \sum_{u}^{M} g_u v_{l,u}(t). \tag{17}$$

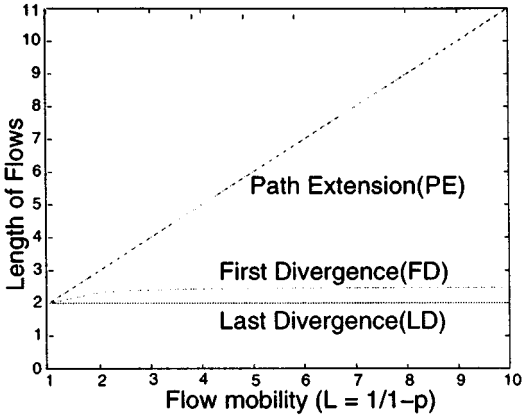

Fig. 9. The route lengths for the example in Fig. 8.

In Fig. 8, the cost of link AB may be greater than the cost of link AC. In our example, assuming $g_u = 1$ for all u, the route lengths of both OD pairs are 2, $2 + p/(1-p)$, and $2 + p/(1+p)$, for LD, PE and, FD algorithms, respectively. Fig. 9 shows the route length versus the user mobility, as measured by $L = 1/(1-p)$, which is the average number of cells a mobile users visits during a call lifetime. It is obvious that the PE path rerouting has the route length linearly increased as user mobility increases. Using FD algorithm, only a slightly increment in route length due to user mobility. While using the LD algorithm, the route length is supposed to be a constant relative to the user mobility.

When link costs g_u are not uniform, for example, $g_{AB} = g_{CB} = 5, g_{AC} = g_{CA} = 1, g_{BD} = 1$, the lengths of traffic flows for LD, PE, and FD algorithms will be 6, $6+p/1-p$ and $6+p/1+p$. The relative increasing rate of the lengths for the PE and FD algorithms will be smaller compared with the original length of traffic flows.

In the case of unbalanced traffic, the performance cannot be evaluated only by the route length of OD pairs. The traffic flows ψ_1 and ψ_2 may reach their peak values at different times. If the new path that takes handoff traffic flow is long, the extra traffic load caused by user mobility will be large. For example, if the peak value of m_1 is $1.8Tm_0$, the extra cost of using LD algorithm is $0.8g_{AB} = 4$ while only $0.8g_{AC} = 0.8$ for FD algorithm. Thus in the case of unbalanced traffic, we can evaluate the performance of path rerouting algorithms using the route length of handoff flow

$$|U_l(t)^+|_g = \left| \left(\sum_{k=0}^{K} \Delta_{lk}(t) V_l(t) \right)^+ \right|_g \tag{18}$$

where $+$ means only the positive elements are considered. In our example, suppose at time t_m, $p_{12} = p_{\max}$ and $p_{21} = p_{\min}$, at time t_a, $p_{12} = p_{\min}$ and $p_{21} = p_{\max}$. Also, $g_{AB} = g_{CB} = 5, g_{AC} = g_{CA} = 1, g_{BD} = 1$. We also call the route length of handoff flow as handoff new path length, which is shown in Fig. 10 as a function of the imbalance ratio $p_{\max}/p_{\min}$. The route length of handoff flow can also reflect the delay of handoff path rerouting. That is the longer the new path, the greater the delay.

For a given path rerouting algorithm, it is desirable to have both the route length and new path length minimized in order to reduce the extra network resource demand for supporting handoff control. Through simple examples, we have obtained analytical results on the performance of typical path rerouting algorithms. They are reflected by the route lengths of OD pairs and the length of handoff new paths. We have observed that the PE algorithm is not a good choice if the user mobility is high, and the LD algorithm may not be a good choice if traffic is unbalanced. For both concerns, it seems the FD algorithm is a good choice since it has the smallest handoff new path route length and only a slightly greater overall route lengths in OD pairs. We now verify the analytical observations through simulations of a larger network configuration.

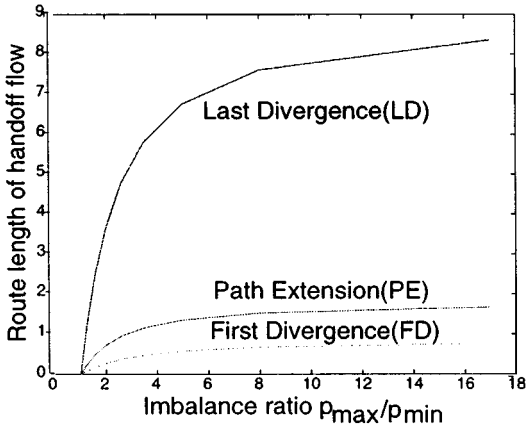

Fig. 10. The new path lengths for the analytical analysis on example in Fig. 8.

IV. Experiments Through Simulations

In the previous section, we have seen that the route lengths of OD pairs and the route lengths of handoff new paths can measure the performance of path rerouting algorithms. The route lengths of OD pairs reflect overall traffic load in the network and the route lengths of handoff new paths reflect the extra traffic load when handoff traffic flows are unbalanced.

The method we are going to use is to have several calls active in the network. For each call, one party is mobile terminal in the mobile ATM subnetwork and the other party is the fixed host sitting outside of the mobile ATM subnetwork. Then we will let the mobile terminals handoff to another cell according to certain path rerouting algorithm. The length of the whole connection path for each call is measured after each handoff, in order to estimate the route lengths of OD pairs. The length of the new path of each call is measured in order to estimate the route lengths of handoff new paths.

The network used in the simulation is as shown in Fig. 11. It is an example in ATM Forum PNNI specifications [24] for PNNI hierarchical structure. In this network, there are 27 switches and three hierarchical levels. We use peer group C as a fixed backbone network and peer group A and B as the mobile ATM access network. Every switch in group A and B is considered a basestation. Mobile terminals can be attached through any switch in A, B and move (handoff) from one to another. Although the network topology in this example is given as hierarchical, our approach applies to any arbitrary network topology.

A. Simulation Results with Fixed Mobility Pattern

In the first simulation, we assume the user mobility pattern matches the network topology, i.e., any two nodes within A, B are neighboring access points (having an $M-$edge) if there is a link (C-edge) between them. We call this setup as fixed mobility pattern.

The simulation is conducted as follows. Three mobile terminals starting calls at $A.1.2$, $A.3.4$, and $B.2.3$ to a remote server at $C.2$ in the fixed network. Each call starts with an exponential distributed lifetime represented by t_f with $E[t_f] = L$ s and performs a handoff every second. Hence the average number of handoffs per call lifetime is L, which represents the user mobility p as $L = 1/1 - p$. when a call is terminated at an access point, a new call is started at the same access point of the mobile terminal. The procedure continues until each mobile terminal performs 10 000 handoffs. A mobile terminal will randomly handoff to one of the neighboring cells with equal probability.

In the simulation, we use three path rerouting algorithms: path extension (PE); first-cross-to-CT (FXCT, as an approximation of LD algorithm, which uses first crossover switch toward the CT as the COS); and first-cross-to-oldBS (FXBS, as an approximation of FD algorithm, which uses first crossover switch toward the OldBS as the COS).

Fig. 12 shows the route lengths for the three path rerouting algorithms. It is obvious that PE has linear increasing in route length. Higher user mobility implies more handoffs per call lifetime. It is not surprising that FXBS has only a little extra path length over the FXCT algorithm and is almost constant as mobility increases. This is consistent with our analytical example in Fig. 9. Using the FXBS algorithm, the network needs about 10% more capacity than using FXCT algorithm, however, the capacity increment does not change much as the user mobility increases.

Fig. 13 shows the route lengths of handoff new paths for the three algorithms. They do not change much with the user mobility. However, they will cause the handoff traffic flow to increase at different rates when imbalance ratio increases. The FXCT algorithm has a new path length about 0.5 hops more than that of the FXBS algorithm. When handoff flows are not balanced, the capacity increment on the new path for the FXCT algorithm can be much larger than for the FXBS algorithm. Since the new path takes the handoff flows, if handoff traffic flows double the original traffic load on the new paths, the FXCT algorithm will need one more hop $(2 * 0.5)$ on average than the FXBS algorithm, which is also about 10% of total average path length. So depending upon the mobility pattern, one algorithm may consume less of the network resources.

B. Simulation Results with Random Mobility Pattern

In the second simulation, the neighborhood of the access network is randomly generated among the nodes in peer groups A, B. For each node x, a node y is a neighbor of x with a probability of $16/N$, $4/N$ and $1/N$, if y and x belong to same peer group at $G0$, $G1$, and $G2$ level, respectively, where N is the

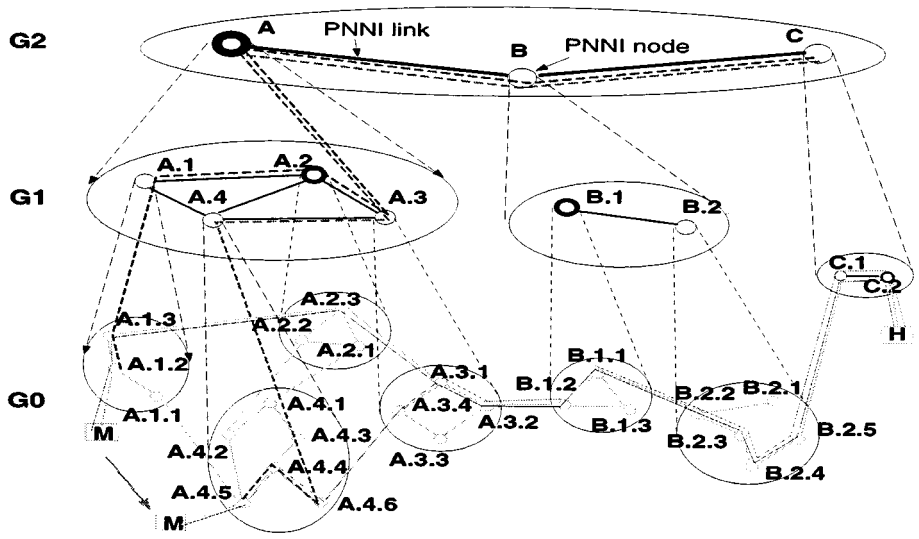

Fig. 11. A sample PNNI network.

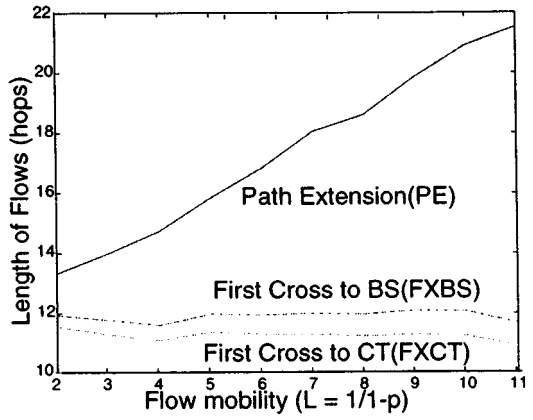

Fig. 12. Average OD whole path length with fixed mobility.

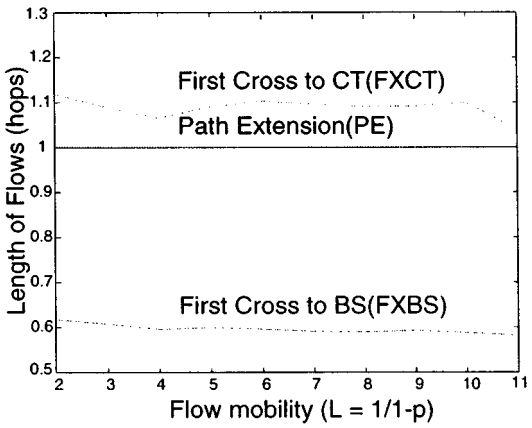

Fig. 13. Average new path length with fixed mobility.

total number of nodes. Based on this selection, we have each node having about four neighbors. If a node x is closer to y in terms of PNNI hierarchy, the probability that x is a neighbor of y is higher. The probability a handoff is to which neighbor of a node is again same as in last example. That is 1/number of neighbors. We call this neighborhood setup as random mobility pattern.

From Fig. 14, we can see the route lengths for FXCT and FXBS are about the same as the first example. The FXBS needs about 10% more hops than FXCT algorithm. However, in Fig. 15, we observe the difference between the route length for handoff new paths for FXCT and FXBS is bigger about 0.7 hops. Therefore when handoff flows are imbalance, the handoff flows may result in greater capacity requirements for the FXCT algorithm than the FXBS algorithm.

Again in this experiment, we can see the PE algorithm is costly. Since the neighborhood is random, not only the whole path length is linearly incrementing with user mobility, but also the average new path length is the highest within three algorithms. For FXCT algorithm, the whole path length is the shortest but the new path is larger. The overall performance

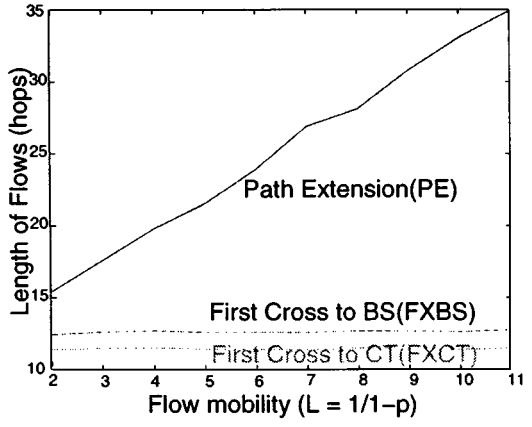

Fig. 14. Average whole path length with random mobility.

of the FXBS algorithm is good since it keeps both the whole path length and the new path length relatively short, which implies that the extra network resources required for supporting handoff is small and the handoff latency is small among three algorithms.

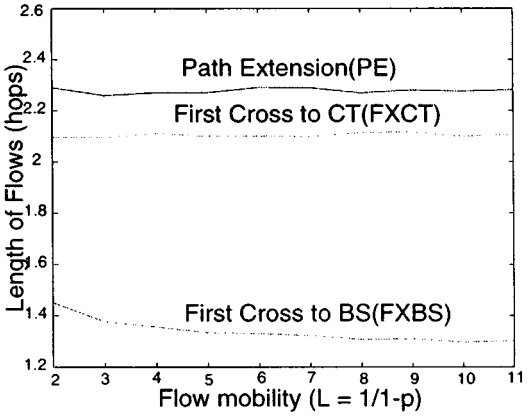

Fig. 15. Average new path length with random mobility.

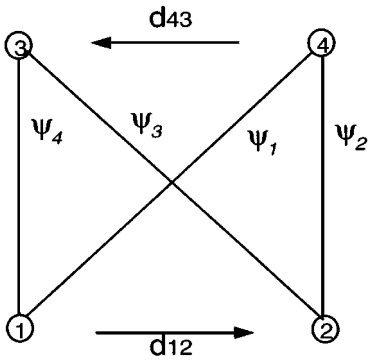

Fig. 16. Transition of traffic flows.

V. Conclusion

We have investigated the traffic distribution in mobile ATM networks through a dynamic flow model. We observed that user mobility will produce two effects over network traffic. One is the temporal variations in network traffic loads, which can cost network resources when the handoff traffic is unbalanced, and the other is the smoothing effect of the network traffic distribution, which can relieve traffic load, especially when handoff traffic is balanced. Typical handoff path rerouting algorithms for handoff control are evaluated. By measuring the route lengths of OD pairs and the route length of handoff new paths, we conclude that the FD (or FXBS) algorithm has advantages over other algorithms in most situations, especially when the link capacity in the access network costs less than that in the core network. The performance analysis can also help us redesign the network topology by adding more links or increasing capacity in the network based on the user mobility pattern. Furthermore, the results may suggest a rearrangement of access points to balance handoff traffic flows. Future work will include analytical models for PNNI hierarchical networks and the simulation in a large-scale network setup.

Appendix A

Usually, the mobility is measured through the users' movement among access points, and the mobility pattern is specified only by user transition matrix D. The flow transition matrix A is not given directly. However, we can derive A from D as follows. In Fig. 16, traffic flows are directional, from nodes 1 and 2 to nodes 3 and 4. A part of traffic flow ψ_1 can be handed off to traffic flow ψ_2, ψ_3 or ψ_4 due to: 1) end user at access point 1 moves to 2; 2) end user at access point 4 moves to 3; or 3) the end users at both 1 and 4 move to 2 and 3 at the same time.

Suppose an active user in node i departs to node j with probability d_{ij}. The time period that a user stays at an access point i is a random variable T_i with the exponential distribution $F_{T_i}(t) = 1 - e^{-\gamma_i t}$. Note that γ_i is the user departure rate at access point i. Suppose a typical call departure rate is Γ. A call departing from the access point 1 is either because of the termination or handoff, so a call departure rate from the access point 1 is

$$\nu_1 = \gamma_1 + \Gamma. \tag{19}$$

Therefore, the traffic flow ψ_1 has departure rate $\mu_1 = \nu_1$. Within the departure traffic flow from the access point 1, a fraction Γ/ν_1 is terminated and a fraction γ_1/ν_1 is added to the flows at other access points, we have $p_{k0} = \Gamma/\nu_i$, and the flow transition probability to a flow 2,3,4 are

$$\begin{aligned} p_{12} &= \frac{\gamma_1}{\nu_1}\, d_{12} = \frac{\gamma_1}{\gamma_1 + \Gamma}\, d_{12} \\ p_{13} &= \frac{\gamma_1}{\nu_1}\, d_{12} d_{43} = \frac{\gamma_1}{\gamma_1 + \Gamma}\, d_{12}\, d_{43} \\ p_{14} &= \frac{\gamma_1}{\nu_1}\, d_{43} = \frac{\gamma_1}{\gamma_1 + \Gamma}\, d_{43} \end{aligned} \tag{20}$$

respectively, assuming the user movement from access point 1 to 2 is independent from the user movement from access point 4 to 3. Given mobility pattern of an access network $G_w = [Y, D]$, and the average call lifetime $T(1/\Gamma)$, the flow transition matrix can be obtained.

Appendix B

For traffic flows $\Phi(t)$ over network OD pairs, balance equation is satisfied

$$\dot{\Psi}(t) = A(t)\Psi(t) + \dot{W}(t). \tag{21}$$

When $W(t)$ is a K-dimensional Brownian motion, i.e., the arrival traffic flow is a K-dimensional Gaussian random processes $\Omega(t) = \dot{W}(t)$. By applying statistical expectation on both side of Equation (1) and exchanging derivative and expectation, we have the mean vector of traffic flows is

$$\dot{m}_\Psi(t) = A(t) m_\Psi(t) + m_\Omega(t). \tag{22}$$

The covariance of $\Psi(t)$ is defined as

$$R_\Psi(t) = E[(\Psi(t) - m_\Psi(t))(\Psi(t) - m_\Psi(t))^T]. \tag{23}$$

By applying derivative to both sides and exchange with the expectation operator, we obtain derivative of the covariance matrix satisfying

$$\begin{aligned} \dot{R}_\Psi(t) = E[&(\dot{\Psi}(t) - \dot{m}_\Psi(t))(\Psi(t) - m_\Psi(t))^T \\ &+ (\Psi(t) - m_\Psi(t))(\dot{\Psi}(t) - \dot{m}_\Psi(t))^T] \end{aligned} \tag{24}$$

The first term on the right-hand side is

$$\begin{aligned} &E[(A(t)\Psi(t)+\Omega(t)-A(t)m_\Psi(t)-m_\Omega(t))(\Psi(t)-m_\Psi(t))^T] \\ &\quad = E[A(t)(\Psi(t)-m\Psi(t))(\Psi(t)-m\Psi(t))^T \\ &\qquad + (\Omega(t)-m_\Omega(t))(\Psi(t)-m\Psi(t))^T] \\ &\quad = A(t)R_\Psi(t) + E[\int_0^s (\Omega(s)-m_\Omega(s))(\Omega(t)-m_\Omega(t))^T ds]. \end{aligned} \tag{25}$$

Since $W(t)$ is Brownian motion, the derivative process $\Omega(t)$ is a white noise. Hence the power spectrum is an impulse. The integration becomes

$$\int_0^s \frac{1}{2}\sigma_\Omega(s)\sigma_\Omega^T(t)\delta(t-s)ds = \frac{1}{2}\sigma_\Omega(t)\sigma_\Omega^T(t). \tag{26}$$

The second term is symmetric to the first one, so we have

$$\dot{R}_\Psi(t) = A(t)R_\Psi(t) + R_\Psi(t)A^T(t) + \sigma_\Omega(t)\sigma_\Omega^T(t). \tag{27}$$

REFERENCES

[1] D. Raychaudhuri and N. D. Wilson, "ATM-based transport architecture for multiservices wireless personal communication networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1401–1414, Dec. 1994.

[2] A. Iwata, D. Raychaudhuri, R. Yuan, and H. Suzuki, "Rationale and framework for wireless ATM specification," in *Proc. ATM Forum/95-1646/PLEN*, 1995.

[3] R. Yuan, S. K. Biswas, L. J. French, J. Li, and D. Raychaudhuri, "A signaling and control architecture for mobility support in wireless ATM networks," *ACM/Baltzer Mobile Networks and Applications*, vol. 1, no. 3, Dec. 1996.

[4] A. Acharya, J. Li, and D. Raychaudhuri, "Mobility management in wireless ATM networks," *IEEE Commun. Mag.*, vol. 35, pp. 100–109, Nov. 1997.

[5] A. Acampora and M. Naghshineh, "An architecture and methodology for mobile-executed handoff in cellular ATM networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1365–1375, Dec. 1994.

[6] C. K. Toh, "Crossover switch discovery for wireless ATM LANs," *ACM/Baltzer Mobile Networks and Nomadic Applicatios*, vol. 1, no. 2, Dec. 1996.

[7] M. Veeraraghavan, M. Karol, and E. Engr, "Mobility and connection management in a wireless ATM LAN," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 50–68, Jan. 1997.

[8] H. Mitts, H. Hansen, J. Immonen, and S. Veikkolainen, "Lossless handover for wireless ATM," *ACM/Baltzer Mobile Networks and Applications*, vol. 1, no. 3, Dec. 1996.

[9] B. Akoyl and D. Cox, "Reroutihg for handoff in a wireless ATM network," *IEEE Personal Commun.*, vol. 3, Oct. 1996.

[10] S. Seshan, H. Balakrishnan, and R. H. Katz, "Handoffs in cellular wireless networks: The daedalus implementation and experience," *IEEE Wireless Personal Commun.*, vol. 4, pp. 141–162, Mar. 1997.

[11] J. Li, R. Yates, and D. Raychaudhuri, "Unified handoff control protocol for dynamic path rerouting in mobile ATM networks," in *Proc. 9th IEEE Int. Symp. Personal Indoor and Mobile Radio Communications*, Boston, MA, Sept. 1998.

[12] J. Li, A. Acharya, and D. Raychaudhuri, "Signaling syntax extensios for handoff control in mobile ATM," in *Proc. ATM Forum/96-1625*, Dec. 1996.

[13] A. Acharya, J. Li, and D. Raychaudhuri, "Primitives for location management and handoff control in mobile ATM networks," in *Proc. ATM Forum/96-1121*, Aug. 1996.

[14] B. Rajagopalan, H. Mitts, K. Rauhala, and G. Bautz, "Proposed handover signaling architecture for release 1.0 WATM baseline," in *Proc. ATM Forum/97-0845*, Sept. 1997.

[15] The ATM Forum, "Baseline txt for wireless ATM specifications," in *Proc. ATM Forum/BTD-WATM-01.16*, Feb. 1998.

[16] J. Filipiak, *Modeling and Control of Dynamic Flows in Communication Networks*. Berlin, Germany: Springer-Verlag, 1988.

[17] M. Schwartz, *Broadband Integrated Networks*. Englewood Cliffs, NJ: Prentice Hall, 1996.

[18] B. Maglaris *et al.*, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834–844, 1988.

[19] P. A. Skelly *et al.*, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Trans. Networking*, vol. 1, pp. 446–459, 1993.

[20] M. Minoux, "Network synthesis and optimum network design problems: Models, solution methods and applications," *Networks*, vol. 19, pp. 313–360, 1989.

[21] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.*, pp. 1598–1608, Dec. 1988.

[22] R. Guerin *et al.*, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select Areas Commun.*, vol. 9, pp. 968–981, Sept. 1991.

[23] W. R. Keith, *Multiservice Loss Models for Broadband Telecommunication Networks*. New York: Springer-Verlag, 1995.

[24] The ATM Forum, "Private network-network interface specification (PNNI) version 2.0," in *Proc. ATM Forum/BTD-PNNI 02.00*, Sept. 1997.

Jun Li received the B.Sc. and M.Sc. degrees from Xidian University, Xi'an, China, in 1983 and 1985, respectively, and received the Ph.D. degree from Rutgers University, Piscataway, NJ, in 1999, all in electrical engineering.

He was with Beijing Information Technology Institute as a Research Engineer from 1986 to 1990, conducting research on speech synthesis and recognition. During October 1987 to December 1988, he worked in NEC's C&C Central Labs, Kawasaki, Japan, as a Visiting Scholar. Since 1990, he worked as a software engineer for several network software companies, in Tokyo and San Francisco, respectively, for developing and localizing network software products. He started his Ph.D. program in Rutgers University in the Fall of 1993, studying mobile and wireless ATM networking systems. Since the summer of 1995, he joined NEC C&C Reseach Laboratories in Princeton, New Jersey, now as a Research Staff Member, working on the network protocol design for next-generation mobile communication systems.

Roy Yates received the B.S.E. degree in 1983 from Princeton University, Princeton, NJ, and the S.M. and Ph.D. degrees in 1986 and 1990 from Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

Since 1990, he has been with the Wireless Information Networks Laboratory (WINLAB) and the Electrical and Computer Engineering Department at Rutgers University, Piscataway, NJ, where he is currently an Associate Professor. His research interests include power control, interference suppression, and media access protocols for wireless communications systems.

Dipankar Raychaudhuri (S'78–M'79–SM'87–F'95) received the B.Tech (Honors) degree from the Indian Institute of Technology, Kharagpur, in 1976 and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Stony Brook, in 1978 and 1979, respectively.

He was with the David Sarnoff Research Center (formerly RCA Laboratories), Princeton, NJ, as Member of Technical Staff from 1979 to 1987, Senior Member of Technical Staff from 1988 to 1989, and Head, Broadband Communications Research, from 1990 to 1992. At Sarnoff, he worked on a range of R&D topics including: very small aperture terminal (VSAT) satellite networks, direct broadcast satellite (DBS), packet video, digital HDTV, and wireless data networks. During the period 1990–1992, he led the multicompany project team responsible for system design and specification of the "Advanced Digital HDTV" prototype tested by the U.S. Federal Communication Commission in 1992. Since January 1993, he has been with NEC USA, C&C Research Laboratories, Princeton, NJ, where he is currently Assistant General Manager and Department Head (Systems Architecture), with focus on multimedia networking technologies including IP and ATM switching/protocols, broad-band wireless, and distributed multimedia software. In 1995, his research group demonstrated one of the first proof-of-concept wireless ATM networks ("WATMnet") capable of delivering multimedia services to portable computing devices. He is an active participant in related standardization activities and has been Vice-Chair of the ATM Forum's Wireless ATM Working Group since its inception in 1996. He has authored over 100 technical papers and 13 U.S. patents.

Dr. Raychaudhuri is currently a Technical Editor of the IEEE MULTIMEDIA MAGAZINE, and has previously served as Editor, IEEE TRANSACTIONS ON NETWORKING (1993–1998), Editor, IEEE TRANSACTIONS ON COMMUNICATIONS (1992–1993), Technical Editor, IEEE COMMUNICATIONS MAGAZINE (1980–1994), and IEEE Communication Society Distinguished Lecturer (1992-1996).