

GENERALIZED MODELING FRAMEWORK FOR HANDOFF ANALYSIS

Ashutosh Dutta, Bryan Lyles Henning Schulzrinne Tsunehiko Chiba, Hidetoshi Yokota, Akira Idoue
Telcordia Technologies Columbia University KDDI R&D Labs
Piscataway, New Jersey, USA New York, USA Japan

ABSTRACT

A mobility event is the result of one network connection path being replaced by another via the rebinding of common system properties. The rebinding is a sequential process that may potentially involve multiple protocol layers of the mobile and require multiple network interactions. This overall process results in a period of time in which network service is degraded by transient data loss and increased end-to-end delay. Optimizations of the handover process mitigating these service degradations have been developed without a formal or systematic framework for mobility solutions. We develop a systematic systems model of the basic properties associated with a mobility event and design a framework around these properties that can provide methodologies for optimizing the handoff components. We then summarize the experimental results from a 3GPP2-based mobility testbed and highlight the delays associated with the functional components of the handoff event. We apply two types of optimization techniques and compare the results.

I. INTRODUCTION

Universal wireless connectivity to communications and information has advanced the world towards ubiquitous computing. But we need handoff mechanisms, often at multiple protocol layers, to allow a mobile terminal to move from cell to cell and maintain service continuity. The simple description of mobility is that as a mobile terminal moves, it releases its binding to the cell that it is leaving and establishes a binding to the cell that it is entering while preserving the existing session. The cellular telephony community has long implemented service and technology specific mobility protocols that handoff voice sessions as the user moves from cell to cell. Because voice service quality is highly sensitive to service interruptions, cell to cell handoffs in a cellular environment have been highly optimized and are not noticed by the public. Tripathi et al [1] describe some of the handoff technologies associated with cellular mobility. For IP traffic, the IETF has defined mobility protocols for both IPv4 and IPv6 [2], [3]. However, the Internet is dramatically more diverse than cellular voice in the range of link layer technologies used to support IP-based traffic, the number of economic units supplying IP service, the authentication protocols and services running above IP. This diversity meant that the IETF could not easily design into the mobility standards the handoff optimization seen in cellular voice without the underlying support from layer 1 assisted soft-handoff. As a result, standard mobile IP-based handoffs can take few seconds to complete a handoff operation and degrades the desired quality of service in the process. IP's transformation from a service supporting email and file transfer to the base layer for network convergence means that the constraints on handover performance are becoming much more stringent. The mechanisms and design principles needed

for building optimized handovers in the context of mobile Internet services are poorly understood and need better analysis. We analyze the handover event, identify the functional components that contribute to the delay during handoff, and design a model that can characterize the handoff event and represent the components responsible for handoff delay and packet loss.

The remainder of the paper is organized as follows. We describe related work in Section II. Section III analyzes the functional components associated with a handoff process. Section IV introduces Petri Net-based models to describe the handoff event, and compare with the experimental handoff related results. Finally, we conclude the paper in Section V.

II. RELATED WORK

Several mobility protocols such as MIP [2], SIP [4], and MIPv6 [3] have been designed to function at different layers and there are relevant optimization techniques SIPFast [5], FMIPv6 [6], HMIPv6 [7] for each of these mobility protocols. However, to the best of our knowledge, there is no prior work defining a formal representation of a mobility event. We cite a few examples that have attempted to model certain wireless access characteristics and Mobile IPv6. Marshan et al [8] have used a Petri net-based model to analyze the performance characteristics of wireless Internet access for GSM/GPRS system. Amadio et al [9] have modeled the IP mobility using a process calculi approach and has applied this to a specific protocol such as Mobile IPv6. The above work has used a software agent approach to model the mobility. None of these has actually attempted to model the atomic processes associated with a handoff. We address this issue by systematically analyzing the handoff event and a Petri net model that can represent the functional components.

III. ANALYSIS OF HANDOFF PROCESS

In this section, we analyze the functional components that comprise a mobility event and highlight the associated common properties. Analysis of these properties provides a systematic approach to the design of a formal mobility model. In order to provide an acceptable quality of service for interactive VoIP and streaming traffic, one needs to limit the end-to-end delay, jitter and packet loss. Based on the type of application (e.g., interactive, streaming, data) standards organizations define threshold limits for these metrics. For example, for one-way delay, ITU-T G.114 recommends 150 ms as the upper limit for most of the applications, and 400 ms as generally unacceptable delay. One way delay tolerance for video conferencing is in the range of 200 ms to 300 ms. According to ETSI TR 101 a normal voice conversation can tolerate up to 2% packet loss in general. A mobile subjected to handoff needs to follow similar guidelines.

Handover process can be homogeneous and heterogeneous. Heterogeneous handover includes movement between different types of wireless networks and administrative domains. Supporting terminal handovers across heterogeneous access networks such as CDMA (Code Division Multiple Access), IEEE 802.11, WiMAX (IEEE 802.16) and GPRS (General Packet Radio Service) needs to take into account different QoS, security, and bandwidth characteristics associated with each type of access network. Similarly, movement between two different administrative domains will need to re-establish authentication and authorization in the new domain. By experimental analysis, we have shown that a real-time traffic is impaired when the mobile is subjected to heterogeneous and homogeneous handoff in the absence of any optimization technique [10]. It takes from 3 seconds up to 15 seconds depending on the type of handover. Thus, it is important to identify the elementary set of operations that contribute to this delay and associated packet loss.

As part of the handoff analysis, we investigated and analyzed several cellular and IP-based mobility protocols. In particular, we studied the handoff process associated with the cellular protocols such as GSM, CDMA, 802.11, network layer mobility protocols, such as Mobile IP over IPv4, and IPv6 and application layer mobility protocol such as SIP. After a careful study of these mobility protocols, we identify the basic functions and systems properties that are affected during a mobility event. By analyzing the atomic operations involved in each of these basic functions, we design a generalized modeling framework that can represent a handoff event. This model can also formalize the rules of optimization to enhance the overall systems performance during a mobility event for any specific type of handoff.

We categorize the handoff event primarily into the following six phases:

1) *Handover initiation* 2) *Network and resource discovery* 3) *Network selection* 4) *Network attachment* 5) *Configuration* and 6) *Media redirection*.

Each of these phases includes several sub-phases. We describe the details of each these phases in the following sections.

A. Handover Initiation

During this phase, the mobile or the network determines the need for the handoff based on the network condition, and sends trigger to the lower layers to start the handoff process. A handover initiation process can be either mobile-initiated handover, network-initiated handover, mobile controlled handover or network-controlled handover. For example, during a mobile-initiated handoff, based on Signal-to-Noise Ratio (SNR) or the quality of service of the existing application, the mobile decides regarding the impending handoff, and starts the network discovery process to determine the best available target network. In a network controlled handover, network triggers the mobile for handoff.

B. Network and Resource Discovery

In the second phase, the mobile or the network discovers possible new points of attachment. This process involves discovering both the neighboring networks and the resources within the network. Once the target network is discovered, several resource parameters within the target network need to be retrieved including channel number, bandwidth, encryption algorithm, authentication server, registration server, and configuration server. Resource discovery process helps the mobile to associate with the correct channel number and proper authentication parameters in the new network so that it can communicate successfully. Based on the type of access technology, the discovery process could be passive network discovery where the mobile listens for network announcements or active network discovery where the mobile solicits network announcements. The type of discovery process varies for each type of network. In a GSM network, the mobile discovers a new Location Area, whereas a GPRS network includes discovery of a new Routing Area. In case of IP-based network, the discovery process can span across all the layers, and include cell, subnet or domain discovery. Each layer 2 access provides different ways of discovering the networks and resources. For example, GSM uses the BCCH (Broadcast Control Channel), CDMA uses a pilot channel, and 802.11 uses active and passive scanning to discover the new point-of-attachment. Similarly, foreign agent advertisement and router advertisement can help discover the new point of attachment at layer 3, such as a router or MIP Foreign Agent.

C. Network Selection

Network selection is a process by which a mobile node or a network entity analyzes the information discovered about its neighboring networks, and then selects a network to connect to. The selection may be based on criteria such as required QoS, cost, and user preferences. Thus, an appropriate selection mechanism helps the overall resource optimization process and increases the probability of successful handoff.

D. Network Attachment

After the mobile has selected the target network, it attempts to connect to the new network point of attachment. Sudden lapse of an existing connection or the availability of the new point of attachment is usually sent as event notification trigger to the upper layers to expedite any further handoff related functions. The IETF is currently working on standardizing a protocol for Detection of Network Attachment (DNA) [11] that involves mechanisms both at layer 2 and layer 3, that can notify the upper layer about the detection of a new network attachment. The IEEE 802.21 working group is currently defining several event service primitives such as "Link Up", and "Link Down" that can be used to provide the status of lower layer events to expedite the handoff process. For example, since layer 2 association takes place before any upper layer operations, it helps to send layer 2 trigger such as "Link up" to execute the upper layer mobility functions such as layer 3 configuration. Signal-to-Noise Ratio (SNR) threshold is one such possible layer 2 trigger event.

E. Configuration

During the configuration phase, the mobile connects to the new point of attachment in the network and establishes the mapping of its connection identifier with the appropriate network entity in the network. The configuration phase can be divided into the following sub-phases:

Identifier Configuration: This process allows a mobile to configure a new temporary connection identifier either at layer 2 or at layer 3 at the new point of attachment. As part of the identifier configuration process, the mobile obtains a new identifier, tests its uniqueness, and finally, assigns it to the mobile's interface. As an example, it configures Care-of-Address (CoA) in case of IP environment or TMSI (Temporary Mobile Subscriber Identity) in case of GSM. In general, completion of these processes involves a series of signaling messages between the mobile and the server in the network contributing to the overall handoff delay.

Registration: Registration establishes the mapping between the permanent locator such as SIP URI or home address and the temporary identifier such as care-of-address for proper location management function. An optimized or hierarchical registration process helps to expedite location management and speeds up the delivery of the new traffic.

Authentication and Authorization: The authentication process allows the mobile to get access to the network resources. During handoff, the re-authentication process involves a signaling exchange between the mobile and the authentication server in the network. In case of GSM, it uses the SRES and A3 algorithm for the authentication. For 802.11 access networks, the mobile could use open system authentication, shared authentication such as WEP or stronger authentication such as 802.11i in layer 2. Fathi et al [12] demonstrate how different authentication mechanisms affect the layer 2 access delay. Similarly, layer 3 authentication protocol such as PANA (Protocol for carrying Authentication to Network Access) add delay during layer 3 authentication. Georgiades [13] shows that it takes up to 4 seconds to complete both authentication and authorization process. These operations can use a combination of EAP (Extensible Authentication Protocol) [14] over layer 2 for IEEE 802.1x-based authentication and EAP-TLS (Transport layer Security) [15] for layer 3 authentication. For inter-domain mobility case, the authentication process is typically followed by an authorization process that involves interaction with authorization server, adding further delay.

Security Association: Security association can be defined as a secure binding between two endpoints that applies security policy and keys to protect information. Before a new communication path is established between the end-points, the mobile node needs to authenticate itself, and establish a security association with other network nodes. The security association may take place at several layers of the protocol stack. Establishing a security association involves a key distribution procedure that includes exchange of messages

between the mobile and any centralized security server. The security association in an IP-based environment can be uniquely identified by a tuple consisting of a Security Parameter Index (SPI), an IP destination address, and a security protocol (AH or ESP) identifier as defined in ISAKMP (Internet Security Association and Key Management Protocol) [16].

Encryption: After a mobile is authenticated to the network, protection for data and signaling is provided by means of encryption. For IP-based mobility, encryption can take place at different layer such as WEP (Wired Equivalent Privacy) or WPA (Wi-Fi Protected Access) at layer 2, IPSEC/ESP (Encapsulating Security Payload) at layer 3 and SRTP (Secured RTP) at layer 4. Similarly GSM uses SRES algorithm to derive the encryption key where as IS-41 uses CAVE algorithm to derive the encryption key. The process of encryption and decryption adds to the one-way transmission delay for the packet and packet loss [16]. Extent of delay and packet loss depends upon the type of encryption algorithm used. The process of obtaining the new key to encrypt the data and signaling after the handoff, adds delay to the normal data transfer process. Also, if the encryption algorithm in the new access network is different than the previous network, the encryption and decryption process contribute to the added delay.

F. Media Redirection

A media redirection completes the handover process. This phase involves re-routing of the media so that the delivery of data from the old connection path to the new connection path is re-established according to the pre-defined service guarantee. Following are some operations that constitute the media redirection phase:

Binding Update: Binding update is the process by which a mobile updates its new network identifier so that the data in flight can be rerouted to the new destination. As the mobile connects to a new point of attachment and obtains a new temporary network identifier (e.g., TMSI in GSM, COA in MIPv6, FA-COA in MIPv4) in the new network, it needs to update the communicating host to reroute the packet to the new destination. This process associates the new network identifier with the permanent identifier of the mobile. Until the re-association of the new identifier is complete, the data in transit still goes to the old network and is considered lost in the absence of any optimization mechanism, such as buffering or packet forwarding. In some cases, this binding update needs to be authenticated. For example, MIPv6 requires a return routability procedure and adds two additional messages such as CTI (Care-of-test Init) and HTI (Home-test Init) to obtain the binding key so that binding update can be authenticated as well. This delays the binding update procedure.

Media Rerouting: Once the binding update is complete, data from the correspondent node gets routed to the mobile's new location. Media redirection process includes several elementary operations such as encapsulation, decapsulation,

tunneling, buffering, and copy-forwarding. During the media re-routing process, data in flight may get lost or may get delayed because of these operations. Thus, there is a need to investigate the ways of optimizing these operations to reduce media delivery delay. Certain operation such as buffering adds delay to the packet traversal, but helps to reduce the packet loss. Thus, the buffering delay needs to be adjusted to compromise between packet loss and one-way packet delay. Table 1 shows how the basic handoff functions are taken care of at each of the layers in an IP-based handoff environment. Based on the type of handover one or more layers get affected.

Table 1: Handoff operation in layers 2, 3 & 7

Handoff Operations	Layer 2	Layer 3	Application Layer
Discovery	Scanning	Router advertisement	Domain Advertisement
Authentication	EAPoL	IKE, PANA	S/MIME
Security Association	802.11i	IPSEC	TLS SRTP
Configuration	ESSID	DHCP Stateless	URI
Address Uniqueness	MAC Address	ARP DAD	SIP Registration
Binding Update	Cache Update	Update CN, HA	SIP Re-INVITE
Media Routing	IAPP	Encapsulation Tunneling	Direct media routing

IV. MOBILITY MODELING

A mobility model provides a more abstract representation of handoff events potentially enabling exploration of new optimizations. In this section, we develop the mobility systems model by analyzing the state transition associated with each layer, and represent each transition using a set of finite state machines. The mobility event is viewed as a perturbation to the steady state of a communicating node. As a communicating node is subjected to a mobility event, it goes through a set of intermediary states before it attains the steady state again by returning to the connected state. All sets of atomic operations described in Section III are performed during the state transition. Thus, it can be viewed as a Discrete Event Dynamic Systems (DEDS). We use Deterministic Time Petri nets [18] to model the mobility event. Depending upon the type of mobility, each layer in the protocol stack gets affected and the mobile goes through a series of similar transition processes within each layer. The mobile's communication is interrupted because of the delay associated with each transition. Since most of these state transitions during the mobility event appear to be sequential in nature, the mobility system exhibits similar behavior often observed by the Flexible Manufacturing System. Zuberek et al [19] model and analyze simple schedules for manufacturing cells. Mobility systems model can use similar techniques for performance analysis for handoff events. We model a mobility event as a series of sequential states, where there are several sub-events within each of these states. In case of series of consecutive handoffs, it can be modeled as a cyclic event. Figure 1 shows the high level state machines associated with a mobility event using a Timed Petri net

approach [16]. Each place (P) represents various stages of the mobility event and the transition (t) represents the time taken due to different set of operations between the stages.

The Petri net model representing the general mobility systems is actually a decision free Petri net, where a minimum cycle time is an indicator of maximum performance. The cycle time is represented as $C = \max T_k/N_k$; $k=1,2,3...q$, where T_k = sum of the execution times of the transmissions in circuit k and N_k is the total number of tokens in the places in circuit k and q is the number of circuits in the net. In case of a systems model involving mobility event, these values can vary depending upon the number of transitions involved in a cycle. For example, Figure 1 also introduces a new place as an alternate transition path that provides a faster transition to the connected state by allowing local buffering and local media redirection. Introduction of new place such as P7 shows how one can take advantage of Peteri net's lateral fusion reduction rule to represent this specific optimization method. Table 2 shows respective cycle time for different optimization techniques applied to Fig. 1.

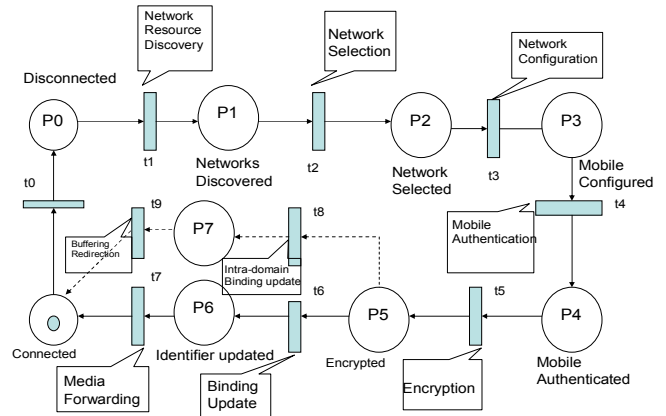


Figure 1: Petri Net model for handoff event

Table 2: Cycle time for various optimization

Type of Optimization	Loops in the state transition	D_i Execution Time	N_i Tokens	D_i/N_i Cycle Time
No Optimization	p0t1p1t2p2t3p3t4 p4t5p5t6p6t7p10	24x	1	24x
Limit Binding update	p0t1p1t2p2t3p3t4 p4 t5p5t8p7t9p10	19x	1	19x
Proactive	p0t9p10	2x	1	2x
Maintain Security binding	p0t1p1t2p2t3 p5t6p6t7p10	19x	1	19x

Figure 2 shows Petri net representation and component level state transitions at each layer. Several sub-processes in each of the layers, such as scanning, layer 2 authentication, layer 3 configuration, layer 3 authentication, encryption, and binding update are illustrated here. A mobile is subjected to different

amount of handoff delay based on the types of handover it is subjected to. For example, an intra-subnet, intra-technology handover takes much less time than the inter-subnet and inter-technology handover, because the mobile goes through a smaller number state transitions during the handoff process. Parallelization of state transitions within each layer, reduction of number of states, and proactive operation of the sequential events associated with the handoff are some techniques for achieving optimization. In addition, the mobility model can also evaluate the effectiveness of any mobility protocol. We apply this mobility model and validate the handoff properties by way of experimental results in a testbed. We use IETF-based protocols, such as DHCP, SIP, MIP, IPsec, and PPPoE to implement these functions. Figure 3 shows the breakdown of these handoff components and categorize the delays associated with layer 2, layer 3, and application layer operations.

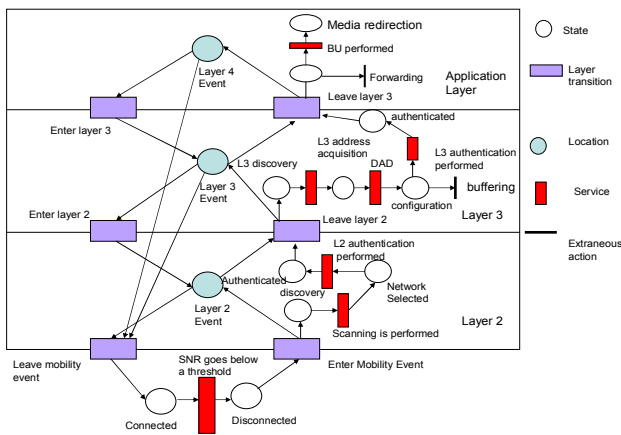


Figure 2: Layered approach to Petri Net model

For example, the delays associated with layer 2, layer 3, discovery, address acquisition, SIP server discovery, registration, binding update, and media redirection are shown

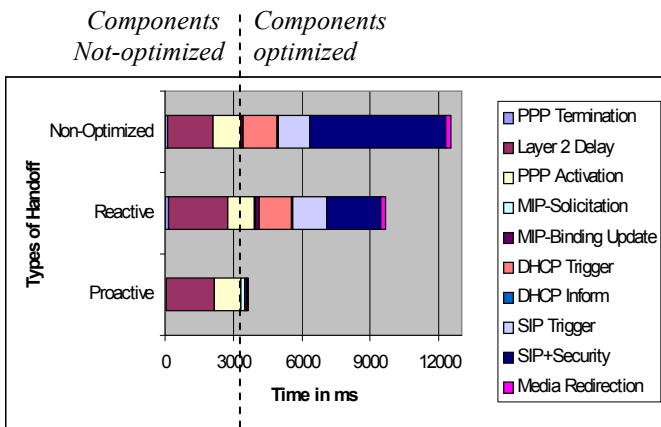


Figure 3: Experimental results of handoff components

for each type of optimization categories. As separated by dotted line in Figure 3, we have not applied any optimization technique to layer 2 related functions in the experiment. In proactive handoff mode of the experiment, many of the handoff functions such as security association and server

discovery operations are performed prior to handoff or are done in parallel, whereas in non-optimized case, the handoff operations are done in a sequential manner. Reactive mode uses context transfer technique to reduce the delay. Thus, the handoff delays are reduced for proactive and reactive cases. These handoff delay components for different operations can be co-related with the cycle time resulting out of series of transitions from one state to another state (e.g., t1, t2, t3, t4, t5, t6, t7, t8, t9) in the Petri Net model shown in Figure 1.

V. CONCLUSION

During the process of handoff from one network to another network, the mobile goes through a sequence of operations that cause delay and packet loss. Existing mobility protocols provide their own optimization techniques to attain the desired level of quality of service. Thus, the existing optimization solutions are very much ad hoc in nature and lack a set of formal methodologies. A thorough analysis of these handoff related operations and generalized model for handoff management introduced in this paper can provide a systematic approach to identify the required handoff components and design the rules of optimization that could be applied based on application and mobile's movement pattern.

REFERENCES

- [1] N. Tripathi, J. Reed, and H. F. Van Landingham, Handoff in cellular systems, *IEEE Personal Communications Magazine*, vol. 5, Dec. 1998.
- [2] C. E. Perkins, IP mobility support for IPv4, RFC 3344, Internet Engineering Task Force, Aug. 2002.
- [3] D. B. Johnson, C. E. Perkins, and J. Arkko, Mobility support in IPv6, RFC 3775, Internet Engineering Task Force, June 2004.
- [4] H. Schulzrinne and E. Wedlund, "Application Layer Mobility using SIP", *ACM MC2R*, vol 4, July 2000.
- [5] A. Dutta et al, "Fast handover Schemes for Application Layer Mobility Management," in *Proc. IEEE PIMRC* September 2004, Barcelona, Spain.
- [6] R. Koodli et al, "Fast Handover for Mobile IPv6," RFC 4068, July 2005
- [7] H. Soliman et al, "Hierarchical Mobile IPv6 Mobility Management," RFC 4140, IETF, August 2005.
- [8] M. Marshan et al, "International Workshop on Petri net and Performance Models," *On Petri Net-based Modeling Paradigms for the Performance Analysis of Wireless Internet Access*, 2001.
- [9] R. M. Amadio and S. Prasad, Modeling IP mobility, tech. rep., INRIA, Sophia Antipolis, 1997.
- [10] A. Dutta et al, "Experimental Analysis of Multi Interface Mobility Management with SIP and MIP," *IEEE Conference on Wireless Networks, Communication, and Mobile Computing* Vol 2, June 2005, Maui, HI
- [11] R. Johnson et al, Goals of detecting network attachment in IPv6, RFC 4135, IETF, Aug. 2005.
- [12] H. Fathi et al, "On the impact of Security on the Latency in WLAN 802.11b: Analytical Study," *IEEE Globecom*, 2005
- [13] M. Georgiades, Context transfer support for IP-based mobility management, tech. rep. 2004.
- [14] B. Aboba et al, "Extensible Authentication Protocol," IETF RFC 3748, June 2004.
- [15] B. Aboba and D. Simon, "PPP EAP TLS Authentication," IETF RFC 2716, Oct 1999.
- [16] D. Maughan et al, "Internet Security Association and Key Management Protocol," IETF RFC 2408, Nov. 1998.
- [17] H. Xiao and P. Zarella, "Quality effects of wireless VoIP using security solutions," *IEEE MILCOM* 2004, CA.
- [18] T. Murata, Petri nets: Properties, analysis and applications, *Proceedings of IEEE*, vol 77, issue 4, Apr. 1989
- [19] W.M. Zuberek and W. Kubiak, Timed petri nets in modeling and analysis of simple schedules for manufacturing cells, *Elsevier Science Journal, Computers and Mathematics with Applications*, Aug. 1999.