

## Adaptive Slot Allocation to Control Queuing Delay in TDMA Wireless Base Station\*

MONG-FONG HORNG<sup>1,3</sup>, YAU-HWANG KUO<sup>1,2</sup>, JANG-PONG HSU<sup>4</sup>  
AND REN-HAO CHENG<sup>4</sup>

<sup>1</sup>*Center for Research of E-life Digital Technologies*

<sup>2</sup>*Department of Computer Science and Information Engineering  
National Cheng Kung University  
Tainan, 701 Taiwan*

<sup>3</sup>*Department of Computer Science and Information Engineering  
Shu-Te University  
Kaohsiung, 824 Taiwan*

<sup>4</sup>*Advanced Multimedia Internet Technology Inc.  
Tainan, 710 Taiwan*

An Adaptive Slot Allocation (ASA) scheme for controlling the queuing delay of packet delivery in a TDMA base station is presented in this paper. ASA utilizes the multi-queue architecture in a base station to support differentiated services for mobile hosts. The services required by each host are divided into quality-guaranteed type and best-effort type, which are served by separate queues. Another mechanism to realize service differentiation is the scaling factors used to differentially affect the determination of the outgoing rate for each queue in ASA. Also, the queue status is used as another parameter to determine the outgoing rate so that the adaptation of ASA can be achieved. Based on the allocated data rate and real channel quality, ASA further allocates time slots among the mobile hosts to control their packet delay. Such an adaptive slot allocation scheme can track the variation of input traffic and channel capacity adequately but still controls the queuing delay of target packets.

In this analysis, we illustrate how the proposed ASA controls the queue dynamics. Then we investigate the relationship between queue dynamics and queuing delay. As a result, we can conclude that the queuing delay of services for each host is effectively controlled by the parameters of the ASA queuing model. Moreover, the controllability of the service queues relieves the task of buffer management in a base station. A simulation of real streaming traffic traveling across hops is made to evaluate the queue dynamics and delay performance of ASA. The simulation results confirm the expected properties, even under heavy traffic variation.

**Keywords:** wireless TDMA base station, delay controllability, quality-of-service, adaptive slot allocation, queue dynamics

### 1. INTRODUCTION

Nowadays, development of quality-guaranteed services over the Internet has attracted much research interest. The objective of Quality-of-Service (QoS) technology is to offer network services with different service level agreements (SLAs) through a dif-

---

Received December 25, 2002; revised April 29, 2003; accepted August 25, 2003.

Communicated by Yu-Chee Tseng.

\* The partial work of this paper was presented in the IEEE International Symposium on Computer Communication, Antalya, Turkey, 2003.

ferentiated service mechanism. The service level agreement results from the negotiation between user and network, in which the quality guarantees include packet delay, packet loss and packet jitter. According to the negotiated SLA, the routers on the delivery path allocate network resources such as bandwidth, time slots and packet buffer to accomplish the guarantee of network quality.

The QoS problem becomes important when multimedia services are involved. Multimedia data transmission has some significant features including in-time delivery, higher tolerance to packet loss, more restricted buffer management and low tolerance to packet jitter. These phenomena of multimedia traffic demand a sophisticated QoS mechanism to ensure the availability of resources of various service classes.

The popularity of wireless communication adds new complexity to the QoS problem, where the frequent variation of a wireless link results in the inefficiency of bandwidth control. Two factors cause the variation of wireless links: radio characteristics and user mobility. The channels of a wireless network operate at the Super High Frequency (SHF). Because of the properties of line-of-sight in SHF, channel fading, multipath effect and radio interference [1-3], the channel characteristics depend on the locations of the mobile host and base station, as well as their distance. The change of location caused by host mobility causes further variation of the channel characteristics. Moreover, the link conditions of mobile hosts to a base station are not only time varying but also different from each other. Therefore, to support quality-guaranteed wireless multimedia services, we need a more efficient scheduling mechanism of the base station to allocate resources so that the QoS requirements promised to mobile hosts are accomplished as well as possible.

The third-generation code-division multiple access (CDMA) high data rate (HDR) system is the most promising technology for wide-area wireless multimedia networks. In this kind of system, the mobile hosts in the same cell share the same CDMA channel to connect the base station. The downlink from the base station to the hosts adopts time division multiple access (TDMA), where time is divided into fixed-size time slots to serve the hosts in a cell. The base station allocates the time slots for the served mobile hosts to meet their QoS requirements. Andrews *et al.* [4, 5] proposed a throughput-optimized scheduling approach for a shared wireless link with variable channel conditions. Mirhakkak *et al.* [6] suggested an approach using dynamic adaptation of QoS levels in which applications can specify their QoS needs as ranges rather than the scalar values and tolerate transient periods of degraded service in a wireless network. In controlling resource utilization, queue dynamics is often used to compare with the defined threshold so that the targeted traffic conditions are aware. Hahne and Choudhury [7] established a threshold mechanism with multiple loss priorities to solve the memory-sharing problem in packet switching. The works presented in [8, 9] use a threshold to detect queue state and to allocate the bandwidth of various service classes. Choosing the best threshold value is difficult. Although the concept of dynamic thresholds [7-9] has been introduced, the calculation of thresholds is time-consuming and complicated. In this paper, we focus on the development of an efficient approach for TDMA data scheduling. Instead of using a threshold, we adopt the queue occupancy as the state variable to adapt the bandwidth and slot allocation in the downlink from the base station. The queue status in the base station is used as a feedback to determine the required data rate of mobile hosts so that their queueing delay is under control.

Some frameworks [4, 10-15] have been proposed to realize the QoS objective on wired and wireless networks. In these frameworks, prioritization is the fundamental issue to construct a QoS mechanism and to offer the functionality of service differentiation. Most frames follow the process of classification, queuing and scheduling to realize the prioritization. All received packets are classified according to some pre-defined criteria such as packet source/destination, packet type and data type. A set of packet buffers, called service class queues, keeps the classified packets until the packet scheduler sends them out. Therefore, the architecture of multiple queues becomes one of the fundamentals of differentiated networks for handling incoming packets with their desired service quality. However, this architecture raises the issue of how to allocate each queue with an adequate bandwidth, that is, how to determine the delivery order and bandwidth sharing such that the queuing delay and loss of each class queue are guaranteed. The investigation presented in this paper also adopts a multi-queue architecture as the queuing mechanism for multimedia services in a TDMA base station, and then develops a packet scheduler with delay-controllable slot allocation capability.

The remainder of this paper is organized as follows. Section 2 describes the operational model of the proposed delay-controllable slot allocation scheme. Section 3 analyzes the properties of the proposed model and the effects of the operational parameters. Section 4 illustrates and analyzes the simulation results for real multimedia traffic. Finally, the conclusion is given in section 5.

## **2. MODEL OF PACKET DELAY-CONTROLLABLE SLOT ALLOCATION SCHEME**

Delay control is the major target of this investigation. End-to-End (E2E) packet delay is defined as the time taken by a packet traveling from the source host to the destination host, and is usually composed of propagation delay, processing delay and queuing delay. Propagation delay is caused by the propagation of packet signals over transmission media. Processing delay is the time spent by networking devices (e.g., routers), to determine how to forward packet, for example, packet routing and packet filtering. Queuing delay is the waiting time of packets stored in networking devices before being transmitted.

Propagation delay of a packet depends on transmission media and packet length. For example, the propagation delay of packets carried on an optical link is usually faster than on electrical link. Processing delay of packet depends on the computation power of networking devices and the applied protocol stack. Low efficiency of interpreting the protocol header in a router degrades the processing speed. The length and hop count of delivery path are two other factors affecting propagation and processing delays. But, for a specific delivery route, the propagation and processing delays of each packet are almost invariant. Only the queuing delay heavily depends on the traffic conditions of the traveling route.

In the past, Best-Effort (BE) service is the major kind of Internet services, where all packets in a router compete for being delivered in a First-In-First-Out (FIFO) manner. As a result, the queuing delay is not controllable and thus varies. This scheme leads to no guarantee or control on the delivery quality, particularly when a link congestion occurs. Therefore, we need to improve the controllability of queuing delay of packets to help

control of the E2E delay. In this section, we present an adaptive slot allocation (ASA) scheme to well control the queuing delay of packets forwarded by a TDMA wireless base station so that the QoS requirement is satisfied. In a wireless base station, slot allocation becomes complicated because of the varying channel capacity. How to allocate time slot among the mobile hosts with different and time-variant channel capacity is the focus of our investigation.

## 2.1 Queue Architecture of ASA Scheme

Basically, the proposed ASA scheme adopts a two-class queuing mechanism in the base station for each mobile host, where the two queues are denoted as the Guaranteed-Service (GS) class and Best-Effort (BE) class. During delivery, the GS queue has higher priority than the BE queue. Because each mobile host may have a different link condition and QoS requirement, the base station constructs individual GS for each registered mobile host. However, all mobile hosts share the same BE queue. The BE class queue will not be served until all GS queues have been served. In summary, ASA applies a multiclass queuing architecture as shown in Fig. 1.

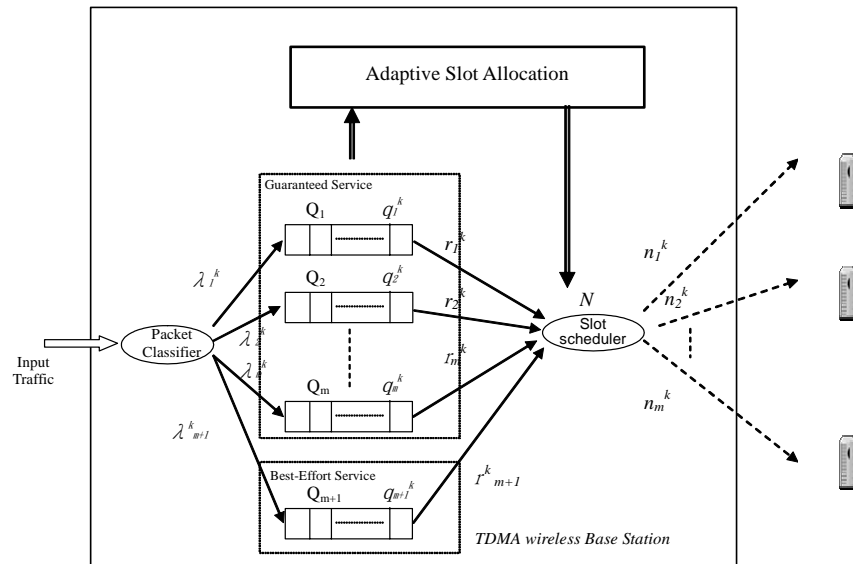


Fig. 1. The multiclass queuing architecture utilized by the ASA scheme.

## 2.2 Operational Model of ASA Scheme

Since multiclass queuing is the basic component part of differentiated service for realizing QoS-enabled networks, the slot allocation of each class queue directly impacts the scheduling performance and the delivery quality in a base station. The packet scheduler serves all class queues in a round robin way. In each service round, a four-phase procedure is performed to realize the ASA-based scheduling for the GS queues:

**Phase I:** Allocate the data rate of each GS queue  $Q_i$  at the  $k^{\text{th}}$  round according to

$$r_i^k = \frac{\omega_i q_i^k}{T} \quad (1)$$

where  $r_i^k$ ,  $\omega_i$ ,  $q_i^k$  and  $T$  are the allocated data rate, scaling factor and queue length of  $Q_i$  at the  $k^{\text{th}}$  round and service round time, respectively.

**Phase II:** Calculate the required number of time slots of each GS queue  $Q_i$  for the allocated data rate by considering the link capacity:

$$n_i^k = \begin{cases} \frac{r_i^k T}{C_i^k t} & \text{if } C_i^k > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where  $n_i^k$ ,  $C_i^k$  and  $t$  are the required number of time slots of  $Q_i$ , the link capacity between the mobile host corresponding to  $Q_i$  and base station, as well as the period of time slots, respectively. If  $\sum_{i=1}^m n_i^k \leq N$ , where  $N$  denotes the total number of available time slots ( $= T/t$ ), go to Phase IV directly. Otherwise, go to next phase.

**Phase III:** Reallocate the time slots to each GS queue as follows:

```

/* N : total number of time slots
/* N_unused : the number of unused time slots
/* n_i : the number of time slots required by Q_i
/* n_hat_i : the number of time slot actually allocated for Q_i
/* m : number of GS queues
sort {n_i} in an increasing order
reset variable n_hat_i
N_unused = N
for (j = 1; j <= m; j++)
{
if N_unused > (m + j - 1) * Min({n_i, j <= i <= m})
{
n_hat_i = n_hat_i + Min({n_i, j <= i <= m}) for j <= i <= m
N_unused = N_unused - (m + j - 1) * Min({n_i, j <= i <= m})
}
}
else
{
n_hat_i = n_hat_i + N_unused / (m - j + 1)
N_unused = 0
}
}

```

**Phase IV:** Deliver the packets in GS queues according to the result of time slot allocation in Phase II or Phase III.

In the procedure described above, the ASA approach uses queue dynamics and a set of scaling factors to control the delivery rate of each queue. The scaling factor is adopted to reflect the service level agreement. For example, a larger scaling factor will be specified when better delay performance is required. Then, the number of required time slots is obtained from the factors of allocated data rate, service round time, slot time period and channel capacity. If the number of available time slots is less than the total allocated time slots, the time slots are reallocated to meet two objectives: guarantee of minimum bandwidth and fair bandwidth sharing. Thus, in Phase III, the accumulated time slots of all queues grow together from zero at the same pace. The procedure progressively fills the time slot requirements of queues and is terminated when the unused slots are exhausted. This filling procedure obviously offers fair slot allocation for all queues and ensures the queues allocated with a minimum number of time slots are optimal.

In ASA, the scaling factor  $\omega_i \in (0, 1]$  associated with each queue is a user-supplied operational parameter to control the queue dynamics. If  $\omega_i$  is equal to zero, the associated queue is prohibited from delivering data. On the contrary, the queue with  $\omega_i = 1$  will be served with a full delivery of data in each service round. Between the two extremes, according to Eq. (1), the delivered byte volume in a service round is proportional to the occupancy of each queue. The queue occupancy is related to the waiting time spent by the packets in a queue. Thus, for any queue, if the occupancy is regulated through the adaptation of bandwidth allocation, the queueing delay of each queue would be well controlled. The detailed analysis of the relationship among queue occupancy, bandwidth allocation and queueing delay will be illustrated in the following sections. It is noted that an empty queue shares no bandwidth or time slot. ASA suspends the bandwidth allocation for an inactive connection, so that the utilization of bandwidth is greatly improved in comparison with schedulers using constant bandwidth reservation.

### 3. THEORETICAL ANALYSIS OF ASA SCHEME

To perform qualitative and quantitative analyses, we assume that ASA is applied to the TDMA base station under the following conditions:

1. The channel capacity of connecting each mobile host is available to the base station in each round.
2. During a service round, the channel capacity remains constant.
3. The computation time needed for slot allocation is ignored.
4. The overhead time taken in service switching between GS queues (that is, mobile hosts) is ignored.
5. The packets in a queue will not be lost except for the failure of the channel to mobile host; in this case, the queue is removed.
6. The admission control is applied before establishing a connection to ensure the sufficiency of network resources at the base station.

Based on the assumptions mentioned above, we first analyze the queue dynamics and convergence of ASA under normal conditions, in which the time slots of each service round are sufficient to support the required time slots of all queues in the TDMA base station. In normal operation of ASA, we will show that the queue occupancy is explicitly formulated as a function of the scaling factor, service round time and input traffic rate. Second, the convergence of bandwidth allocation of ASA is analyzed. Both random traffic and constant traffic are of concern. Finally, the controllability of queueing delay in ASA is proven. When the queue is at the steady state, the scaling factor and service round time concisely control the delay of each packet in the cases of constant input traffic or slightly varying traffic. Even when the input traffic is regarded as random, the average delay of packets is guaranteed by statistics. In summary, the mean of the queueing delay is independent of the queue occupancy when the queue is at the steady state.

### 3.1 Queue Dynamics and Convergence of ASA

Based on the fluid model [16], the queue dynamics of ASA is modeled as

$$q_i^{k+1} = q_i^k + (\lambda_i^k - r_i^k)T \quad (3)$$

where  $\lambda_i^k$  is the data arrival rate of  $Q_i$  at the end of the  $k^{\text{th}}$  service round. Let us start the performance analysis from the queue state. The convergence property of ASA is stated as follows:

**Theorem 1** In ASA, for a given service round time  $T$ , scaling factor  $\omega_i$  and input rate  $\lambda_i$ , the state of each queue  $Q_i$  is obtained by

$$q_i^k = \sum_{m=1}^k (1 - \omega_i)^{m-1} T \lambda_i^{k-m} \quad (4)$$

**Proof:** Applying ASA bandwidth allocation formula of Eqs. (1) to (3), we have

$$q_i^{k+1} = q_i^k + (\lambda_i^k - \frac{\omega_i q_i^k}{T})T \quad (5)$$

Z-transforming Eq. (5) yields

$$q_i(z) = \frac{\lambda_i(z)T}{(z-1+\omega_i)} \quad (6)$$

where  $q_i(z)$  and  $\lambda_i(z)$  are the z-transforms of  $q_i^k$  and  $\lambda_i^k$ , respectively. To obtain the time sequence of  $q_i^k$ , we apply the inverse z-transform on Eq. (6) and

$$q_i^k = Z^{-1}[q_i(z)] = Z^{-1}\left[\frac{\lambda_i(z)T}{(z-1+\omega_i)}\right] \quad (7)$$

Expanding the division of Eq. (7), we get

$$\begin{aligned}
q_i^k &= T\lambda_i^{k-1} + (1-\omega_i)T\lambda_i^{k-2} + (1-\omega_i)^2T\lambda_i^{k-3} + \dots + (1-\omega_i)^{k-1}T\lambda_i^0 \\
&= \sum_{m=1}^k (1-\omega_i)^{m-1}T\lambda_i^{k-m} \quad \square
\end{aligned} \tag{8}$$

Eq. (4) illustrates the dynamics of queue  $Q_i$  in ASA, which is related to the input rate, service round time and its scaling factor.  $T\lambda^{k-m}$  is the number of bytes received at the  $(k-m)^{\text{th}}$  round. Therefore, the queue state of  $q_i^k$  is regarded as the weighted average of the input traffic rate at all past rounds. The term  $(1-\omega_i)^{m-1}$  converges to zero for any  $\omega \in (0, 1]$ . Moreover, the convergence of  $q_i^k$  depends on the value of  $\omega$ , the queue state depends on shorter traffic history when  $\omega$  is close to 1. On the other hand, as  $\omega$  decreases to near zero, the queue state depends on a longer range of traffic history.

Next, we discuss the convergence of the queue state in ASA, for which a queue reaches and stays within a specific zone around the final state. Such a zone is specified by a percentage value of the final state to indicate the tolerance of a steady state.

**Corollary 1** In ASA, if the input rate of queue  $Q_i$  is constant, then it will maintain a steady state whose value is given by

$$q_i^{ss} = \frac{\lambda_i^{ss}T}{\omega_i} \tag{9}$$

**Proof:** According to Eq. (4), when a queue  $Q_i$  has a constant input rate,  $\lambda_i^{ss}$ , the queue state can be written as

$$\begin{aligned}
q_i^k &= T\lambda_i^{ss}(1 + (1-\omega_i) + (1-\omega_i)^2 + \dots + (1-\omega_i)^{k-1}) \\
&= \frac{T\lambda_i^{ss}(1 - (1-\omega_i)^k)}{1 - (1-\omega_i)} \\
&= \frac{T\lambda_i^{ss}(1 - (1-\omega_i)^k)}{\omega_i}
\end{aligned} \tag{10}$$

Moreover, since  $\omega_i \in (0, 1]$ , the term  $(1 - (1-\omega_i)^k)$  approaches 1 as  $k$  increases with time. Consequently, we have  $q_i^{k+1} = q_i^k = T\lambda_i^{ss} / \omega_i$  when  $q_i$  exponentially approaches the steady state. That is, the queue  $Q_i$  reaches a steady state whose value is determined by the input rate, service round time and scaling factor. When  $Q_i$  has the scaling factor  $\omega_i = 1$ , its queue state reaches a steady state of  $T\lambda_i^k$  in a service round.  $\square$

From Corollary 1, it is obvious that ASA has some advantages in buffer management since we just need to control two coefficients. If the input traffic is generalized to a random process with varying rate, ASA also exhibits good statistical properties.

**Corollary 2** In ASA, if the input rate  $\lambda_i$  is a random process with a distribution function  $P_\lambda$  and its mean is finite, then the average queue state of  $Q_i$  with  $\omega_i = 1$  is equal to  $E[\lambda_i]T$ , otherwise the average queue state of  $Q_i$  approximates to  $E[\lambda_i]T/\omega_i$ .

**Proof:** From the generic dynamics model of class queues given by Eq. (4), to derive the mean of the queue state  $Q_i$ , we apply the mean operation to both sides of Eq. (4) to yield

$$E[q_i] = E[q_i^k] = E\left[\sum_{m=1}^k (1-\omega_i)^{m-1} T \lambda_i^{k-m}\right] \quad (11)$$

In Eq. (11), except for the input rate, both the service round time and scaling factor are constants. Thus, if the mean input rate is finite, we finally have the approximation:

$$\begin{aligned} E[q_i] &= E\left[\sum_{m=1}^k (1-\omega_i)^{m-1} T \lambda_i^{k-m}\right] \\ &= E(\lambda_i) T \sum_{m=1}^k (1-\omega_i)^{m-1} \\ &= E(\lambda_i) T \frac{1-(1-\omega_i)^k}{\omega_i} \end{aligned} \quad (12)$$

Obviously, when the scaling factor is equal to one, the queue state arrives exactly at  $E[\lambda_i]T$ , regardless of the service round index,  $k$ . Otherwise, the queue state of  $Q_i$  exponentially approaches the steady state,  $E[\lambda_i]T/\omega_i$ , given by Eq. (12).  $\square$

Therefore, in the case of random traffic, ASA still offers a guarantee of queue dynamics controlled by the service round time and the scaling factor. In other words, ASA make buffer management easy. In Theorem 1 and associated corollaries, the queue dynamics and steady states in various conditions are specified. In the following, we analyze the convergence time from initial queue state to the steady state. When the traffic rate is a random process, because of the traffic locality, the traffic can be approximated as lots of segments of piece-wise constant traffic. Usually, the duration of each segment is larger than the service round time of ASA. If the input rate is a piece-wise constant,  $\lambda_i^k$ , in the short interval  $(t_0, t_0 + \tau)$ , combining Eqs. (1) and (4), we have

$$r_i^k = (1-(1-\omega_i)^k) \lambda_i^k \quad \text{for all } k \in \left[\left\lceil \frac{t_0}{T} \right\rceil, \left\lceil \frac{t_0+\tau}{T} \right\rceil\right] \quad (13)$$

By the allocation of outgoing data rate according to Eq. (13), the queue  $Q_i$  converges exponentially to the steady state in the duration  $\tau$ . According to Eq. (10), the convergence of  $Q_i$  to steady state is determined by  $\omega_i$  and  $T$ . Table 1 shows the required time of ASA converging to the steady state under different  $\omega_i$ . For the cases of  $\omega_i = 0.5$ , the

**Table 1. Convergence time of ASA for different scaling factors and state tolerances.**

Converge to steady state	$\omega = 0.01$	$\omega = 0.1$	$\omega = 0.2$	$\omega = 0.5$	$\omega = 1$
Within 1%	418T	41T	20T	8T	1T
Within 10%	225T	23T	12T	6T	1T

steady state within  $\pm 1\%$  tolerance is reached within at most six service rounds. Thus, if the service round time  $T$  is sufficiently smaller than the interval  $\tau$ , the allocation of  $r_i^k$  according to Eq. (1) ensures that  $Q_i$  enters the steady state.

### 3.2 Convergence of Bandwidth Allocation of ASA

The other significant feature of ASA is its capability of real-time bandwidth adaptation to track the instantaneous variation in input rate. This property is illustrated by the following theorem.

**Theorem 2** In the case of random traffic, for any steady queue with nonzero scaling factor in ASA, the allocated bandwidth  $r_i^{ss}$  approximates the input rate  $\lambda_i^{ss}$  in its statistics.

*Proof:* Any steady queue in ASA has a queue state given by Eq. (13). Also, according to Eq. (1), the scheduler allocates bandwidth to each queue  $Q_i$  in random traffic in the amount of  $r_i^k = \omega_i q_i^k / T$ . Combining both equations, we have

$$\begin{aligned} r_i^k &= \omega_i (\lambda_i^{k-1} + (1-\omega_i)\lambda_i^{k-2} + (1-\omega_i)^2\lambda_i^{k-3} + \dots + (1-\omega_i)^{k-1}\lambda_i^0) \\ &= \omega_i \sum_{m=1}^k (1-\omega_i)^{m-1} \lambda_i^{k-m} \end{aligned} \quad (14)$$

Since  $\lambda_i$  is a random variable characterized by a random process, we have

$$\begin{aligned} E[r_i^k] &= E\left[\frac{\omega_i q_i}{T}\right] \\ &= \frac{\omega_i}{T} \frac{(1-(1-\omega_i)^k)T}{\omega_i} E[\lambda_i] \\ &= (1-(1-\omega_i)^k)E[\lambda_i] \end{aligned} \quad (15)$$

Then,  $E[r_i^k]$  is equal to  $E[\lambda_i^k]$  if the scaling factor is one. Otherwise, the average allocated bandwidth converges exponentially to the average input rate.  $\square$

Although Theorem 2 states the property of ASA in random traffic, the case of constant traffic has the same property. Because  $E[q_i^k] = q_i^{ss} = \lambda_i^{ss} / \omega_i$ , we get  $r_i^k \approx \lambda_i^{ss}$  in the case of constant traffic

### 3.3 Controllability of ASA Queuing Delay

Next, the latency of a packet waiting in a queue is also consider. In the case of constant traffic, when the queue reaches the steady state, from Corollary 1, there are  $T\lambda_i^{ss} / \omega_i$  packet bytes waiting to be served in front of each arrived packet. At that time, the scheduler serves  $Q_i$  with a bandwidth  $r_i^{ss} \approx \lambda_i^{ss}$  according to Eq. (15). As a result, the latency of a packet in the queue is  $T / \omega_i$ . This property is stated formally in the following corollary.

**Corollary 3** In ASA, the queuing delay in a queue is approximately  $T / \omega_i$

**Proof:** By the Little's formula [17] in statistics, the average delay in queue, input rate and queue occupancy of a FIFO queue are related by

$$E[q_i] = E[d_i]E[\lambda_i] \quad (16)$$

Substituting Eq. (12) into Eq. (16) gives

$$E[d_i]E[\lambda_i] = E(\lambda_i)T \frac{1 - (1 - \omega_i)^k}{\omega_i} \quad (17)$$

and so,

$$E[d_i] = \frac{(1 - (1 - \omega_i)^k)T}{\omega_i} = \begin{cases} T & \text{if } \omega_i = 1 \\ \frac{(1 - (1 - \omega_i)^k)T}{\omega_i} & \text{otherwise} \end{cases} \quad (18)$$

Clearly, the queuing delay caused by queuing is related to the service round time and the scaling factor. In other words, ASA has the ability to guarantee the latency of each incoming packet.

From the above discussion, it is clear that ASA has the ability to control queuing delay if each GS queue can be served with its required rate  $r_i^k$ . In order to achieve the required rate of each queue, there should be a sufficient number of time slots in each round for all mobile hosts. However, the variation in channel capacity makes slot allocation complicated and delay performance unpredictable. ASA will allocate time slots to queues based on the principle of maximizing the satisfaction of all mobile hosts when the total required slots are more than the available slots  $N$  in a service round. In this case, some queues may be temporarily out of control. But, ASA is robust enough to quickly come back to a well-controlled state, after the condition of time slot insufficiency is relieved. The robustness of ASA will be verified in the next section. The condition of delay controllability of ASA is given in the following corollary.

**Corollary 4** In ASA, the condition of delay controllability is given by

$$\sum_{i=1}^m \frac{r_i^k}{C_i^k} \leq 1 \quad (19)$$

In a TDMA wireless network, slot allocation is determined by the required data rate and channel capacity as shown in Eqs. (1) and (2). Moreover, the total required slots should be no more than the slots available in a service round, which is the sufficient condition for ASA to control delay, that is,

$$\sum_i n_i^k \leq N \quad (20)$$

From Eq. (2), the total number of time slots is given by

$$\sum_i n_i^k = \sum_i \frac{r_i^k T}{C_i^k t} \quad (21)$$

Because  $T/t$  is equal to  $N$  and the sufficient condition is true when the required number of slots is not more than the total number of slots in a service round, we have

$$\sum_{i=1}^m \frac{r_i^k}{C_i^k} \leq 1. \quad \square$$

There are two cases leading to slot insufficiency, the traffic burst and the temporal degradation of the radio channel. Both are counter to the condition of delay controllability of Eq. (19) and lead to degradation of delivery quality. Since the degradation is unavoidable, ASA takes a strategy of maximizing the satisfaction of minimum slot requirement of all queues to guarantee both a minimum slot allocation and fairness for GS queues.

### 3.4 Packet Loss Analysis of ASA

Packet loss is one of the typical QoS design issues. In a wireless network, there are two conditions that cause packet loss; one is the temporary failure of a wireless link and the other is queue overflow when the wireless link is too poor to deliver the incoming packets. The former degrades the throughput of the wireless link and the later is the major issue which we discuss in this subsection.

Consider a queue  $Q_i$  with a buffer size,  $B_i$ , the input traffic  $\lambda_i$  and the serving data rate  $r_i$ , a packet loss occurs when the queue state  $q_i$  exceeds  $B_i$ . In this overflow condition, the resulting packet loss rate  $R_{loss}$ , defined by the lost byte count per unit time, is given by the difference between the input traffic rate and the serving data rate. We will investigate the critical condition of packet loss for both constant traffic and random traffic. The relationship between packet loss and the other parameters, including input traffic rate, scaling factor and service round time, is also derived clearly in the following corollaries.

**Corollary 5** In ASA, if the input traffic of  $Q_i$  is constant, then the condition of packet loss and the packet loss rate are given by

$$\lambda_i > \frac{\omega_i B_i}{T} \quad (22)$$

and

$$Loss_i = \lambda_i - \frac{\omega_i B_i}{T} \quad (23)$$

respectively.

**Proof:** When the input traffic is at a constant rate, denoted by  $\lambda_{const}$ , we have shown in Corollary 1 that ASA drives the queue state,  $q_i$ , toward a steady state given by Eq. (9). If the buffer size,  $B_i$ , is insufficient for the steady state, packet loss is certainly unavoidable. Thus the condition of packet loss for a constant input traffic is given by

$$q_i = \frac{\lambda_i T}{\omega_i} > B_i \quad (24)$$

Equivalently, we have

$$\lambda_i > \frac{\omega_i B_i}{T} \quad (25)$$

The allocated data rate of  $q_i$ , according to Eq. (1), resulting from a saturated queue is given by

$$r_i = \frac{\omega_i B_i}{T} \quad (26)$$

and the packet loss rate of a saturated queue is equal to the input traffic rate  $\lambda_i$  minus the serving data rate,  $r_i$ . That is,

$$\begin{aligned} Loss_i &= \lambda_i - r_i \\ &= \lambda_i - \frac{\omega_i B_i}{T} \end{aligned} \quad \square$$

For random traffic, we characterize the input traffic by a random variable with a certain distribution and assume the statistics of this random variable are known. The packet loss probability is investigated in Corollary 5.

**Corollary 6** In ASA, if the input traffic rate  $\lambda_i$  is a random variable with known characteristics, the packet loss probability is bounded and the stochastic bound is given by

$$P(q_i \geq B_i) \leq \frac{E(\lambda_i)T}{B_i \omega_i} \quad (27)$$

**Proof:** The probability of packet loss is defined as the probability of a packet received when a target queue is saturated. Considering Eq. (4), in ASA, the queue dynamic is coupled with the input traffic and operational parameters. Because the operational parameters are constant during operation, the queue dynamic is also regarded as a random variable. When the target queue arrives at its steady state, we have the packet loss probability

$$\begin{aligned} P(q_i \geq B_i) &= P\left(\frac{T\lambda_i}{\omega_i} \geq B_i\right) \\ &= P\left(\lambda_i \geq \frac{B_i \omega_i}{T}\right) \end{aligned} \quad (28)$$

To derive the loss probability, consider the mean rate of input traffic given by

$$\begin{aligned}
 E(\lambda_i) &= \int_0^{\infty} \lambda_i p(\lambda_i) d\lambda \\
 &= \int_0^{\lambda^0} \lambda_i p(\lambda_i) d\lambda + \int_{\lambda^0}^{\infty} \lambda_i p(\lambda_i) d\lambda \\
 &\geq \int_{\lambda^0}^{\infty} \lambda_i p(\lambda_i) d\lambda \geq \int_{\lambda^0}^{\infty} \lambda^0 p(\lambda_i) d\lambda \\
 &= \lambda^0 \int_{\lambda^0}^{\infty} p(\lambda_i) d\lambda = \lambda^0 P(\lambda_i \geq \lambda^0)
 \end{aligned} \tag{29}$$

where  $p(\lambda_i)$  is the probability density function of  $\lambda_i$  and  $\lambda^0$  is an arbitrary non-negative value. From Eq. (29), we have the relationship,

$$P(\lambda_i \geq \lambda^0) = \frac{E(\lambda_i)}{\lambda^0} \tag{30}$$

According to Eq. (30), and letting  $\lambda^0 = (B_i \omega_i) / T$ ,

$$P(q_i \geq B_i) = P(\lambda_i \geq \frac{B_i \omega_i}{T}) \leq \frac{E(\lambda_i) T}{B_i \omega_i}$$

Finally, the packet loss probability for the condition of random traffic is obtained.  $\square$

Consequently, in both constant and random traffic, a short service round time, a large buffer and scaling factor can effectively reduce the packet loss rate and probability. The quantitative relationships shown in Eqs. (22), (23) and (27) can help the design of buffer management and QoS. For the conditions of constrained buffer and differentiated QoS requirement, ASA demonstrates a flexible scheme for satisfying the various quality needs.

## 4. PERFORMANCE EVALUATION AND CHANNEL MODELING

### 4.1 Modeling of Wireless Channel

The channel model used to evaluate the ASA performance is depicted in Fig. 2. The  $J$  states indicate the variation of the target wireless channel. The transition of the channel throughput is between states. The states denoted by the vector  $\{S_i, \text{ for } i = 0, \dots, J - 1\}$  indicate the various channel throughputs, in which  $S_j$  has a higher throughput than  $S_{j-1}$ . I.e.  $S_0$  is the failure state of a wireless link and  $S_{j-1}$  is the state with the highest channel throughput. There are three kinds of state transition models. When the wireless channel remains steady, the state transition is regarded as a self-transition. When the channel fails, the state will change to  $S_0$ . The third kind of transition occurring between  $S_j$  and  $S_{j-1}$  indicates the transient characteristics of an active wireless channel. For example, in a wireless IEEE 802.11 LAN, the throughput of an access point may be 1 Mbps, 2 Mbps, 5

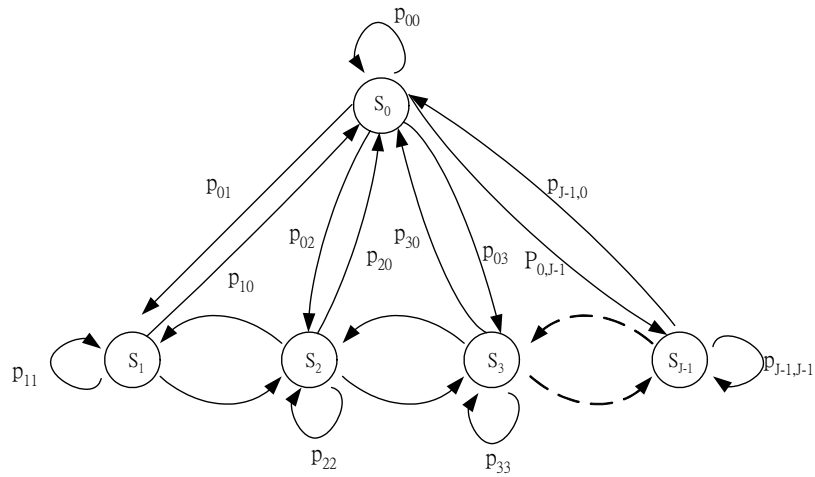


Fig. 2.  $J$ -state model of wireless channel.

Mbps or 11 Mbps. Therefore, we may use the state set  $\{S_0, S_1, S_2, S_3, S_4\}$  to model the channel characteristics. In the proposed ASA, the state probability  $S_i$  and the transition probability between  $S_i$  and  $S_j$  are defined as  $\phi_i$  and  $p_{ij}$ , respectively. Therefore we have the probability vector of all states at the time instance  $k$  given by

$$\Phi^{k+1} = \Phi^k P \tag{31}$$

where

$$\Phi^{k+1} = [\phi_0^{k+1} \ \phi_1^{k+1} \ \dots \ \phi_{J-1}^{k+1}] \tag{32}$$

$$P = \begin{bmatrix} p_{00} & p_{01} & 0 & 0 & 0 & 0 \\ p_{10} & p_{11} & p_{12} & 0 & 0 & 0 \\ 0 & p_{21} & p_{22} & 0 & \vdots & 0 \\ 0 & 0 & p_{32} & \ddots & p_{J-3,J-2} & \vdots \\ \vdots & \vdots & & & p_{J-2,J-2} & p_{J-2,J-1} \\ 0 & 0 & 0 & 0 & p_{J-1,J-2} & p_{J-1,J-1} \end{bmatrix} \tag{33}$$

and

$$\sum_{j=0}^{J-1} p_{ij} = 1 \text{ for } i = 0, \dots, J-1 \tag{34}$$

Then, the probability of the channel state at round  $k$  is given by

$$\Phi^{k+1} = \Phi^0 P^k \tag{35}$$

The transition probability depends on the channel characteristics caused by signal degradation, interference and distortion. For the case of rapid inter-state transitions, the self-transition probability is relatively small so that the channel condition changes and the throughput often varies. In contrast, if  $p_{ii}$  is large enough, even approaching 1, then the channel condition will remain constant. The adopted model is flexible so as to apply to various wireless channels. In [18], Ji *et al.* used a simplified two-state model to characterize the GSM wireless channel, where the state holding time is taken from mixtures of several geometric distributions. Hereafter, the  $J$ -state model is applied to evaluate the performance of ASA.

#### 4.2 Performance Evaluation

In this subsection, we evaluate the performance of the proposed ASA by simulation. The simulated network structure is depicted in Fig. 1. The simulated model is composed of one base station and two mobile hosts in the wireless domain. The service round time and slot time are 1 ms and 1  $\mu$ s, respectively. Thus, there are 1000 slots available in a service round. The test traffic traveling from a wired network is forwarded to the mobile hosts by the base station. To make the simulation closer to a real network scenario, we measure the real traffic caused by a video service connection from an ISP across the Internet as the test data. The simulated scenarios, including varying traffic and varying channel capacity, are used to verify the performance of ASA. Certainly, the queue states during the operation of ASA are of concern. The evaluation of queue dynamics demonstrates the feasibility of simple buffer management to prevent queues from overflowing. Each mobile host is assumed to have two kinds of received traffic, GS service and BE service. In the following simulation, GS service is evaluated.

The first scenario of simulation assumes that ASA has sufficient time slots to satisfy all mobile hosts and the channel capacity is a constant. We use this simulation to demonstrate the effect of scaling factors. This scenario conforms to the condition of Eq. (19), in which case, each GS queue is allocated with the required rate  $r_i$  such that Eq. (4) holds. Then, the states of GS queues are independent. The effects of scaling factor, including the allocated rate, queue state and queueing delay of a GS queue, are illustrated in Fig. 3. Fig. 3 (a) shows the tested input traffic used in simulation, which is a random traffic with a fixed mean. Four scaling factors,  $\omega = 0.01, 0.1, 0.5$  and 1, are applied to verify the performance of ASA. The result of rate adaptation of ASA is demonstrated in Fig. 3 (b). The allocated outgoing data rate exponentially tracks the input traffic, where the converging speed gradually becomes faster as the corresponding scaling factor increases from 0.01 to 1. Moreover, the convergence tendency is basically independent of input traffic variation. After the allocated traffic reaches the steady state, the allocated rate with higher scaling factor even has the ability of tracking a slight variation of the input traffic. Therefore, when the allocated time slots sufficiently support the needs of all GS queues, the rates of GS queues are dynamically and effectively adapted to their corresponding input traffic. Queue dynamics governed by various scaling factors are also depicted in Fig. 3 (c). Because of the tracking property of the ASA, the queue state stays nearly constant even when the input traffic is randomly varying around a mean. Also, the steady state of each GS queue is directly related to its associated scaling factor as given in Eq. (9). Also, the queueing delay is well-controlled by the scaling factor, as shown in Fig. 3 (d).

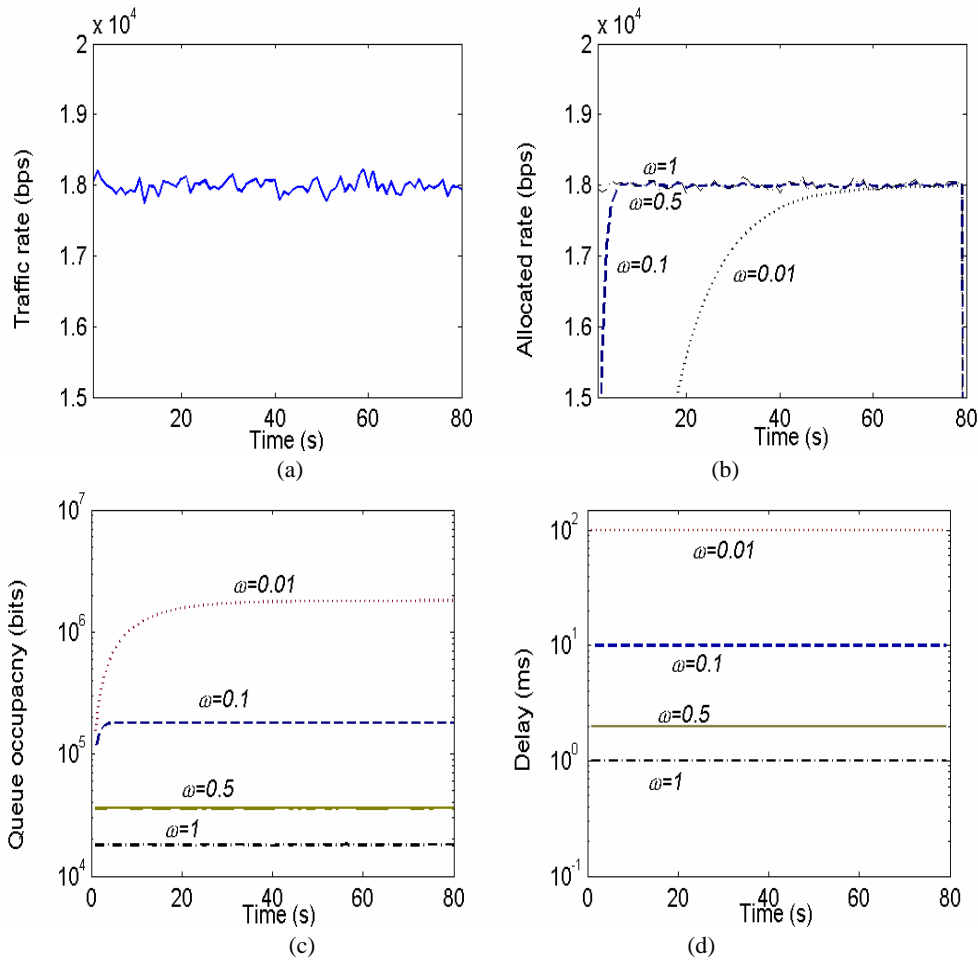


Fig. 3. Performance comparison of ASA in various scaling factor values.

Thus, if the mean and variance of traffic are available in the SLA before the connection is established, ASA is useful to simplify queue management and reduce the possibility of buffer overflow because of the resulting steady queue dynamics. By this simulation, we can confirm that the scaling factor is the key parameter for deciding the convergence of the tracking rate as shown in Eq. (15).

The performance of ASA in varying channel capacity is further evaluated in this section, and the result is shown in Figs. 4 and 5. In this simulation, we use an ON-OFF model to feature the dynamic channel quality, i.e.,  $J = 2$ .  $S_0$  and  $S_1$  indicate ON and OFF states, respectively. The state transition probability is given by

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} \tag{36}$$

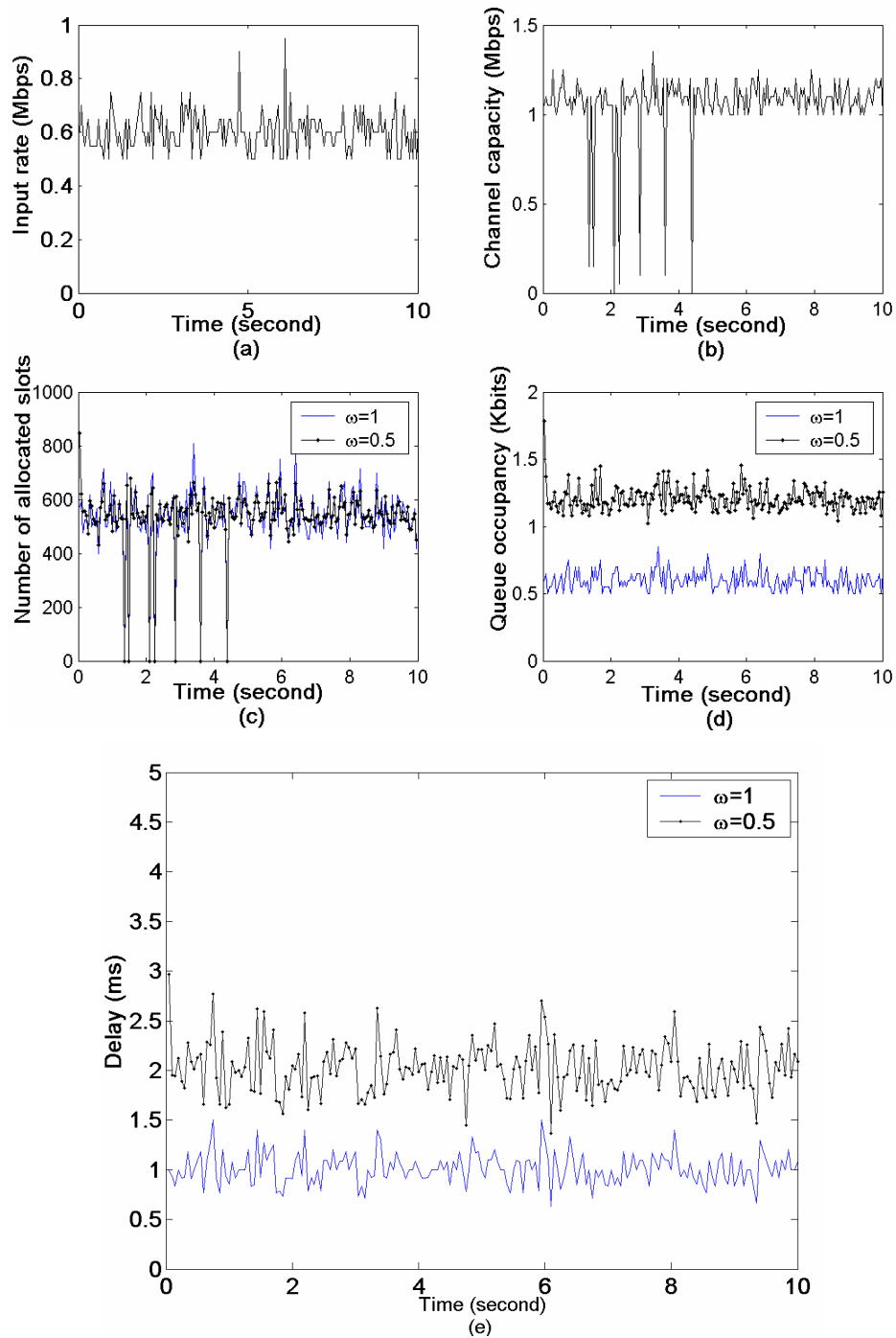


Fig. 4. Performance comparison of two mobile hosts with  $\omega = 1$  (solid line) and 0.5 (dotted line) when the time slots of BS are sufficient, and  $\{p_{10}, p_{01}\} = \{0.1, 0.9\}$ .

subject to

$$p_{00} + p_{01} = 1 \quad \text{and} \quad p_{10} + p_{11} = 1$$

We assume that each transition occurs at the beginning of a time slot and no transition is allowed during the slot time. The holding times of the ON state and OFF state are two independent random variables in one unit of time slot. They are characterized by two independent geometric distributions. From Eq. (36), the average holding times of the ON and OFF states are

$$T_{ON} = \frac{P_{11}}{P_{10}} \quad \text{and} \quad T_{OFF} = \frac{P_{01}}{P_{00}} \quad (37)$$

Then, the effective capacity of this wireless channel is as

$$C_{eff} = \frac{T_{ON} C_{ON}}{T_{ON} + T_{OFF}}$$

where  $C_{ON}$  is the throughput when the channel is in  $S_1$  and  $C_{eff}$  is the effective channel capacity. First, a simulation assuming that the effective channel capacity is sufficient to deliver the corresponding input traffic was performed.

By using the above model of a wireless channel, we obtain the channel dynamics shown in Fig. 4 (b) and use it for the simulation with the input traffic illustrated in Fig. 4 (a). Two scaling factors,  $\omega = 1$  and  $0.5$ , are used to make comparisons of the resulting slot allocation, queue dynamics and queuing delay of packets. As shown in Fig. 4 (c), ASA obviously adapts the slot allocation when the channel degrades. Figs. 4 (d) and 4 (e) show that the consequent queue dynamics and packet delay are kept steady, even with the channel variation. It reveals the phenomenon of queue dynamics implied by Eq. (13). Fig. 4 (e) also confirms the delay controllability denoted by Eq. (19) in this simulation.

The last simulation scenario is to evaluate the real performance of ASA for the condition of slot insufficiency. Either the traffic burst or the temporal degradation of the radio channel causes this condition. Both make the total time slots of a service round insufficient to meet the net requirement of all hosts. When a slot insufficiency occurs, it is not possible to support all GS queues by degrading service. This situation leads to the issue of fair allocation of time slots. As stated in Phase III of the procedure, ASA will progressively fill the slot requirements of all GS queues. As a result, each non-empty GS queue is guaranteed to have a minimum slot allocation.

The simulation results of ASA working with slot insufficiency are shown in Fig. 5. The input traffic rates of the simulated hosts have means of 600 Kbps and 300 Kbps as depicted in Fig. 5 (a). The scaling factors are assigned according to their QoS requirement, with  $\omega = 1$  and  $0.1$ , respectively. The channel characteristics of the two mobile hosts are the same as the previous case as depicted in Fig. 5 (b). The temporarily degrading channel triggers the insufficient condition in which ASA allocates slots to ensure max-min fairness. Fig. 5 (d) depicts the effect of ASA controlling the queue dynamics according to the assigned scaling factors. Even in the case of insufficient slots, ASA

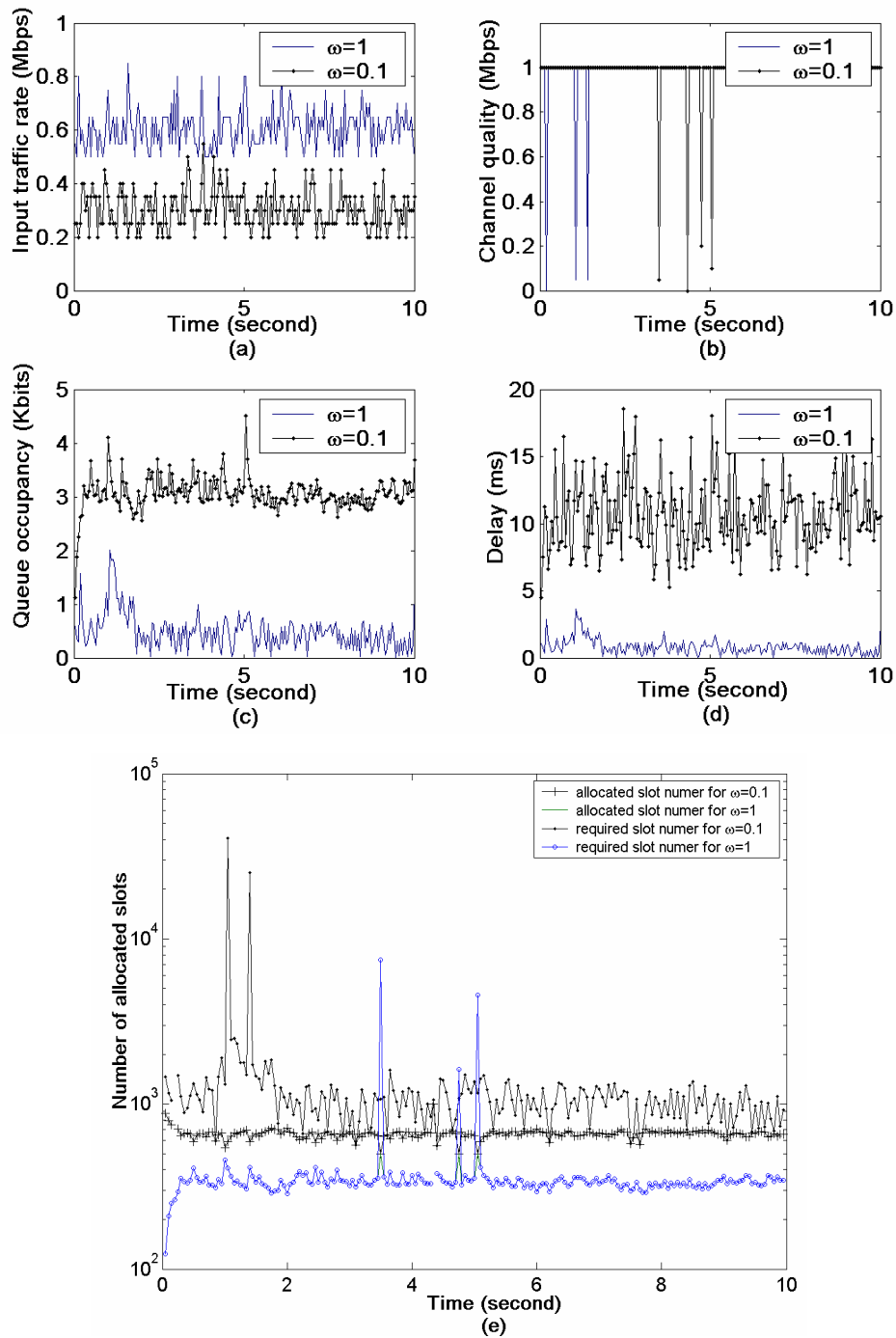


Fig. 5. Performance comparison of two mobile hosts with  $\omega = 1$  (solid line) and 0.1 (dotted line) when the time slots of BS are insufficient and  $\{p_{10}, p_{01}\} = \{0.1, 0.9\}$ .

robustly adapts the slot allocation to keep the average queue occupancy steady as stated in Corollary 2. The delay performance of the two hosts is shown in Fig. 5 (d). ASA has the ability to adapt slot allocation so that the delay is controlled as quickly as possible. Although, the channel and input traffic are varying, the average delay of two GS queues is still controlled. In Fig. 5 (e), the required slot numbers and the allocated slot numbers of two GS queues are shown to demonstrate the fair slot allocation.

The impact of radio links on the ASA performance is also of concern. We use the input traffic shown in Fig. 4 (a) to simulate the ASA performance in varying channel capacity. The variation of channel capacity is modeled as the transition probability set  $\{p_{01}, p_{10}\}$ . When the probability,  $p_{10}$ , is high, the capacity is more subject to change from the ON state to the OFF state, implying that  $p_{11}$  is too small to make the channel capacity stay ON. The same property is also derivable for  $p_{01}$ . The simulation results are given in Table 2. The various transition probabilities cause changes in the channel capacity. When the effective capacity is less than the input traffic rate, obviously the queue occupancy is out of control and packet delay increases, regardless of scaling factors. But, if the effective capacity is enough to transport the input traffic, ASA demonstrates the ability to control the queue occupancy and the delay of packets, even if the channel capacity is temporarily insufficient. In other words, ASA is robust in bringing the over-accumulated queue back to its steady state when the channel capacity scarcity is alleviated.

**Table 2. Average queue occupancy and delay performance in various radio channel conditions.**

$P_{10}$	$P_{01}$	Effective Capacity $EffC_{eff}$ (Mbps)	Input Traffic Rate $\lambda$ (Mbps)	Allocated Data Rate $r$ (Mbps)		Queue Occupancy $q$ (bits)		Delay $d$ ( $\mu$ s)	
				$\omega = 0.9$	$\omega = 0.1$	$\omega = 0.9$	$\omega = 0.1$	$\omega = 0.9$	$\omega = 0.1$
0.1	0.1	0.42	0.6	$\omega = 0.9$	0.42	$\omega = 0.9$	98	$\omega = 0.9$	167
				$\omega = 0.1$	0.42	$\omega = 0.1$	98	$\omega = 0.1$	167
0.1	0.5	0.95	0.6	$\omega = 0.9$	0.61	$\omega = 0.9$	6.9	$\omega = 0.9$	1.5
				$\omega = 0.1$	0.6	$\omega = 0.1$	6.4	$\omega = 0.1$	10.8
0.1	0.9	0.99	0.6	$\omega = 0.9$	0.61	$\omega = 0.9$	0.8	$\omega = 0.9$	1.3
				$\omega = 0.1$	0.61	$\omega = 0.1$	6.1	$\omega = 0.1$	10.3
0.5	0.1	0.44	0.6	$\omega = 0.9$	0.44	$\omega = 0.9$	198	$\omega = 0.9$	334
				$\omega = 0.1$	0.44	$\omega = 0.1$	198	$\omega = 0.1$	334
0.5	0.5	0.52	0.6	$\omega = 0.9$	0.52	$\omega = 0.9$	11	$\omega = 0.9$	18.7
				$\omega = 0.1$	0.52	$\omega = 0.1$	11	$\omega = 0.1$	18.8
0.5	0.9	1	0.6	$\omega = 0.9$	0.59	$\omega = 0.9$	0.81	$\omega = 0.9$	1.4
				$\omega = 0.1$	0.59	$\omega = 0.1$	0.5	$\omega = 0.1$	10.2
0.9	0.1	0.58	0.6	$\omega = 0.9$	0.58	$\omega = 0.9$	140	$\omega = 0.9$	237
				$\omega = 0.1$	0.58	$\omega = 0.1$	140	$\omega = 0.1$	237
0.9	0.5	0.96	0.6	$\omega = 0.9$	0.6	$\omega = 0.9$	3	$\omega = 0.9$	5.1
				$\omega = 0.1$	0.6	$\omega = 0.1$	7.7	$\omega = 0.1$	12.9
0.9	0.9	0.99	0.6	$\omega = 0.9$	0.6	$\omega = 0.9$	0.9	$\omega = 0.9$	1.4
				$\omega = 0.1$	0.6	$\omega = 0.1$	6.1	$\omega = 0.1$	10.2

## 5. CONCLUSION AND DISCUSSION

In this paper, we propose a new scheme of adaptive slot allocation, the ASA scheme, to offer QoS-based differentiated services on datagram networks. According to the operational model presented in this paper, ASA utilizes an adaptive time slot allocation mechanism. The adaptation is realized by using the queue state as feedback to track the variation of input traffic. A scaling factor is adopted in each GS queue to reflect the service level agreement that the mobile host has. Therefore, ASA has the functions of input traffic tracking and differentiated services. When the base station has sufficient slots to meet all requirements of GS queues, the queueing delay and the queue occupancy are well controlled by ASA. If time slots are insufficient for all GS queues, ASA serves the queues on the principle of maximizing the satisfaction of the minimum slot requirement among all mobile hosts. In this condition, ASA also exhibits good robustness. In summary, ASA features the properties of delay/queue controllability, robustness and differentiation.

The above-mentioned properties have been analyzed theoretically in this paper. The relationships between the operational parameters and the QoS indices are derived. The analyses show that the queue occupancy and queueing delay are effectively controlled by the scaling factor and service round time.

Finally, we perform real case simulations with real video streaming traffic caused by VoD service provided by an ISP to demonstrate the real performance of ASA in the condition of highly varying traffic. The simulation shows that ASA rapidly and precisely tracks the variation of the input traffic so that the results are well-controlled queue dynamic and strongly guaranteed packet delay. The well-controlled queue dynamics further eases the buffer management and reduces the possibility of packet loss.

## REFERENCES

1. B. A. Forouzan, *Data Communications and Networking*, McGraw-Hill, Singapore, 2001.
2. A. S. Tanenbaum, *Computer Networks*, Prentice Hall, New Jersey, 1996.
3. W. Turin, R. Jana, C. Martin, and J. Winters, "Modeling wireless channel fading," in *Proceedings of International Conference on Vehicular Technology*, 2001, pp. 1740-1744.
4. R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource reservation protocol (RSVP) version 1 functional specification," RFC 2205, 1997.
5. S. Blake *et al.*, "An architecture for differentiated services," RFC 2475, 1998.
6. M. Andrews *et al.*, "Providing quality of service over a shared wireless link," *IEEE Communication Magazine*, Vol. 39, 2001, pp. 150-154.
7. S. Kapp, "802.11: leaving the wire behind," *IEEE Internet Computing*, Vol. 6, 2002, pp. 82-85.
8. IEEE 802.11 WG, *Draft Supplement to STANDARD FOR Telecommunications and Information Exchange Between Systems -LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)*, IEEE

- 802.11e/D2.0, 2001.
9. P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communication Magazine*, Vol. 39, 2001, pp. 70-77.
  10. S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz, and L. Stibor, "IEEE 802.11e wireless LAN for quality of service," in *Proceedings of International Conference on European Wireless (EW'2002)*, 2002, pp. 32-39.
  11. M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel condition," Bell Labs Technical Memorandum, 2000.
  12. M. Mirhakkak, N. Schult, and D. Thomson, "Dynamic bandwidth management and adaptive application for a variable bandwidth wireless environment," *IEEE Journal on Selected Areas Communications*, Vol. 19, 2001, pp. 1984-1997.
  13. E. L. Hahne and A. K. Choudhury, "Dynamic queue length thresholds for multiple loss priorities," *IEEE/ACM Transactions on Networking*, Vol. 10, 2002, pp. 368-380.
  14. C. G. Park, D. H. Han, and Y. Lee, "Performance analysis of threshold based bandwidth allocation scheme for IP traffic on ATM networks," *IEE Proceeding Communication*, Vol. 149, 2002, pp. 29-33.
  15. H. H. Yoon, H. Kim, C. Oh, and K. Kim, "A queue length-based scheduling scheme in ATM networks," in *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, Vol. 1, 1999, pp. 234-237.
  16. J. Filipiak, *Modeling and Control of Dynamic Flows in Communication Networks*, Springer-Verlag, Heidelberg, 1988.
  17. L. Kleinrock, *Queueing System*, Wiley, New York, 1975.
  18. P. Ji, B. Liu, D. Towsley, and J. Kurose, "Modeling frame-level errors in GSM channels," in *Proceedings of IEEE Global Telecommunications Conference*, Vol. 3, 2002, pp. 2483-2487.



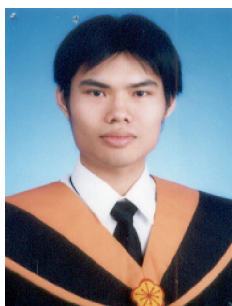
**Mong-Fong Horng (洪盟峰)** received the B.S and M.S. degrees in Control Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1989 and 1991, respectively, and the Ph.D. degree in Computer Science and Information Engineering from National Cheng Kung University in 2003. From 1993 to 2003, he was with the Department of Electrical Engineering, Kao Yuan Institute of Technology, Kaohsiung, Taiwan. From 1999 to 2001, he was the vice chairman of the Computer Center at Kao Yuan Institute of Technology. Since 2003, he has been an Associate Professor with the Department of Computer Science and Information Engineering, Shu-Te University, Kaohsiung, Taiwan. His research interests include network QoS, broadband communication, and Artificial Intelligence.



**Yau-Hwang Kuo (郭耀煌)** was born in Tainan, Taiwan in 1959. He received M.S. and Ph.D. degrees in Computer Engineering from National Cheng Kung University in 1984 and 1988. He was the President of Taiwanese AI Association from 1999 to 2000, the Director of Research Center for Computer System Technology from 1997 to 2000, and the Managing Director of Chinese Fuzzy System Association from 1996-2000. He is currently Professor with the Department of Computer Science and Information Engineering, National Cheng Kung University. He is also the Director of Center for Research of E-life Digital Technology, and the coordinator of Computer Science & Information Engineering Program of National Science Council, R.O.C. His research interests include intelligent computing, knowledge management, broadband communication, information retrieval, pattern recognition and VLSI design.



**Jang-Pong Hsu (許振鵬)** received the M.S. and Ph.D. degrees in Electrical Engineering and Information Engineering from National Cheng Kung University in 1987 and 1998, respectively. Since 2002, he has been the director of RD department at Advance Multimedia Internet Technology Inc. in Tainan, Taiwan, R.O.C. His research interests include pattern recognition, fuzzy neural network systems, intrusion detection systems, and virtual private network.



**Ren-Hao Cheng (鄭人豪)** received the M.S. degree in Computer Science and Information Engineering from National Cheng Kung University in 2003. Since 2003, he has been a RD engineer at Advance Multimedia Internet Technology Inc. in Tainan, Taiwan, R.O.C. His research interests include wireless networks, packet scheduling algorithms, Quality of Service and VoIP systems.