Cross-Layer Design for Dynamic Resource Allocation in Wireless Networks

John Y. Kim, Ali Saidi and Randall J. Landry

The MITRE Corporation
202 Burlington Road, Bedford, MA 01730-1420
{johnkim,asaidi,rlandry}@mitre.org

Abstract— In this paper, a novel analytical cross-layer design framework for dynamic resource management of wireless networks is proposed. First, dynamic bandwidth and time resource allocation policies for a single-user under fading channels that maximize capacity are derived. The analysis is then extended to multi-user environments, where the resource allocation is jointly optimized across both physical and data link layers. A closed-form expression of a QoS measure, mean delay in this case, is derived as a function of layer 2 traffic, multiple access contention from other users, and allocated data-rates at the physical layer. This mean delay expression is then used to efficiently allocate physical layer resources. We also study the effects of various contention mitigation policies on network capacity and average latency under optimum resource allocation strategies.

Index Terms—cross-layer design (CLD), dynamic bandwidth allocation (DBA), dynamic time allocation (DTA), information theory, quality of service (QoS), queing theory, resource management

A. INTRODUCTION

Future communication systems will be characterized by high data-rates, diversity in the type of transmitted data and a rapid evolution towards increasingly network-centric architectures. This places a growing demand on the efficient utilization of system resources including transmitter power, time and assigned channel bandwidth. Conventional fixed mode communication systems are typically designed for worst-case channel conditions to combat the effects of deep fading. Adaptive communication systems, on the other hand, provide an important alternative that can achieve considerably higher efficiency compared to non-adaptive systems. The adaptation can be performed by selective variation of the parameters in the physical layer as well as in the higher layers of the protocol stack.

The advantages of adaptive communication systems for operation in widely varying wireless channel conditions are substantial. However, finding an optimum adaptation strategy for a given set of constraints on network resources is not a trivial problem. Maximum efficiency can only be achieved when protocol layers collaboratively respond to changing network state by dynamically allocating resources subject to the network constraints [1]. The complexity of this joint optimization problem through cross-layer design (CLD) [2] grows very fast as the number of layers and the parameters in each layer considered for optimization increases. Therefore, it is critically important to consider the feasibility of solving the cross-layer joint optimization problem when selecting variables for adaptation [2,3,4]. Recent work in the area of CLD for wireless networks has focused on rather "ad-hoc" approaches in which state information at one layer is used by higher layer protocols to improve network performance relative to a strictly layered methodology [5]. The majority of these approaches do not promote a fundamental analytical treatment of the CLD problem.

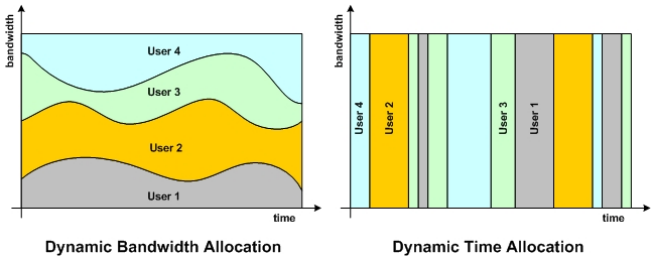

Figure 1 Proposed dynamic resource allocation schemes

In this paper we consider a CLD for the allocation of physical layer resources which jointly optimizes across the physical and data-link layers. The proposed analytical approach focuses on the allocation of bandwidth and time, the two primary shared resources in multiple access systems. The resulting Dynamic Bandwidth Allocation (DBA) and Dynamic Time Allocation (DTA) schemes for given network constraints are represented in a notional way in Figure 1. For DBA, which is based on FDMA, the amount of disjoint bandwidth resource being occupied by each user dynamically changes based on its channel condition. Different users can use the channel simultaneously, subject to the total bandwidth constraint as well as each users average bandwidth constraint. In the TDMA based DTA, time, rather than bandwidth, is dynamically allocated to each user. The transmitting user occupies the entire system bandwidth, which means only a single user can access the channel at a time subject to each user's average time constraint. One practical realization of the proposed schemes could be based on a multi-

carrier system whose bandwidth (carriers) and time resources can easily be managed [12].

Our approach to CLD combines information theoretic techniques with queuing theoretic techniques to develop an analytical framework for joint optimization of resource utilization across multiple layers. Physical layer constraints such as average bandwidth and time in our approach are assumed to be arbitrary in simultaneously deriving the optimum policies at physical and data link layers. The actual values of the average bandwidth and time allocations are then computed based on the network traffic parameters as well as multiple access contention from other users to satisfy a soft QoS measure such as average delay.

We begin by deriving the optimal bandwidth and time allocation solutions for the single user case. In [6], the authors compute the optimal power allocation policy under fading channels which maximizes the capacity for a given average power constraint. We employ a similar approach for bandwidth and time allocations assuming a fixed transmit power level, and obtain the optimal solutions for given average bandwidth and time constraints. The analysis is then extended to the multi-user case where contention and contention resolution polices are discussed. A priority queuing system is employed to model the impact of multiple access contention on traffic arrivals at a single user. An expression is derived for mean delay as a function of user traffic, multiple access contention and average data-rate allocated to the user. Delay constraints are then used to compute the optimal allocation of data-rate, and hence time or bandwidth resources.

The remainder of the paper is organized as follows. In Section B, our system model is described and the optimal DBA and DTA solutions are derived. Various contention policies and corresponding capacities are discussed in Section C. Our CLD approach and delay analysis are presented in Section D. Finally, the paper is concluded with relevant results and some final remarks in Sections E and F, respectively.

B. Optimal Capacity Analysis

In this section, the optimal capacity analysis for DBA and DTA is presented. It is assumed that the transmit power level is fixed and the channel encounters flat fading with unit average gain. It is also assumed that the channel information is known to both transmitter and receiver.

Dynamic Bandwidth Allocation (DBA)

Let $B(\gamma)$ be the instantaneous signal bandwidth which is a function of signal-to-noise ratio (SNR) given by γ. Since the transmit power level is assumed be fixed, the SNR distribution is dictated by the channel fading distribution. The optimal solution for $B(\gamma)$ can then be found by maximizing the average capacity, C_{DBA}, under DBA as follows:

$$maximize\ C_{\text{DBA}} = \int_0^\infty B(\gamma)\log_2\left[1+\gamma\frac{\bar{B}}{B(\gamma)}\right]P(\gamma)\,d\gamma, \quad (1)$$

subject to the average bandwidth constraint

$$\int_0^\infty B(\gamma)P(\gamma)d\gamma = \bar{B}, \quad (2)$$

where $P(\gamma)$ is the pdf of SNR. This optimization leads to the following Euler-Lagrange equation:

$$\frac{\partial}{\partial B(\gamma)}\left(B(\gamma)\log_2\left[1+\gamma\frac{\bar{B}}{B(\gamma)}\right]-\lambda B(\gamma)\right)=0, \quad (3)$$

where λ is the Lagrange multiplier. Expanding and simplifying the above equation leads to

$$\log_2\left[1+\gamma\frac{\bar{B}}{B(\gamma)}\right]-\frac{\gamma\bar{B}}{B(\gamma)+\gamma\bar{B}}-\lambda=0. \quad (4)$$

Letting $X = B(\gamma)/\bar{B}$, the above equation simplifies to

$$\log_2\left[1+\frac{\gamma}{X}\right]-\frac{1}{\frac{X}{\gamma}+1}=\lambda. \quad (5)$$

It follows from this expression that since λ is a constant, the variable X/γ must also be a constant, which suggests that $B(\gamma)$ is directly proportional to γ. Suppose $B(\gamma)=\varepsilon\gamma$. Then we can solve for ε from the constraint equation (2):

$$\varepsilon = \frac{\bar{B}}{E[\gamma]} \quad (6)$$

Therefore, the optimal bandwidth allocation that maximizes the average capacity is $\bar{B}\gamma/E[\gamma]$. Substituting $B(\gamma)$ into (1) we get the maximum average capacity

$$C_{\text{DBA}} = \bar{B}\log_2\left[1+E[\gamma]\right], \quad (7)$$

which is the AWGN Shannon capacity for a given average SNR. The optimum DBA policy allocates more bandwidth when the channel fading is less severe and can be viewed as a bandwidth water-filling technique [7] in time.

Two interesting observations are noteworthy. First, the average capacity is independent of channel fading conditions when the optimal BW allocation policy is employed. Second, the spectral efficiency stays constant for a given SNR distribution implying that no adaptive modulation scheme is required for DBA.

The above derivation assumes that the SNR measurements are made with respect to the average bandwidth constraint, $\bar{B}$. In order to make relative comparisons, the derivation needs to include a reference. Let B_{ref} be the reference bandwidth from which the SNR distribution data is obtained. Then, the instantaneous SNR under instantaneous bandwidth $B(\gamma)$ is $\gamma B_{\text{ref}}/B(\gamma)$. Therefore,

$$C_{\text{DBA}} = \bar{B}\log_2\left[1+E[\gamma]\frac{B_{\text{ref}}}{\bar{B}}\right]. \quad (8)$$

The above equation tells us that, as the average bandwidth allocation increases, the capacity also increases while the spectral efficiency diminishes.

In practice, the allocated bandwidth will be upper-bounded by the total system bandwidth, $B_{\max}$. Then, the optimal instantaneous bandwidth allocation becomes

$$B(\gamma) = \begin{cases} \varepsilon\gamma, & \gamma \le \gamma_{\max} \\ \varepsilon\gamma_{\max}, & \gamma > \gamma_{\max} \end{cases}, \quad (9)$$

where $\gamma_{\max} = B_{\max}/\varepsilon$, and the average bandwidth constraint is expressed as

$$\int_0^{\gamma_{\max}} \varepsilon\gamma P(\gamma)\partial\gamma + \int_{\gamma_{\max}}^{\infty} \varepsilon\gamma_{\max} P(\gamma)\partial\gamma = \bar{B}. \quad (10)$$

The optimal DBA capacity in the presence of a bounded system bandwidth is then given by

$$\begin{aligned} C_{\text{DBA}} = &\int_0^{\gamma_{\max}} \varepsilon\gamma\log_2\left[1+\frac{B_{\text{ref}}}{\varepsilon}\right]P(\gamma)d\gamma \\ &+ \int_{\gamma_{\max}}^{\infty} B_{\max}\log_2\left[1+\frac{\gamma B_{\text{ref}}}{B_{\max}}\right]P(\gamma)d\gamma \end{aligned} \quad (11)$$

Note that the above bounded result does not deviate much from the ideal case when $\bar{B} << B_{\max}$ (as it normally would be in the practical multi-user cases).

Dynamic Time Allocation (DTA)

In DTA, the bandwidth resource is assumed to be fixed, and the only resource being managed/allocated is time. Let the time allocation policy, $T(\gamma)$, be an indicator function such that

$$T(\gamma) = \begin{cases} 1, & \gamma \in \{\text{tx region}\} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

Then, the optimal $T(\gamma)$ is obtained by solving the following constraint optimization problem:

$$\textit{maximize } C_{\text{DTA}} = \int_0^{\infty} T(\gamma)B_{\max}\log_2\left[1+\gamma\right]P(\gamma)d\gamma, \quad (13)$$

subject to the average time constraint

$$\int_0^{\infty} T(\gamma)P(\gamma)d\gamma = \bar{T}, \quad (14)$$

Since the instantaneous capacity is an increasing function of γ, the average capacity is maximized if $\gamma \in \{\text{tx region}\}$ takes on SNR values that exceed a threshold, γ_{tx}, determined by:

$$\int_{\gamma_{\text{tx}}}^{\infty} P(\gamma)d\gamma = \bar{T}. \quad (15)$$

The resulting optimal average capacity is simply

$$C_{\text{DTA}} = \int_{\gamma_{\text{tx}}}^{\infty} B_{\max}\log_2\left[1+\gamma\right]P(\gamma)d\gamma. \quad (16)$$

The average time constraint can be directly translated into the average power constraint. Therefore, the optimal DTA policy can be interpreted as a time domain variant of constant power water-filling proposed in [13],. The optimal DTA performance, unlike DBA, depends on channel fading conditions, and its spectral efficiency changes as a function of SNR. This implies that achieving maximum average capacity in the DTA scheme requires adaptive modulation. Again, it is noted that the above analysis is based on the assumption that the SNR is measured with respect to the total system bandwidth and needs to be adjusted when making relative comparisons.

C. CONTENTION AND EFFECTIVE CAPACITY

We have derived the optimal capacity solutions for DBA and DTA given average resource constraints for the single user case. In this section, we study multiple-access systems and the impact of multi- user contention on the performance compared to the optimal single user capacities. Under fixed resource allocations, there is no contention as long as the sum of individual allocations does not exceed the total system resource pool. However, when there are two or more users in the DBA/DTA based systems, competing for the same resource could lead to contention. Hence, the actual delivered capacity, or *effective capacity*, is a function of this contention probability and the employed contention-resolution policy.

DBA

Let N be the number of users in the system sharing the common resource pool with independent SNRs. In DBA, a contention arises when the sum of individual instantaneous bandwidth allocations exceeds the total system bandwidth. Let β_{DBA} be the contention probability for a DBA based multi-user system. Then,

$$\beta_{\mathrm{DBA}} = P\left[\sum_{i=1}^{N} \frac{\bar{B}_i \gamma_i}{E[\gamma_i]} > B_{\max}\right], \tag{17}$$

where γ_i and $\bar{B}_i$ denote the instantaneous SNR and average bandwidth constraint for user i, respectively,. Given that B_i has the same distribution as γ_i, the distribution of the sum of B_i over all i can be characterized from the pdf of γ_i [8][9]. For example, assuming each γ_i is Rayleigh distributed, the above contention probability can be calculated as follows [8]:

$$\beta_{\mathrm{DBA}} = \int_{B_{\max}}^{\infty} P(s)ds$$

$$P(s) = \begin{cases} \left(\frac{1}{\bar{B}_i}\right)^N \frac{s^{N-1}}{(N-1)!}\exp\left[-\frac{s}{\bar{B}_i}\right] & \text{if } \bar{B}_i = \bar{B}_j \\ \sum_{i=1}^{N} \bar{B}_i^{N-2} \exp\left[-\frac{s}{\bar{B}_i}\right] \prod_{j\neq i}^{N} \frac{1}{\bar{B}_i - \bar{B}_j} & \text{if } \bar{B}_i \neq \bar{B}_j \end{cases} \tag{18}$$

where $P(s)$ is the pdf of the sum of the N random variables B_i.

The effective capacity for User i, $\tilde{C}_i$, is a function of contention probability and the contention mitigation policy

$$\begin{aligned} \tilde{C}_i &= E\left[C_i(\gamma_i)\middle|\text{no contention}\right](1-\beta_{\mathrm{DBA}}) \\ &\quad + E\left[C_i(\gamma_i,\text{policy})\middle|\text{contention}\right]\beta_{\mathrm{DBA}} \\ &= \tilde{C}_i^1 + \tilde{C}_i^2 \end{aligned} \tag{19}$$

$C_i(\gamma_i)$ is the instantaneous capacity resulting from the optimal DBA policy specified by (6), whereas $C_i(\gamma_i,\text{policy})$ is the instantaneous capacity based on the specific contention resolution policy being employed.

The choice of the contention mitigation policy greatly impacts the system performance. Consider the following two different choices of policies and their effects on capacity:

$$C_i(\gamma_i,\text{policy 1}) = \frac{B_i(\gamma_i)}{\sum_{i=1}^{N} B_i(\gamma_i)} B_{\max}\log_2[1+\tilde{\gamma}_i] \tag{20}$$

$$C_i(\gamma_i,\text{policy 2}) = \frac{\bar{B}_i}{\sum_{i=1}^{N} \bar{B}_i} B_{\max}\log_2[1+\tilde{\gamma}_i] \tag{21}$$

Policy 1 implements a discipline where the allocation during contention is based on optimal instantaneous allocations, whereas Policy 2 divides the bandwidth based on average target bandwidth requirements, analogous to fixed allocation. $\tilde{\gamma}_i$ is the corresponding SNR resulting from altered bandwidth allocation policies. The effective capacity can be obtained by solving

$$\tilde{C}_i^1 = \int_0^{B_{\max}} P(s')ds' \int_0^{\xi_i - s'} C_i(\gamma_i) P(\gamma_i) d\gamma_i \quad (22)$$

$$\tilde{C}_i^2 = \int_0^{\infty} P(s')ds' \int_{\xi_i - s'}^{\xi_i} C_i(\gamma_i, \text{policy}) P(\gamma_i) d\gamma_i \ , \quad (23)$$

where $s' = \sum_{j \neq i}^{N} B_j(\gamma_j)$ and $\xi_i = B_{\max} E[\gamma_i] / \bar{B}_i$. The contention policy choice must depend upon specific QoS needs of individual users in the system. Our proposed resource management approach offers two additional degrees of freedom to control user QoS outcomes by assigning resource constraints and enforcing contention policies. Therefore, our approach offers a flexible means to satisfy different user QoS requirements under various channel conditions. For example, it is known that the optimal "sum of rate" capacity for FDMA is achieved when each user's bandwidth allocation is proportional to its instantaneous SNR [10]. However, the resulting policy does not always adhere to the resource constraints and favors users with "good" channels over others.

DTA

In DTA, a contention arises when more than one user is allowed to transmit at a given instance. Unlike in DBA, the users can have different contention probabilities depending on $\bar{T}$ assignments. For example, the probability of contention for User *i*, β_i, is

$$\begin{aligned} \beta_i &= 1 - P[\text{no contention}] \\ &= 1 - \prod_{j \neq i}^{N} P\left[\gamma_j < \gamma_{tx}^j\right] \\ &= 1 - \prod_{j \neq i}^{N} (1 - \bar{T}_j), \end{aligned} \quad (24)$$

where $\bar{T}_i$ is the time constraint for User *i*. Similar to DBA, the effective capacity for User *i*, is $\tilde{C}_i^1 + \tilde{C}_i^2$ defined in a similar way. For DTA, however, $\tilde{C}_i^1$ is simply

$$\tilde{C}_i^1 = [1 - \beta_i] C_{\text{DTA\_optimal}} \ . \quad (25)$$

The value of $\tilde{C}_i^2$ again depends on the chosen contention-resolution policy. Assuming the policy randomly selects a user to transmit during contention, then

$$\begin{aligned} \tilde{C}_i^2 = &\sum_{j=1}^{N-1} P[\text{contention due to } j \text{ other users}] \\ &\times \frac{C_{\text{DTA\_optimal}}^{\text{i}}}{j+1} \end{aligned} \quad (26)$$

The policy which maximizes the aggregate capacity for TDMA is to have the user with the highest SNR transmit during contention [11].

Residual Bandwidth Utilization

Up until this point we have ignored the residual bandwidth/time resources during non-contention periods:

$$\begin{aligned} &B_{\text{residual}}(\boldsymbol{\gamma}) = B_{\max} - \sum_{i=1}^{N} B_i(\gamma_i), \quad \sum_{i=1}^{N} B_i(\gamma_i) < B_{\max} \\ &T_{\text{residual}}(\boldsymbol{\gamma}) = \prod_{i=1}^{N} (1 - \bar{T}_i) \end{aligned} \quad (27)$$

These residual resources can be utilized to make up for the capacity loss due to contention and/or to service "best-effort" traffic. This means that additional residual bandwidth utilization (RBU) policies are needed for distributing the residual resources among active users. The resulting effective capacity becomes

$$\begin{aligned} \tilde{C}_i = &E\left[C_i(\gamma_i, \text{RBUpolicy}) \middle| \text{no contention}\right](1 - \beta) \\ &+ E\left[C_i(\gamma_i, \text{RBUpolicy}_{\text{c}}) \middle| \text{contention}\right]\beta \end{aligned} \quad (28)$$

The same policy can be employed for both non-contention and contention periods, in which case the effective capacity simply becomes

$$\tilde{C}_i = E\left[C_i(\gamma_i, \text{policy})\right] \ . \quad (29)$$

D. Delay Analysis of Cross-Layer Design

In this section, we present the cross-layer aspects of our work, which demands that we consider the impact of network traffic and the Quality of Service (QoS) delivered to User *i* in the presence of multiple access contention. In order to model the behavior of User *i* in the presence of multiple users, we adopt a 2-priority

queuing system with a Head-of-Line (HoL) service discipline. Let λ_i denote the mean arrival rate of the low-priority customers (packets) arriving to the queue. These arrivals represent the actual data awaiting transmission from User *i*. Let λ_C denote the arrival of *contention jobs* to the queue. These arrivals can be seen as "dummy packets" intended to model the impact on local packets of physical layer resources, either bandwidth or time, that have been allocated to other users. Since the arrival of such a job results in local traffic being delayed for some time, we choose to give these jobs HoL priority over User *i* arrivals. Let $\overline{R}_i$ denote the mean service rate of the queue. The service rate distribution is assumed to be general. Note that $\overline{R}_i$ represents the average data-rate allocated to User *i*. Using well-known results for the mean delay of such a HoL priority queue [14] and substituting for the quantities defined above, we arrive at the following equation for the mean delay of User *i* packets:

$$\overline{D}_i = \frac{\lambda_i + \lambda_C}{2E\left\{R_i^{\,2}\right\} \cdot \left[1 - \frac{\lambda_C}{\overline{R}_i}\right] \cdot \left[1 - \frac{1}{\overline{R}_i}\left(\lambda_i + \lambda_C\right)\right]} \qquad (30)$$

We now apply this mean delay analysis to our DBA/DTA multiple access scenarios. We can substitute $C_{\text{opt},i}$ for $\overline{R}_i$ since it is the average capacity (data rate) assigned to User *i* under no contention condition. In the previous section we have defined the effective capacity, which is the actual delivered capacity in the presence of multiple-access contention. Since λ_C represents the capacity loss due to contention, it is simply

$$\lambda_C = C_{\text{opt},i} - \tilde{C}_i \qquad (31)$$

Substituting for $\overline{R}_i$ and λ_C in (30) yields the following expression for the mean delay of DBA/DTA User *i* packets:

$$\overline{D}_i = \frac{\lambda_i + \beta_i' C_{\text{opt},i}}{2E\left\{C_i^{\,2}\right\} \cdot \left[1 - \beta_i'\right] \cdot \left[1 - \beta_i' - \frac{\lambda_i}{C_{\text{opt},i}}\right]}, \qquad (32)$$

where

$$\beta_i' = \frac{\lambda_C}{C_{\text{opt},i}} = 1 - \frac{\tilde{C}_i}{C_{\text{opt},i}}. \qquad (33)$$

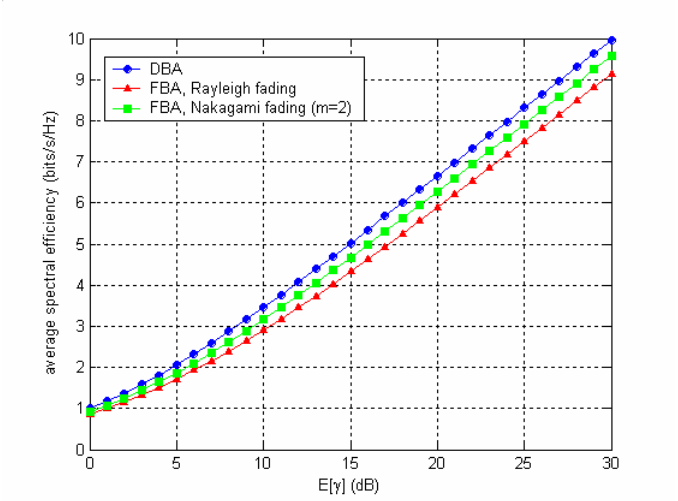

Figure 2. Average spectral efficiency comparison between dynamic and fixed bandwidth allocations

β_i' is an indicator of severity of the contention for User *i*. As β_i' approaches zero, the mean delay performance improves.

Our CLD approach is captured by (32), which describes the user mean delay as a function of variables in both physical and data link layers consisting of assigned resource constraint, contention mitigation policy, residual bandwidth utilization policy, channel fading and data arrival statistics. For given user mean delay requirements and arrival rates, one can utilize the network performance and physical resource relationship shown in (32) to assign appropriate resource constraints and implement a suitable contention policy. Not only is our CLD approach useful for resource allocation, but it can also be used to aid admission/scheduling decisions, since it can accurately predict the impact that additional users(s)/transmission(s) will have on the QoS performance of existing users.

E. RESULTS

Figures 2 and 3 illustrate the average capacity performances of the proposed dynamic bandwidth and time allocations, respectively. The figures also show the corresponding results for fixed bandwidth and time allocations. Fixed bandwidth allocation (FBA) assigns static bandwidth $\overline{B}$ while fixed time allocation (FTA) assigns the same portion of each frame, given by $\overline{T}$, to every user for transmission, both regardless of the channel condition. Therefore, the average capacities for FBA and FTA are:

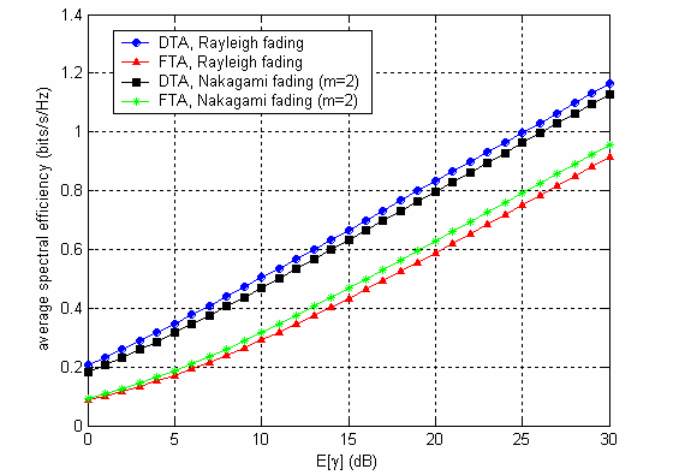

Figure 3. Average spectral efficiency comparison between dynamic and fixed time allocations; $\overline{T} = 0.1$

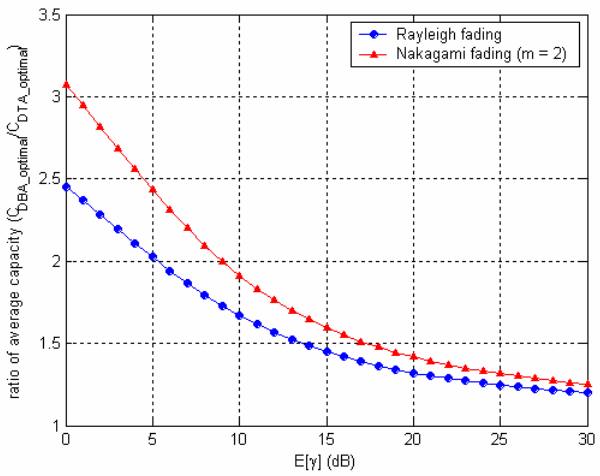

Figure 4. Ratio of dynamic bandwidth and time allocation average capacities; $\overline{B}/B_{\max} = \overline{T} = 0.1$

$$C_{\text{FBA}} = \int_0^\infty \overline{B} \log_2(1+\gamma) P(\gamma) d\gamma$$
$$C_{\text{FTA}} = \int_0^\infty \overline{T} B_{\max} \log_2(1+\gamma) P(\gamma) d\gamma \qquad (34)$$

In Figures 2 and 3, it is observed that both dynamic bandwidth and time allocations provide better average capacity performance than their fixed counterparts. As our analysis has shown, the optimal DBA capacity is independent of the channel fading condition, while the DTA capacity does depend on the channel condition. It is observed in both figures that the capacity gain over the fixed allocation schemes diminishes as the level of channel variance decreases.

Figure 4 shows the ratio of DBA and DTA capacities as a function of $E[\gamma]$, which is measured with respect to $\overline{B}$. In order to make a fair comparison between the two schemes, $\overline{B}$, $B_{\max}$ and $\overline{T}$ are chosen such that $\overline{B}/B_{\max} = \overline{T}$. This ensures that the average resource (time × bandwidth) allocation is the same for both schemes. It is observed that DBA outperforms DTA regardless of the channel condition. This is because both schemes are assumed to use the same constant maximum power when transmitting (which is a practical assumption). The discrepancy in performance stems from employing continuous transmission (DBA) vs. on-and-off transmission (DTA). This means DBA, on average, consumes $1/\overline{T}$ times more power than DTA. In short, DBA offers better average capacity performance while DTA consumes less power.

Figures 5 and 6 show the effects of multiple-access contention on average DBA and DTA capacity performances, respectively. Rayleigh fading with $E[\gamma]$ = 30 dB (with respect to $\overline{B}$) is used and $\overline{B}/B_{\max}$ is set equal to $\overline{T} = 0.05$. It is assumed that all users have the same target resource requirement and SNR (channel) distribution. In both cases the residual capacities are ignored. It can be seen that the average capacity performance suffers as the number of users (i.e. contention probability) increases. The figures also compare the performances of several contention policies described in Table 1, as well as the optimal single user and fixed allocation capacities. Note that for fixed allocation schemes, there is no contention as long as

$$\sum_{i=1}^{N} \overline{B}_i \le B_{\max}, \text{ for FBA}$$
$$\sum_{i=1}^{N} \overline{T}_i \le 1, \text{ for FTA} \qquad (35)$$

Policy A for both cases prevents all users from transmission during contention. Therefore, Policy A produces the lower bounds on the average performance of our DBA/DTA schemes, whereas the optimal capacity curves correspond to the upper bounds. Multiple-access performance for DBA/DTA, which depends on specific contention policies being employed, always lies between these two bounds as shown in the figures. It is observed that there exist policies that generally outperform the fixed allocation schemes; utilizing the residual capacities would improve the DBA/DTA capacities further. While there is no "optimal' contention policy, the contention

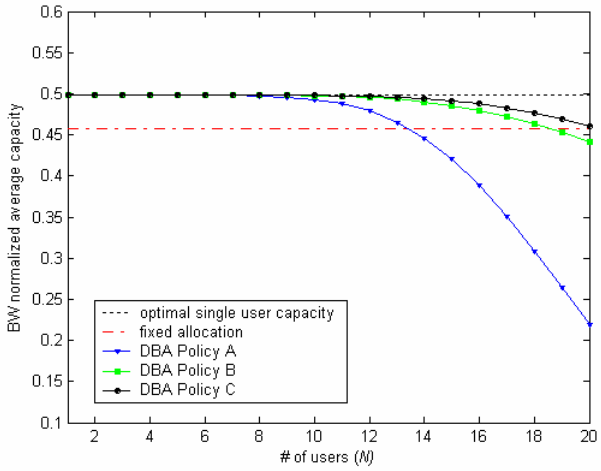

Figure 5. Average capacity performance comparison among various DBA contention policies; Rayleigh fading

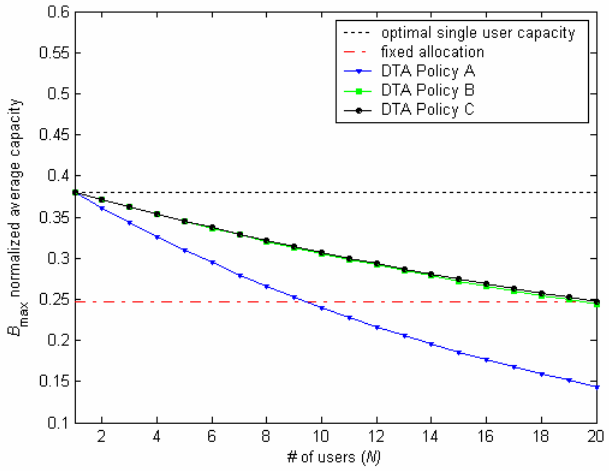

Figure 6. Average capacity performance comparison among various DTA contention policies; Rayleigh fading

policy should be chosen based on user QoS requirements. The proposed schemes offer two distinct ways to control QoS outcomes; resource constraints and contention policy.

Table 1. Contention policies whose results are shown in Figures 5 and 6

	DBA	DTA
Policy A	No transmission	No transmission
Policy B	Allocation based on target requirements	Random allocation
Policy C	Allocation based on both target and SNR	Highest SNR transmits

Figures 7 and 8 show how the proposed cross-layer resource management techniques can be used to satisfy different user QoS requirements. The figures illustrate the DBA and DTA average delay results for a system of ten users. Users 1 through 6 have less stringent delay requirements than users 7 through 10, which would be the case for instance if the former were predominantly data users and the latter were voice users. It is assumed that all ten users have the same channel (Rayleigh) and traffic arrival distributions. Optimal allocation based contention polices are used to obtain both results. It is observed that the cross-layer technique can shape the delay outcome to meet specific QoS requirements by jointly utilizing data and channel statistics. The average resource constraint for each user is generated by the network layer, and the physical layer dynamically adjusts the user's instantaneous resource allocation based on the derived optimal rules. Without cross-layer exchange, if resource allocation is solely based on channel statistics (as in [11]), all users experience the same average delay in both cases.

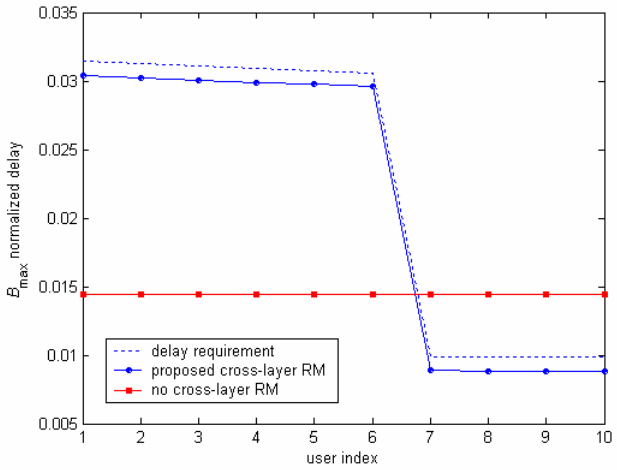

Figure 7. DBA user delay requirement adaptation via cross-layer implementation

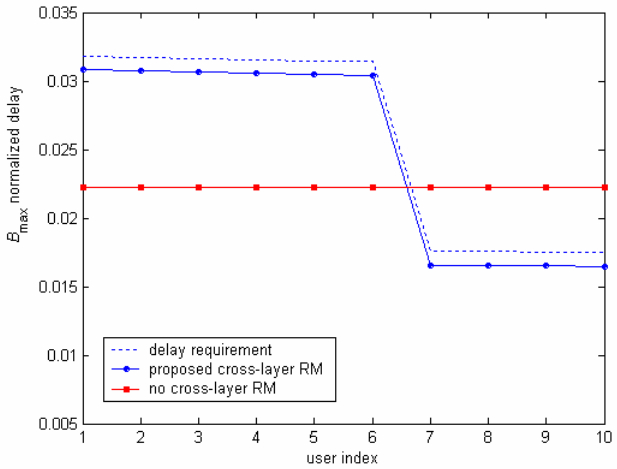

Figure 8. DTA user delay requirement adaptation via cross-layer implementation

F. CONCLUSIONS AND FUTURE STUDIES

In this paper, we have proposed a novel CLD approach based on the fundamental relationship between physical layer resource (bandwidth and time) and network performance. In this proposed scheme, the QoS-compliant average resource constraint for each user is determined by the network layer taking channel and traffic statistics into account. The physical layer then dynamically adjusts the user's instantaneous resource assignment based on the derived optimal allocation rule. Our method is inherently flexible satisfying different user QoS requirements.

Our future studies include a plan to extend our approach to cover wider physical channel environments and additional network performance measures. We are currently studying the application of our DBA/DTA framework to frequency-selective channels. One approach may be to divide the channel into several frequency flat sub-channels and apply DBA/DTA schemes on an individual sub-channel basis. We are also working on developing analytical models to incorporate additional QoS measures such as outage probability and delay jitter. These measures, which involve identifying the second moments, are especially relevant for dynamic resource allocation schemes whose performance variability is greater than their fixed allocation counterparts.

REFERENCES

[1] I. E. Telatar and R. G. Gallager, "Combing queuing theory with information theory for multi-access," *IEEE JSAC*, Vol. 13, No. 6, pp. 963-969, Aug. 1995

[2] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad hoc networks," *Wireless communications, IEEE Transactions on* , Vol. 3, No. 1, pp. 74-85, Jan. 2004

[3] A. Maharshi, T. Lang and A. Swami, "Cross-layer designs of multichannel reservation MAC under Rayleigh fading," *Signal Processing, IEEE Transactions on*, Vol. 51, No. 8, pp. 2054–2067, Aug. 2003

[4] S. Toumpis and A. J. Goldsmith, "Performance, optimization, and cross-layer design of media access protocols for wireless ad hoc networks," *ICC '03*, pp. 2234-2240, May 2003

[5] S. Shakkottai and T. Rappaport, "Cross-Layer Design for Wireless Networks", *IEEE Comm. Mag.*, pp. 74-80, Oct. 2004

[6] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. on Communications*, Vol. 45, No. 10, pp. 1218-1230, Oct. 1997

[7] R. G. Gallager, *Information theory and reliable communication*, Wiley & Sons, New York, NY, 1968

[8] Y.-D. Yao and A. Sheikh, "Investigations into cochannel interference in microcellular mobile radio systems," *IEEE Trans. on Vehicular Technology*, Vol. 41, No. 2, pp. 114-123, May 1992

[9] A. Abu-Dayya and N. C. Beaulieu, "Outage probabilities of cellular mobile radio systems with multiple Nakagami interferers," *IEEE Trans. on Vehicular Technology*, Vol. 40, No. 4, pp. 757-768, Nov. 1991

[10] W. Yu and J. M. Cioffi, "FDMA capacity of Gaussian multiple-access channel with ISI," *IEEE Trans. on Communications*, Vol. 50, No. 1, pp. 102-111, Jan. 2002

[11] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," *Proc. IEEE ICC'95*, Seattle, Wash., 1995

[12] D. Kivanc, G. Li and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. on Wireless Communications*, Vol. 2, No. 6, pp. 1150-1158, Nov. 2003

[13] W. Yu and J. M. Cioffi, "On constant power water-filling," *Proc. IEEE ICC'2001*, June 2001

[14] L. Kleinrock, *Queueing Systems: Volume II: Computer Applications*, John Wiley and Sons, 1976

[15] A. Safwati, H. Hassanein and H. Mouftah, "Optimal cross-layer designs for energy-efficient wireless ad hoc and sensor networks," *Performance, Computing, and Communications Conference*, 2003

[16] G. Carneiro, J. Ruela and M. Ricardo, "Cross-layer design in 4G wireless terminals," *Wireless Communications, IEEE Transactions on*, Vol. 11 , No. 2, pp. 7-13, Apr. 2004