# On the Relationship Between Queuing Delay and Spatial Degrees of Freedom in a MIMO Multiple Access Channel

Sriram N. Kizhakkemadam, Dinesh Rajan, Mandyam Srinath
Dept. of Electrical Engineering
Southern Methodist University
Dallas, TX - 75205
Email: {skizhakk, rajand, mds}@engr.smu.edu

***Abstract*—In this paper, we study the relationship between queuing delay for a random packet arrival process and physical layer parameters in a multiple access channel with multiple antennas at the input and output. Our main contribution is the derivation of a simple, analytical approximation for the average delay that clearly indicates the effect of number of transmit and receive antennas, transmission power, the packet arrival rate and the desired reliability. Comparison with numerical analysis indicates that the proposed analytical approximation of the delay is accurate for medium and large SNRs.**

## I. Introduction

The information age has seen a profusion of wireless mobile devices that has allowed users to communicate using voice, data and even multimedia. A proper characterization of the achievable performance is required so that users can then be given various service guarantees against a price differential. Some of the parameters of the Quality of service include data rate of transmission, the fidelity of reproducing the information at the receiver(distortion), reliability of the data link (Probability of error), the delay in the availability of the information and the power required for information transmission.

Until the early 90's, majority of the research in both network information theory and networking implicitly assumed the Open Systems Interconnection Basic Reference Model (OSI Model for short) of a network. The OSI model is an layered, abstract description for communications and protocol design. In this model, a networking system is divided into various layers. Within each layer, one or more entities implement its functionality. Each entity interacts directly only with the layer immediately beneath it, and provides facilities for use by the layer above it. The communication between entities in different layers and possibly different hosts is through protocols [2]. The study of the impact of physical layer parameters on the network layer has gained importance of late in the field of cross-layer optimization.

The delay at the physical layer was typically considered to be the time required to transmit and decode a codeword with or without ARQ (Automatic Repeat Request). The effect of random arrival of messages on the physical layer was not given importance. For multi-user systems, Telatar and Gallager laid the foundation of an elegant formulation to determine the queuing delay in a single antenna multiple access channel (MAC) in their seminal paper [3]. This paper effectively dispensed of with the independence notion of the different layers of the OSI model. A processor sharing queue was used to model the queuing in a MAC. In this model, the service process is constant with time. The case where the service rate fluctuates was studied in [4]. Details of the Telatar and Gallager model will be discussed later. Recently more efforts on understanding the queuing delay for MIMO systems have been undertaken [5], [6]. In [6], we studied the queuing delay of a MIMO MAC with single user decoding (SUD). Our results from numerical analysis showed that depending on the number of receive antennas, increasing the number of transmit antennas can either increase or decrease the delay. Is this a limitation of single user decoding? Is there a simple relationship between queuing delay and the spatial parameters? Motivated by these questions, we study the queuing delay with joint decoding (JD) in this paper. JD helps us give a lower bound on the minimum achievable delay. In SUD, the received signal from other users is treated as noise. Instead, in JD, we use apriori knowledge of the set of all possible codeword from all users at the receiver for say, successive interference cancellation [7].

Our main contribution is the derivation of a closed form expression for the queuing delay with joint decoding that accurately matches with numerical evaluation in the limit of large SNR. The analytical expression for delay clearly shows the dependence on system parameters like SNR, number of antennas, probability of error, message size and arrival rate of packets. We observe that the delay is inversely proportional to the minimum of number of transmit and receive antennas.

The system model we espouse is in close concordance with that given in [3]. Notably, the model chosen in [3] bridges models in information theory and queuing theory with a sense of engineering intuitive appeal. After describing the system model in section 2, we summarize the relationship between queuing delay and random coding error exponent in section 3. In section 4, we derive an approximate analytical relationship between the delay and number of transmit and receive antennas that is valid in the high SNR regime. We investigate the accuracy of the approximations in section 5 and conclude in section 6.

## II. System Model

Consider a symmetric MIMO MAC with multiple users each of which has $M$ transmit antennas. All the users transmit

their information to a single receiver with $N$ receive antennas. Messages are generated according to a Poisson process with rate $\lambda$. Each new message is considered to represent a virtual user; each message is of size $\log K$ nats. Thus, we have in effect possibly infinite number of transmitters. This abstraction of a user transmitting a single message is useful in modeling the system as a processor sharing queue with the queue at the transmitter being effectively modeled as a queue at the receiver for analytical tractability [3]. Due to this virtual user/transmitter model, we use the terms user and transmitter interchangeably. We denote by $u(t)$, the random variable associated with the number of users in the system at time $t$. Each user encodes its message into an infinite length codeword for transmission. However, the entire codeword is not transmitted. After successfully decoding the transmitted codeword, the receiver sends a signal to the appropriate user to cease transmission. In the limit of large codeword lengths, this 1 bit of feedback is negligible.

The received signal $\mathbf{y} \in \mathbb{C}^N$ at the $N$ receive antennas at time $t$ depends on the transmitted signal according to,

$$\mathbf{y}_t = \sum_{i=1}^{u(t)} \sqrt{\frac{\mathsf{SNR}}{M}} \mathbf{H}_i \mathbf{x}_{i,t} + \mathbf{z}_t, \tag{1}$$

where $\mathbf{x}_{i,t}$, $\mathbf{H}_i$ and $\mathbf{z}_i$ are the normalized input, channel and noise random variables. The normalized input vector of the $i^{th}$ user, $\mathbf{x}_i$ has entries distributed according to a complex Gaussian distribution with 0 mean and unit variance. The real and imaginary entries of the channel matrix $\mathbf{H}_i \in \mathbb{C}^{N \times M}$ are Gaussian distributed with 0 mean and unit variance. The additive uncorrelated white Gaussian noise, $\mathbf{z}$ is distributed according to $\mathcal{CN}(0,1)$. The bandwidth at the receiver is $W$ Hz. The factor of $\mathsf{SNR}$ takes into account arbitrary channel gains and non-unit receiver noise power. We assume that the transmitter has no knowledge of the channel and hence the transmitter of each user allocates the total available transmit power equally to all the transmit antennas. However, the receiver has perfect channel state information (CSIR). We also denote by $\mathcal{H}_u \triangleq \{\mathbf{H_1}, \ldots, \mathbf{H_u}\}$.

The symmetric MIMO MAC with same number of transmit antennas and power constraints for all users has been chosen for analytical convenience. Only natural logarithms are considered in this paper. In this paper, we consider the channel to be fast fading where the channel codeword is coded across multiple coherence time intervals (fading blocks). We now give a summary description of modeling the queuing delay from an information theoretic perspective in the following section.

## III. Queuing Delay

The theory of error exponents is used in [3] to relate queuing and physical layer parameters. The error exponent term $E_o$ is related to the service rate while the specification on a desired probability of error and message size is related to the demand. Based on a symmetric queue processor sharing model as given in [8], a numerical evaluation of the end-to-end delay was obtained in [3]. Specifically, in [3], the random coding bound on the probability of error is used to draw a parallel between the random coding exponent term $E_o$ and the service rate for the multiple access queue. From definition of random coding error exponent for continuous channels [9], the random coding bound is given by

$$P_e \leq \exp \max_{0 \leq \rho \leq 1} \left[ \rho \log K - \sum_{i=1}^{u(t)} E_o\left(\rho, \mathcal{H}_u, q_x\right) \right] \tag{2}$$

The error exponent term $E_o$ is,

$$E_o = \max_{q_x} -\log \int_{\mathcal{H}_u} \int_y \left[ \int_x q_x p\left(y, \mathcal{H}_u | x\right)^{1/(1+\rho)} dx \right]^{1+\rho} dy$$

Since the channel is known at the receiver, $E_o$ simplifies to,

$$E_o = \max_{q_x} -\log \mathcal{E}_{\mathcal{H}u} \int_y \left[ \int_x q_x p\left(y|x, \mathcal{H}_u\right)^{1/(1+\rho)} dx \right]^{1+\rho} dy \tag{3}$$

where, the maximum is over all input distributions $q_x$ of the codeword and

$$p\left(y|x, \mathcal{H}_u\right) = \frac{1}{|\pi \mathbf{I}_N|} \exp \left[ - \left( \mathbf{y} - \sum_{i=1}^{u(t)} \mathbf{H_i x} \right)^{\dagger} \left( \mathbf{y} - \sum_{j=1}^{u(t)} \mathbf{H_j x} \right) \right] \tag{4}$$

In general, the $q_x$ that maximizes the error exponent is given by a distribution concentrated on a "thin spherical shell" [10]. However, choosing $q_x$ as the Gaussian distribution lends itself useful for simplified expressions and a convenient lower bound on the $E_o$(and consequently an upper bound on the probability of error). Therefore, we consider the distribution of the $\mathbf{x}$ to be zero mean Gaussian with covariance $\mathbf{Q} = \mathcal{E}\left[\mathbf{x}\mathbf{x}^{\dagger}\right]$. Since, we assume that channel knowledge is not available at the transmitter, $\mathbf{Q} = \frac{\mathsf{SNR}}{M} \mathbf{I}_M$. The function $E_o$ is related to the queuing parameters as given below.

By making use of the formalism in [3], each user who communicates with the receiver is modeled as a job in a processor sharing queue. For a queuing delay analysis, we need to express random message arrival parameters and channel parameters in terms of demand and supply. By taking logarithm on both sides of (2) and rearranging the expression, we can write the demand per unit bandwidth for a tolerable error probability and message length ($\log K$) as,

$$S = -\log P_e + \rho \log K \tag{5}$$

The remaining term after taking logarithm on both sides of (2), ($\sum_{i=1}^{u(t)} E_o$) can be treated as accumulated service. The service rate at time $t$ can be obtained by evaluating the error exponent at time $t$ and scaling it with the bandwidth. Thus, the units of the service rate will be bits/sec.

The service rate at time $t$ with joint decoding can therefore be written as,

$$\phi(u) = W E_o\left(\rho, \mathbf{Q}, \mathcal{H}_u\right) \tag{6}$$

In [3], since SUD was considered, the sum service rate was the sum of the service requirement of the individual users. Assuming identical channel distribution for all users, the sum service rate for SUD was given as $\phi(u) = u(t).W.E_o(\rho, \mathbf{Q}, \mathbf{H}_1)$. However, since we are considering joint decoding, all the users are decoded together and hence we compute the joint decoding error exponent $E_o(\rho, \mathbf{Q}, \mathcal{H}_u)$.

The average delay in servicing these users(or the average transmission duration) is obtained by Little's Law [11] which relates the delay to the average number of users in the system and the arrival rate of the users. The average delay is given by,

$$\bar{D} = \mathcal{E}[U]/\lambda, \tag{7}$$

where the distribution of the number of users is given by [3],

$$Pr\{\text{u jobs in the system}\} = \frac{1}{C\phi_!(u)} (\lambda \mathcal{E}[S])^u \tag{8}$$

with,

$$\phi_!(u) = \prod_{\nu=1}^{u(t)} \phi(\nu) \text{ and } C = 1 + \sum_{u=1}^{\infty} (\lambda \mathcal{E}[S])^u / \phi_!(u) \tag{9}$$

A condition for stability of the system in this case can be derived in a manner similar to [3] and is not shown here. The loading of the queue is defined as, $l \stackrel{\Delta}{=} \lambda \mathcal{E}[S]$. The expectation in (9) is over the distribution of the number of codewords for different users. If we make a simplifying assumption that the cardinality of the set of codewords for each user is $K$ and substitute for $\mathcal{E}[S]$ from (5),

$$l = \lambda [\rho \log K - \log P_e] \tag{10}$$

The average delay can be numerically evaluated by computing the error exponent averaged over various channel realizations for a given probability of error, arrival rate and number of codewords. In order to obtain a closed form expression for the delay, we need a closed form expression for the error exponent. The following section makes use of the scaling law in the limit of large SNR in order to obtain an approximate expression for the average queuing delay.

## IV. Analytical Approximation

We now derive analytical approximations for the average delay for a MIMO MAC. A closed form expression for the service rate and consequently the error exponent term $E_o$ is the key to a closed form expression for the delay. The error exponent term $E_o$ is akin to the sum rate capacity expression of a MIMO MAC except for a scaling of the SNR term. By making use of Jensen's inequality in (3), it can be easily shown that the error exponent for the MIMO MAC with joint decoding is given by [12],

$$E_o \leq \mathcal{E}_{\mathcal{H}}\ \rho \log \left| \mathbf{I}_N + \frac{\sum_{i=1}^{u(t)} \frac{\mathsf{SNR}}{M} \mathbf{H_i Q H_i^\dagger}}{1+\rho} \right|, \tag{11}$$

where the free parameter $\rho$ can be numerically selected over $[0, 1]$ to either minimize delay [3] or to minimize the probability of error [9]. Numerical analysis indicates that $\rho = 1$ maximizes the error exponent and hence we set $\rho = 1$. We now make use of scaling laws in the limit of large SNR to obtain closed form expressions for $E_o$. The scaling law for a fast fading MIMO MAC was mentioned in [10] as,

$$\lim_{\mathsf{SNR}\to\infty} \mathcal{E}_{\mathbf{H}} \log \left| \mathbf{I} + \sum_{i=1}^{u} \frac{\mathsf{SNR}}{M} \mathbf{H_i H_i^\dagger} \right| \approx \ \min(uM, N) \log \frac{\mathsf{SNR}}{M} \tag{12}$$

The large SNR assumption is useful in obtaining performance limits for characterizing the queuing delay.

Let us define by

$$\alpha_r = W.N. \log \frac{\mathsf{SNR}}{M},\ \alpha_t = W.M. \log \frac{\mathsf{SNR}}{M} \tag{13}$$

In terms of $\alpha_t$ and $\alpha_r$, the service rate can be expressed as,

$$\begin{aligned} \phi(u) &= W E_o(\rho, \mathbf{Q}, \mathcal{H}_u) \\ &\approx W \min(uM, N) \log \frac{\mathsf{SNR}}{M(1+\rho)} \\ &\approx \min(u.\alpha_t, \alpha_r) \end{aligned} \tag{14}$$

In the limit of large SNR, we neglect the scaling of the SNR term by $(1+\rho)$ in (14).

*Theorem 1*: The average delay for a MIMO MAC under fast fading with no CSIT and full CSIR is approximately given by

$$\begin{aligned} \overline{D} &\approx \frac{l}{\lambda[\alpha_r - l]} \text{ for } N \leq M \\ &\quad \frac{1}{\lambda}\left(\frac{l}{\alpha_t}\right) \text{ for } M < N \end{aligned} \tag{15}$$

The approximation is accurate for large SNR and if the stability conditions are satisfied, viz.,

$$l < \min(\alpha_t, \alpha_r) \tag{16}$$

*Proof:*

The average delay is given by

$$\overline{D} = \frac{1}{\lambda} \sum_{u=1}^{\infty} \frac{u(\lambda \mathcal{E}[S])^u}{C\phi_!(u)} \tag{17}$$

By making use of (14), the delay expression can be simplified further. Since the scaling law for the service rate depends on the number of users, we consider two cases, $N < M$ and $\delta M < N \leq (\delta+1)N$, $\delta \in \mathbb{Z}^+$.

*Case A:* If $N < M$,

$$\phi(u) \approx \alpha_r\ \forall\ u\ \Rightarrow\ \phi_!(u) \approx \alpha_r^u \tag{18}$$

and

$$C = 1 + \sum_{j=1}^{\infty} \left(\frac{l}{\alpha_r}\right)^j = \frac{1}{1 - \frac{l}{\alpha_r}} \tag{19}$$

The convergence of the infinite geometric series in (19) is valid when $l/\alpha_r < 1$. The average delay is therefore,

$$\begin{aligned}\overline{D} &= \frac{1-\frac{l}{\alpha_r}}{\lambda}\sum_{j=1}^{\infty} u.\left(\frac{l}{\alpha_r}\right)^u \\ &= \frac{1-\frac{l}{\alpha_r}}{\lambda}\frac{\frac{l}{\alpha_r}}{\left(1-\frac{l}{\alpha_r}\right)^2} \\ &= \frac{l}{\lambda\left[\alpha_r - l\right]} \end{aligned} \quad (20)$$

*Case B*: $\delta M \leq N, (\delta+1)M > N$, $\delta \in \mathbb{Z}^+$
If $u > \delta$,

$$\begin{aligned}\phi_!(u) &= \left[WM\log\frac{\mathsf{SNR}}{M}\right]^{\delta}\delta!\left(WN\log\frac{\mathsf{SNR}}{M}\right)^{u-\delta} \\ &= \left[WN\log\frac{\mathsf{SNR}}{M}\right]^{u}\delta!M^{\delta}N^{-\delta} \\ &= \alpha_r^u\delta!\beta^{\delta} \end{aligned} \quad (21)$$

where we have denoted $(M/N)$ as $\beta$.
For $u < \delta$,

$$\begin{aligned}\phi_!(u) &= \left[W\ M\ \log\frac{\mathsf{SNR}}{M}\right]^{u}.u! \\ &= \alpha_t^u\ u! \end{aligned} \quad (22)$$

The normalizing constant $C$ can therefore be written as,

$$C = 1+\sum_{j=1}^{\delta}\left(\frac{l}{\alpha_t}\right)^j\frac{1}{j!}+\sum_{j=\delta+1}^{\infty}\left(\frac{l}{\alpha_r}\right)^j\frac{1}{\delta!\beta^{\delta}} \quad (23)$$

$$\approx \exp\left(\frac{l}{\alpha_t}\right)+\frac{1}{\delta!\beta^{\delta}}\frac{\left(\frac{l}{\alpha_r}\right)^{\delta+1}}{1-\left(\frac{l}{\alpha_r}\right)} \quad (24)$$

The approximation of the first power series summation in (23) by an exponential series is valid for $l/\alpha_t < 1$. The convergence of the second term in power series expansion in (23) is valid for $l/\alpha_r < 1$.

The average delay can therefore be written as,

$$\begin{aligned}\overline{D} &= \frac{1}{\lambda\ C}.\left[\sum_{u=1}^{\delta}\frac{u\ (l/\alpha_t)^u}{u!}+\sum_{u=\delta+1}^{\infty}\frac{u\ (l/\alpha_r)^u}{\delta!\ \beta^{\delta}}\right] \\ &= \frac{1}{\lambda\ C}.\left[\frac{l}{\alpha_t}\ \exp\left(\frac{l}{\alpha_t}\right)+\right. \\ &\quad \left.\frac{\left[(\delta+1)\left(\frac{l}{\alpha_r}\right)^{\delta+1}-\delta\left(\frac{l}{\alpha_r}\right)^{\delta+2}\right]}{\delta!\beta^{\delta}\left[1-\frac{l}{\alpha_r}\right]^2}\right] \\ &\approx \frac{1}{\lambda}\left(\frac{l}{\alpha_t}\right) \end{aligned} \quad (25)$$

By combining the criteria for stability in Case A and B, we get the stability condition as $l < \min(\alpha_t, \alpha_r))$.

*Comment 1:* If $M < N$, the delay is inversely proportional to the number of transmit antennas. Thus, increasing the number of transmit antennas decreases the delay. This decrease in delay is unlike the non-monotonic behavior seen if single-user decoding is used [6]. Single-user decoding limits the maximum service rate to 1 [3]. Increasing the number of transmit antennas with SUD decreases the service rate further. In contrast, with joint decoding, the delay decreases with addition of antennas at user terminals as seen in (25) and (20). Therefore, to harness the benefit of multiple antennas for delay in a multiple access setting, joint decoding is critical. The plot of the average delay v/s the number of transmit and receive antennas is shown in Fig. 1 and Fig. 2.

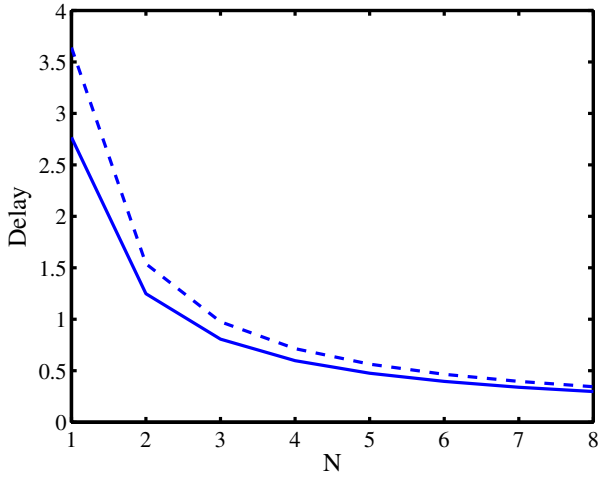


Fig. 1. Plot of delay v/s number of receive antennas, N with 8 transmit antennas at SNR of 40dB and a desired probability of error of $10^{-5}$. The solid line is numerical evaluation while the dashed line is analytical approximation.

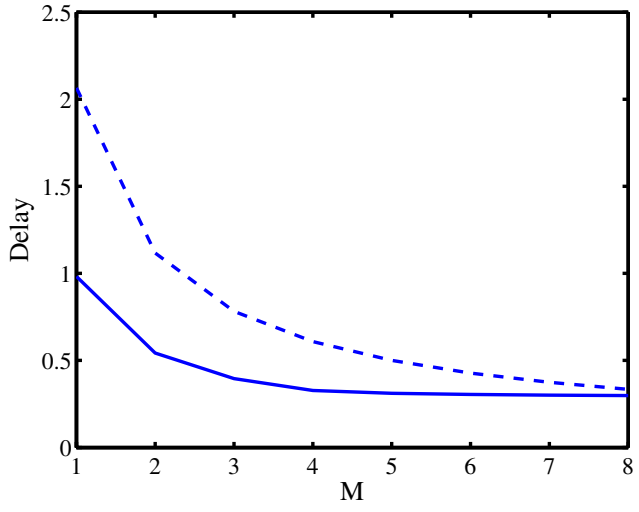


Fig. 2. Plot of delay v/s number of transmit antennas, M with 8 receive antennas at SNR of 40dB and a desired probability of error of $10^{-5}$. The solid line is numerical evaluation while the dashed line is analytical approximation.

## V. Numerical Evaluation

In Figs.1 and 2, the plots of the average delay v/s the number of transmit and receive antennas indicate that the approximation is valid for high SNR. In Fig. 3, we plot the delay v/s SNR in dB for a $2 \times 2$ and a $4 \times 4$ system with a desired probability of error of $10^{-5}$. The arrival rate chosen was $0.1$ which satisfied the stability condition. As expected, we observe that the delay decreases with increase in the SNR due to the inverse relationship with number of antennas (20). The closed form expression is accurate for medium to large SNR. If $N > M$, the same trend is seen.

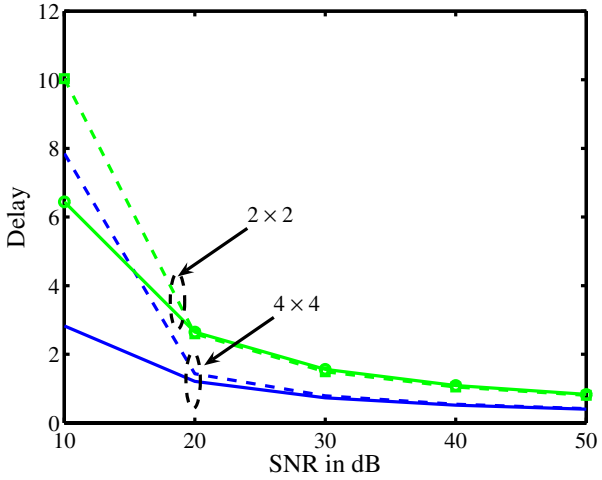


Fig. 3. Plot of Average delay v/s SNR in dB for a $2 \times 2$ and $4 \times 4$ system. The solid lines are numerical evaluation while the dashed lines are the analytical approximation. $\lambda = 0.1$

In Fig. 4, we plot the average delay v/s the probability of error by making use of (20). Increasing the specification on the probability of error causes the delay to increase. An arrival rate of 0.1 was chosen. With decrease in $P_e$, the loading of the system, $l$ increases resulting in an increase in the value of the numerator in the approximate expression (20), (25). The probability of error term appears in the denominator too. But, it is scaled by the arrival rate and is negligible compared to the $\alpha_r$ term that is asymptotic in SNR. Intuitively speaking, a lower value of probability of error means that the receiver should accumulate successively more values of $E_o$ to meet the specification. This would mean that the transmitter has to send more bits and consequently the increase in the delay. Following the trend of Fig. 1, increasing the SNR, decreases the delay and the approximation becomes more tighter. If $N > M$, the trend is again similar, but the approximate expression is tighter as $M$ increases.

## VI. Conclusion

In this paper, we derived a simple closed form approximation for the queuing delay in a MAC with multiple transmit and receive antennas. Our results show that the average delay is inversely proportional to the minimum of the number of

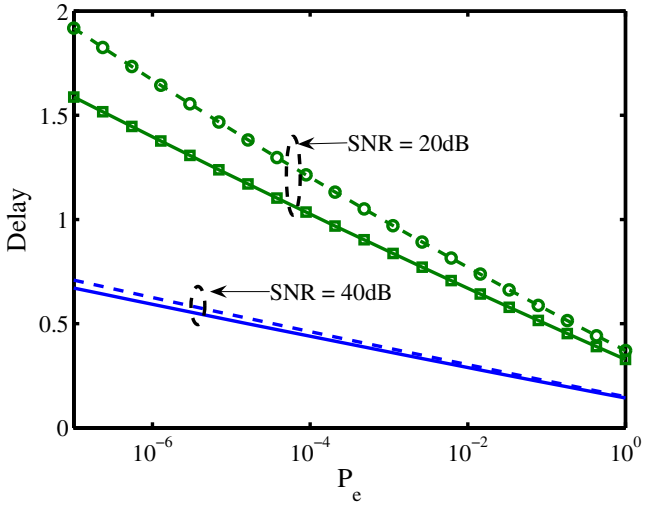


Fig. 4. Plot of Average Delay v/s Probability of error for a $4 \times 4$ system at 20dB and 40dB SNR. The solid lines are numerical evaluation while the dashed lines are the analytical approximation. $\lambda = 0.1$

transmit and receive antennas. Comparison with numerical analysis indicates that the proposed analytical approximation of the delay is accurate for medium and large SNRs. The existence and design of coding strategies that achieve the performance predicted by using multiuser decoding in the proposed scenario should be investigated in future work. Future work should also consider other arrival processes with long range dependence.

## Acknowledgment

This work has been supported in part by NSF under grant CCF 0546519.

## References

[1] S. Shenker, C. Partridge, and R. Guerin, "Specification of guaranteed quality of service," in *RFC 2212: Internet Eng. Task Force*, Sept. 1997.

[2] J. Kurose and K. Ross, *Computer Networking: A Top Down Approach Featuring the Internet*. Addison-Wesley, 2002.

[3] I. Telatar and R. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE Journal on Selected Areas in Comm.*, vol. 13, pp. 963–969, Aug. 1995.

[4] R. Prakash and V. V. Veeravalli, "Centralized wireless data networks with user arrivals and departures," *IEEE Trans. on Info. Theory*, vol. 53, pp. 695–713, Feb. 2007.

[5] P. Elia, S. Kittipiyakul, and T. Javidi, "On the responsiveness-diversity-multiplexing tradeoff," in *WiOpt*, April 2007.

[6] S. N. Kizhakkemadam and D. Rajan, "Queuing aspects of multiantenna multiple access channels," in *IEEE Globecom*, (San Francisco), Nov.-Dec. 2006.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[8] F. P. Kelly, *Reversibility and Stochastic Networks*. New York: Wiley, 1979.

[9] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[10] I. E. Telatar, "Capacity of multi-antenna Gaussian channel," *European Tran. on Telecom.*, vol. 10, pp. 585–595, Nov./Dec. 1995.

[11] D. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice Hall, 2 ed., 1992.

[12] T. Guess and M. Varanasi, "Error exponents for the Gaussian multiple-access channel," in *IEEE ISIT*, p. 214, Aug. 1998.