An Approximate Queueing Model for Multi-rate Multi-user MIMO systems*

Boris Bellalta,Vanesa Daza, Miquel Oliver

Abstract

A queueing model for Multi-rate Multi-user MIMO systems is presented. The model is built upon the assumption that the probability distribution of available destinations among the buffered frames at the Base Station (BS) is approximately the same as the probability distribution of the traffic arriving to the BS, this is, the amount of traffic directed to each MN with respect to the total traffic load. This assumption leads to a simple, but accurate, queueing model for Multi-user MIMO systems that accounts for the impact of a finite number of active MNs in non-saturated conditions. The model is easily applicable to any Multi-user MIMO scenario given that the probability density function of the post-processing SINR (Signal to Interference and Noise Ratio) for each MN is known.

Information about this paper:

- This is an updated version of the paper published in IEEE Communications Letters: Boris Bellalta, Vanesa Daza, Miquel Oliver; 'An Approximate Queueing Model for Multi-rate Multi-user MIMO systems'. IEEE Communications Letters, 2011.
- The Matlab code for the described model can be found at: http://www.dtic.upf.edu/∼bbellalt/. For any comment about, please contact Boris Bellalta (boris.bellalta@upf.edu).

*This work has been partially supported by the Spanish Government under projects TEC2008-0655 (Plan Nacional I+D), TEC2009-13000 (Plan Nacional I+D+i), CSD2008-00010 (Consolider-Ingenio Program) and by the Catalan Government (SGR2009#00617).
The authors are with the Wireless Networks Research Group, Dept. de Tecnologies de la Informacio i les Comunicacions, Universitat Pompeu Fabra, Spain (email:{boris.bellalta,vanesa.daza,miquel.oliver}@upf.edu).

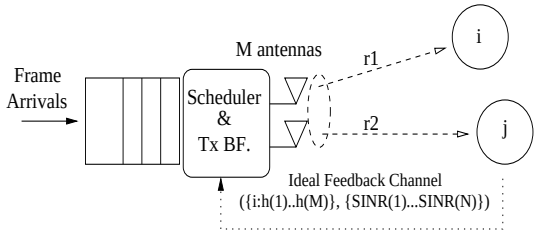

Figure 1: Specific scenario with $M = 2$ antennas serving frames to N MNs

1 Introduction

In a Multi-user MIMO system, a BS equipped with M antennas is able to transmit simultaneously up to M frames to M different single-antenna MNs, achieving a maximum spatial multiplexing gain of M [1]. However, this is only possible if two conditions are satisfied: *i*) the BS has at least M frames stored in its transmission queue and *ii*) these stored frames are destined to at least M different MNs. On the contrary, the number of stored frames or the available destinations among them will limit the maximum number of frames that can be transmitted in parallel. Therefore, the queueing process (how the number of queued frames evolves with the time, which depends on both the arrival and service processes) has to be considered in detail to really understand the performance that Multi-user MIMO systems can provide.

In this letter, a queueing model for Multi-rate Multi-user MIMO systems in non-saturated conditions and with a finite number of active MNs is presented. From the physical layer, the model relies on the knowledge of the post-processing SINR distribution [2] at each MN, independently of the precoding technique is used at the BS.

2 System Model and Assumptions

A BS with M antennas and a finite-buffer of size K frames are considered (Figure 1). Frames of length equal to L bits (constant) directed to the N single-antenna MNs arrive to the BS following a Poisson process with aggregate rate λ, equally distributed among all active destinations.

A general precoding scheme is used at the BS, fed with the Channel State Information (CSI) acquired at each MN and sent to the BS through an ideal and instantaneous feedback channel. Based on those CSI values, a transmission rate, picked from a finite set of rates $\mathcal{R}$, is used for the communication between the BS and the MNs in a frame-by-frame basis.

A FIFO-based scheduling (frame selection) algorithm is considered. It selects $\varsigma \in [1, min(q, M)]$ frames from those $q \in [1, K]$ frames stored in the queue when a new transmission is scheduled, which happens just after the previous one has finished or, if the queue has become empty, immediately after the arrival of a new frame. Specifically, the BS always schedules the first frame waiting for transmission and then, it selects sequentially up to $min(q, M) - 1$ frames, directed to not yet selected destinations. The group of selected frames at each transmission is called a space-batch.

In Figure 2, a specific example of the queue behavior is shown for $M = 2$ antennas, $K = 3$ frames and $\mathcal{R} = \{r_1, r_2\}$ rates. Observe that the $(i + 1)$-th space-batch comprises two frames, one transmitted at r_1 and one at r_2 respectively, resulting in a space-batch duration of L/r_1 seconds as r_1 is the minimum transmission rate in the $(i + 1)$-th space-batch. In general, let $r = min\{r_1, \ldots, r_\varsigma\}$ be the lowest transmission rate assigned to one of the frames included in a certain space-batch and $p(r, \varsigma)$ its associated probability. The dependence with ς in $p(r, \varsigma)$ is included to account for the impact that the number of spatially multiplexed streams can have on the assigned transmission rates.

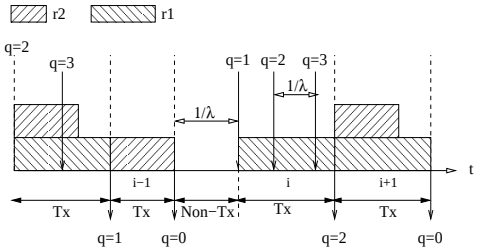

Figure 2: Temporal evolution of the BS's queue. The q values are the queue state (number of frames in the queue) just after a frame arrival or a space-batch departure. Observe that $r_2 > r_1$.

3 The Queueing Model

A $M/G^{[1,M]}/1/K$ batch-service queue [3] is used to model the BS (for further information, refer to [4] and references therein). The goal of the presented model is to analytically obtain the system delay (average time that a frame remains in the BS) and the blocking probability (probability that a new arriving frame can not be stored in the queue as there is no free space in it) for the considered Multi-user MIMO system. Two steps are followed to

solve the presented model: i) the departure distribution, $\boldsymbol{\pi}^d$, is computed by using the discrete-time embedded Markov chain method and ii) the PASTA (Poisson Arrivals See Time Averages) property of the Poisson arrivals is applied to find the probability distribution at arbitrary times, $\boldsymbol{\pi}^s$, as a function of $\boldsymbol{\pi}^d$.

3.1 Distribution of eligible frames in the queue (DEFQ)

As the traffic load is uniformly distributed among all destinations, the probability that an arriving frame is destined to a target MN is $1/N$. Therefore, to compute the probability of selecting $\varsigma \in [1, min(q, M)]$ frames from those q frames stored in the queue, it is assumed that, given a randomly chosen frame from the BS queue, it is destined to a target MN with probability $1/N$ too, that is, the same as its arrival probability.

Let $A_{q,\varsigma,N}$ be the event that computes, when q frames are stored in the queue, the possible choices of ς frames directed to different destinations, among a set of N possible destinations. The probability of the event $A_{q,\varsigma,N}$ can be computed as the quotient of favorable and total events. The number of total events is N^q, that is, the total number of different combinations of q elements if each element can take N different values. To compute the number of favorable events, we should consider, given the presence of q frames, those ς of them directed to different destinations. To do so, we first fix the ς destinations, that is, $\binom{N}{\varsigma}$. Then, for each one of these destinations, all possible ordered partitions of ς different elements in a queue with q frames have to be considered. That is $\sum_{(\mu_1,\ldots,\mu_\varsigma)\in\Psi_{q,\varsigma}} PR^q_{\mu_1,\ldots,\mu_\varsigma}$, where $\Psi_{q,\varsigma}$ is the set defined as $\Psi_{q,\varsigma} = \{(\mu_1,\ldots,\mu_\varsigma) \in \mathbb{Z}^\varsigma_+ \,|\, \mu_1 + \ldots + \mu_\varsigma = q\}$ and $PR^q_{\mu_1,\ldots,\mu_\varsigma}$ denotes the permutation with repetition of q elements in sets of $\mu_1,\ldots,\mu_\varsigma$ elements. Thus, it derives in the following analytic formula:

$$p(A_{q,\varsigma,N}) = \frac{\binom{N}{\varsigma}}{N^q} \cdot \sum_{(\mu_1,\ldots,\mu_\varsigma)\in\Psi_{q,\varsigma}} PR^q_{\mu_1,\ldots,\mu_\varsigma} \tag{1}$$

3.2 Distribution at departure epochs

The probability distribution at departure epochs, $\boldsymbol{\pi}^d$, is computed solving the linear system $\boldsymbol{\pi}^d = \boldsymbol{\pi}^d\mathbf{P}$, together with the normalization condition $\boldsymbol{\pi}^d\mathbf{1}^T = 1$. $\boldsymbol{P}$ is the probability transition matrix, where each $i, j \in [0, K]$ position, represents the probability $\overline{p}_{i,j}$ to move from any state i, with i the number of frames in the queue at the moment that a new space-batch is

scheduled, to any state j, with j the number of frames in the queue just after the space-batch departure. Each $\overline{p}_{i,j}$ is computed averaging all the different space-batch sizes and transmission rates that make that transition possible, that is

$$\overline{p}_{i,j} = \sum_{\varsigma=1}^{\max(1,\min(i,M))} p^{\bullet}(A_{i,\varsigma,N}) \sum_{\forall r \in \mathcal{R}} p(r,\varsigma) p_{i,j}(r,\varsigma) \tag{2}$$

where $p^{\bullet}(A_{i,\varsigma,N})$ is computed from

$$p^{\bullet}(A_{i,\varsigma,N}) = \begin{cases} p(A_{1,\varsigma,N}), & i = 0 \\ p(A_{i,\varsigma,N}), & i \geq 1 \end{cases} \tag{3}$$

as it takes into account the specific case in which the system is empty, $p(r,\varsigma)$ is the probability that a space-batch involving ς frames is transmitted at rate r, as defined in Section 2, and $p_{i,j}(r,\varsigma)$ is the transition probability from any state i to any state j given that a space-batch of ς frames is transmitted at rate r.

From any state $i \geq 1$ to any state $j \in [i-\varsigma, K-\varsigma-1]$, $p_{i,j}(r,\varsigma)$ is computed taking into account the probability of $v = (j-i)+\varsigma$ arrivals during the service time of the on-going space-batch, this is $p_{i,j}(r,\varsigma) = \frac{(\lambda\frac{L}{r})^v}{v!} e^{-\lambda\frac{L}{r}}$ where L/r is the space-batch service time and λ is the Poisson aggregate arrival rate. The transition probability from any state $i \geq 1$ to the last possible state in which the queue can depart, $j = K-\varsigma$, is computed as the complementary probability of not moving to any of the other states, this is $p_{i,K-\varsigma}(r,\varsigma) = 1 - \sum_{j=i-\varsigma}^{K-\varsigma-1} p_{i,j}(r,\varsigma)$. Finally, from the queue empty state, $i = 0$, to any state j, the transition probability $p_{0,j}(r,\varsigma)$ is given by the same probability of departing in state $q = 1$, $p_{0,j}(r,\varsigma) = p_{1,j}(r,\varsigma)$, as the system remains inactive while the queue is empty.

3.3 Distribution at arbitrary times

The distribution at arbitrary times, $\boldsymbol{\pi}^s$, is obtained using the PASTA property, which states that the probability to be at state u at any arbitrary time is equal to the probability that a new arrival finds the queue at this state. Note that this is equivalent to say that the system has been in the u-th state for $1/\lambda$ seconds on average in a given interdeparture epoch (Figure 2).

Additionally, from Figure 2, it can be observed also that the queue evolves between two macro-states, the non-transmitting and transmitting states, with average duration $E[T_t]$ and $E[T_{nt}]$ respectively. The steady-state probability that a new arrival finds the queue in one of these two macro-states depends on the proportion of time that the system is in each one, that

is $\phi_t = \frac{E[T_t]}{E[T_{nt}]+E[T_t]}$ for the transmitting macro-state and $\phi_{nt} = 1 - \phi_t$ for the non-transmitting one. $E[T_t]$ depends on the transmission rate at which a space-batch of size ς frames is transmitted. Then, averaging over all possible cases, $E[T_t] = \sum_{i=0}^{K-1} \pi_i^d \sum_{\varsigma=1}^{\max(1,\min(i,M))} p^{\bullet}(A_{i,\varsigma,N}) \sum_{\forall r \in \mathcal{R}} p(r,\varsigma) \frac{L}{r}$. $E[T_{nt}]$ is computed taking into account that the time between two packet arrivals is in average $1/\lambda$ and the system is in the non-transmitting state only if previous departure has finished in the empty state (otherwise the non-transmitting macro-state has a duration equal to zero). Then, $E[T_{nt}] = \frac{1}{\lambda}\pi_0^d$.

The probability that the system is in u-th state, $u \in [1, K-1]$, at any arbitrary time is shown in Equation 4. Note that, given a space-batch starting at the i-th state, the transition probabilities that guarantee that an arrival has observed the queue at the u-th state before the system departs at state j, depend on the size of the space-batch, the transmission rate r, and the relative position of the u-th state with respect to the i-th state. This is,

$$\pi_u^s = \phi_t \frac{1}{E[T_t]} \frac{1}{\lambda} \sum_{i=0}^{u} \pi_i^d \left(\sum_{\varsigma=1}^{\max(1,min(i,M))} p^{\bullet}(A_{i,\varsigma,N}) \sum_{\forall r \in \mathcal{R}} p(r,\varsigma) \sum_{j=u+1-\varsigma}^{K-\varsigma} p_{i,j}(r,\varsigma) \right) \tag{4}$$

The probability to be in the empty state at any arbitrary time is the probability that a new arrival finds the system in the non-transmitting state, this is $\pi_0^s = \phi_{nt}$. Finally, the probability to be at the K-th state at any arbitrary time is computed as $\pi_K^s = 1 - \sum_{k=0}^{K-1} \pi_k^s$, given that $\sum_{k=0}^{K} \pi_k^s = 1$.

4 Performance Results

Results from the queueing model are compared with the ones obtained by simulation. Simulations are done in C++, carefully considering the described scenario. The considered parameters are: $L = 1872$ bits (constant), $\mathcal{R} = \{r_1, r_2\}$, with $r_1 = 50$ Kbps and $r_2 = 150$ Kbps. Without loss of generality, here it is assumed that the probability that r_1 is the minimum rate when ς frames are scheduled is $p(r_1,\varsigma) = \left(\frac{\varsigma-1}{\Upsilon-1}\right)$, with $\Upsilon \geq M$ a scaling parameter. Then, the probability that a space-batch will be transmitted at r_2 is $p(r_2,\varsigma) = 1 - p(r_1,\varsigma)$.

In Figure 3 the blocking probability ($P_b = \pi_K^s$) is shown for $K = 25$ frames, two different number of antennas ($M = 4$ and $M = 8$) and two different number of MNs ($N = 8$ and $N = 80$) against the aggregate traffic load in bits/second. In Figure 4 the average delay (queueing plus transmis-

sion time, computed by applying Little's Law, $E[D] = \frac{\sum_{k=0}^{K} k \cdot \pi_k^s}{\lambda(1-P_b)}$) is shown for $M = 8$ antennas, two queue sizes, $K = 25$ and $K = 75$ frames, a variable number of users and a fixed traffic load equal to 350 Kbps. In both cases, Υ is set to 8 to penalize further the presence of $M = 8$ antennas (i.e. the chances to transmit at r_2 with $M = 4$ antennas are higher).

The model shows a very good accuracy for the considered traffic loads, number of antennas, number of active MNs and queue sizes. The points where the queueing model shows the worst accuracy are those in which the number of users and the number of antennas are similar. In such situation, the DEFQ approximation is optimistic as the predicted chances to schedule larger space-batches are higher than in the real system. Additionally, there are some other conclusions that can be observed in the results. They can be explained by the low number of eligible frames at each transmission, caused by both the consideration of non-saturated traffic sources and the presence of a finite buffer. First, increasing the number of antennas does not mean a proportional gain on the effective system capacity (see Figure 3); second, as the number of antennas grows, the performance gains are highly influenced by the number of active MNs (see Figure 3 and observe that with $M = 4$ antennas the same results are achieved for $N = 8$ and $N = 80$ MNs); and third, increasing the number of active MNs allows the system to perform more efficiently until a certain value where it stabilizes (see Figure 4). From this value on, increasing the number of MNs does not provide any substantial gain.

5 Conclusions

A new queueing model for Multi-user MIMO systems in non-saturated conditions has been presented. The assumptions done, together with the applied modeling methodology, result in a simple, but accurate, model specially indicated to be considered for the design of cross-layer MAC/PHY protocols for such systems.

References

[1] David Gesbert, Marios Kountouris, Robert W. Heath Jr., Chan-Byoung Chae, and Thomas Salzer. Shifting the MIMO paradigm. From single-user to multiuser communications. *IEEE Signal Processing Magazine. Vol. 24, Page(s). 26-46*, 2007.

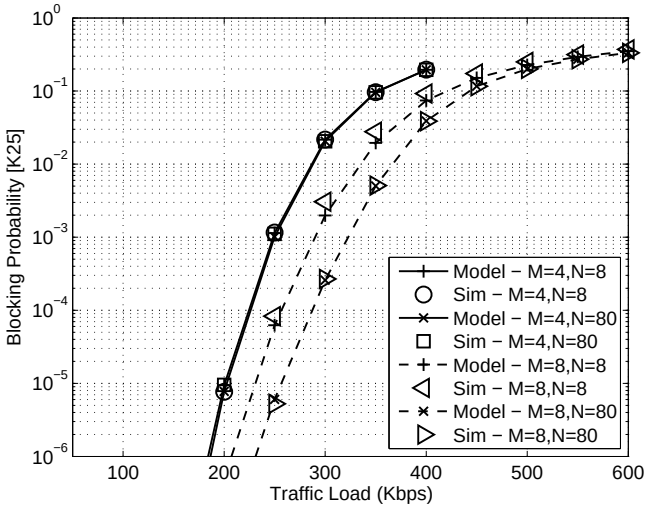

Figure 3: Blocking Probabiltiy

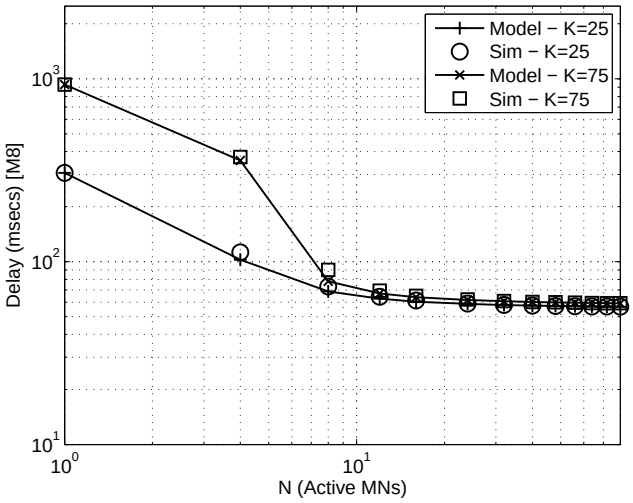

Figure 4: Delay

[2] Chiung-Jang Chen, and Li-Chun Wang; Performance Analysis of Scheduling in Multiuser MIMO Systems with Zero-Forcing Receivers. *IEEE Journal on Selected Areas in Communications, Vol.25, Number:7, Page(s): 1435-1445*, 2007.

[3] Donald Gross, and Carl M. Harris; Fundamentals of Queueing Systems, 3er Ed.. *John Wiley & Sons*, 1998.

[4] Boris Bellalta, and Miquel Oliver; A Space-Time Batch-service Queuing Model for Multi-user MIMO communication systems. In *In Proceedings of the 12-th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM), Tenerife, Spain*, 2009.