A Comparison of Beam Refinement Algorithms for Millimeter Wave Initial Access

Citation for the original published paper (version of record):

Guo, H., Makki, B., Svensson, T. (2017)
A Comparison of Beam Refinement Algorithms for Millimeter Wave Initial Access
28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, IEEE

A Comparison of Beam Refinement Algorithms for Millimeter Wave Initial Access

Hao Guo, Behrooz Makki, Tommy Svensson
Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden
{hao.guo, behrooz.makki, tommy.svensson}@chalmers.se

***Abstract*—Initial access (IA) is identified as a key challenge for the upcoming 5G mobile communication system operating at high carrier frequencies, and several techniques are currently being proposed. In this paper, we extend our previously proposed genetic algorithm (GA)-based beam refinement scheme to include beamforming at both the transmitter and the receiver, and compare the performance with alternative approaches in the millimeter wave multi-user multiple-input-multiple-output (MU-MIMO) networks. Taking the millimeter wave communications characteristics and various metrics into account, we investigate the effect of different parameters such as the number of transmit antennas/users/per-user receive antennas, beamforming resolution as well as hardware impairments on the system performance employing different beam refinement algorithms. As shown, our proposed GA-based approach performs well in delay-constrained networks with multi-antenna users. Compared to the considered state-of-the-art schemes, our method reaches the highest service outage-constrained end-to-end throughput with considerably less implementation complexity. Moreover, taking the users' mobility into account, GA-based approach can remarkably reduce the beam refinement delay at low/moderate speeds when the spatial correlation is taken into account.**

I. Introduction

The next generation of cellular systems (5G) requires both higher data rates (in the order of 10-100 Gbps) and lower end-to-end latencies (down to 1 ms) than previous generations [1]. For this reason, it is aimed to utilize frequency bands in the 30-300 GHz range in order to obtain sufficiently large bandwidths/data rates. Due to power limitation and high path loss at these frequencies, the coverage range is typically small so that highly directional transmissions are required for such millimeter wave (MMW) communications. On the other hand, the physical size of antennas at the MMW band is relatively small, such that large scale beamforming can be performed in practice [2] [3].

Employing large-scale beamforming during the initial access (IA) procedure can be a good way to overcome the increased path loss experienced at higher frequencies. The most challenging task of IA is that the base stations (BSs) make omnidirectional cell searches with directional beams and at the receiver side the users choose their best beam direction to detect the BSs. Successful access means, e.g., that the received power or the signal-to-noise ratio (SNR) is beyond certain thresholds. Connection being established, the BSs and the users begin exchanging message and beam refinement procedure need to be developed here to further improve the beam directions. The user mobility can also be handled by beam refinement.

IA beamforming at MMW is different from the conventional one since it is hard to acquire the channel state information (CSI) at these frequencies. For this reason, codebook-based beamforming has been recently proposed as an efficient method to reduce the dependency on CSI estimation/feedback (for detailed literature review, see Section II). Several works have been presented on both physical layer and procedural algorithms of codebook-based beamforming. However, in these works either the algorithms are designed for special metrics, precoding/combining schemes and channel models or the implementation complexity grows significantly by an increasing number of BSs/users. Moreover, the running delay of the algorithm has been rarely considered in the performance evaluation. On the other hand, generic machine learning-based schemes have been recently proposed for IA which can be effectively applied for different channel models with acceptable implementation complexity.

In this paper, we study the effect beam refinement on the performance of MMW networks. The contributions of the paper are three fold. 1) We extend our previously proposed genetic algorithm (GA)-based beamforming approach to include beamforming at both the transmitter and receiver side. Also, 2) we compare different machine learning based analog beamforming approaches for the beam refinement during IA, including GA-based beamforming, Tabu search beamforming [4], link-by-link beamforming [5] and two-level codebook beamforming [6] [7] in large-but-finite multi-user multiple-input-multiple-output (MU-MIMO) MMW communication systems. Moreover, 3) we analyze the effect of various parameters such as the number of transmit/receive

Published in Proc. of IEEE PIMRC Workshops 2017. Typographical adjustments have been made.

antennas, total power budget and the power amplifier (PA) efficiency on the network performance. As opposed to the literature we take the algorithm running delay into account. Thus, there is a trade-off between finding the optimal beamforming matrices and reducing the data transmission time slot, and the highest throughput may be achieved by few iterations, i.e., a rough estimation of the optimal beamformer. We study the system performance in terms of the end-to-end throughput with service outage constraints as well as the implementation complexity. 4) Furthermore, we evaluate and compare the performance of the considered algorithms under various mobile speed of the users.

Our results demonstrate that the running delay of the algorithms and power amplifier inefficiency affect the system performance remarkably, which should be carefully considered in the system design. Moreover, our proposed GA-based approach outperforms the considered state-of-the-art schemes, in terms of throughput, and reaches (almost) the same result as in the exhaustive search based approach with fewer number of iterations. Furthermore, when taking the user mobility into account, the GA-based approach can remarkably reduce the algorithm running delay based on the beamforming results in the previous time slots. Thus, the GA-based beamforming approach can be an appropriate candidate for IA in future wireless networks.

II. Literature Review

In this part, we present some related research work on IA. The reader mainly interested in technical details can skip this section and go to Sections III-V where we present the system model, the algorithm descriptions and the simulation results, respectively. Beamforming techniques at MMW bands have been considered in standard developments IEEE 802.15.3c (TG3c) [8], IEEE 802.11ad (TGad) [9] and ECMA-387 [10]. The problem formulation for IA beamforming at MMW frequencies are introduced in [11] where a fast-discovery hierarchical search method is proposed. Moreover, several design options for MMW IA are presented in [12], where the basic steps in the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) standard are used as references, and the overall delay of each design option as a function of the system overhead is evaluated. Then, [13] compares three approaches, namely, exhaustive search, two-step and context information-based, in terms of miss-detection probability and discovery time. Another comparison work is presented in [14], where it is shown that different IA protocols have a trade-off between delay and average user-perceived throughput.

In [15], we introduce a genetic algorithm-based initial beamforming approach and evaluate the effect of the algorithm running delay on the network performance. There are also previous works using the GA-based selection approach in different communication networks. For instance, in [16] an efficient scheduling scheme is designed based on the genetic algorithm in the return-link of a multi-beam satellite system. A turbo-like beamforming scheme based on the Tabu search algorithm is proposed in [4] to reduce both searching complexity and system overhead. A concurrent beamforming protocol, which we refer to as link-by-link beamforming, is presented in [5] to achieve high capacity in indoor MMW networks. Finally, for multi-stage beamforming, a tree-structured multi-level beamforming codebook is designed for MMW wireless backhaul systems in [6]. Also, in [7], a low-complexity multi-stage codebook is designed to support the IEEE 802.15.3c protocol.

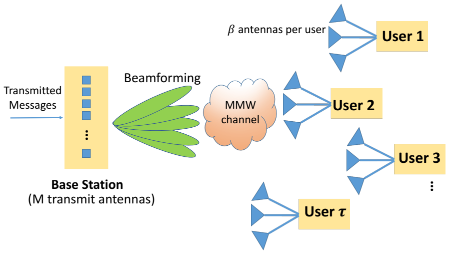


Fig. 1. MMW multiuser MIMO system model.

III. System Model

We consider a MU-MIMO setup with M transmit antennas at a BS and τ multi-antenna users, each with β antennas. As a result, there are $N = \tau \times \beta$ total antennas at the receiver side (see Fig. 1). This is an extension of our work [15] with single receive antennas, and allows for beamforming at the receiver side. We assume that each user has perfect receiver CSI. We set $M > N$. At each time slot t, the aggregated received signal vector $\mathbf{Y}(t)$ at time t over the users after receive beamforming can be described as

$$\mathbf{Y}(t) = \sqrt{\frac{P}{M}}\mathbf{U}(t)^H\mathbf{H}(t)\mathbf{V}(t)\mathbf{X}(t) + \mathbf{Z}(t), \qquad (1)$$

where P is the total power budget, $\mathbf{H}(t) \in \mathcal{C}^{N\times M}$ is the channel matrix with the (i,j)th element given by $H_{i,j}(t) = d_{i,j}^{\gamma}h_{i,j}(t)$, where $d_{i,j}$ is the distance between the receiver antenna i and the transmitter antenna j and γ is a path loss parameter, and $h_{i,j}(t)$ denotes the small scale fading. $\mathbf{X}(t) \in \mathcal{C}^{M\times 1}$ is the intended message signal, $\mathbf{V}(t) \in \mathcal{C}^{M\times M}$ is the precoding matrix at the BS, $\mathbf{U}(t) \in \mathcal{C}^{N\times N}$ is the aggregated combining matrix at the users' side, and $\mathbf{Z}(t) \in \mathcal{C}^{N\times 1}$ denotes the independent and identically distributed (IID) Gaussian noise matrix. For simplicity, we drop the time index t in the following.

Furthermore, the channel model $\mathbf{H}$ is described as

$$\mathbf{H} = \sqrt{\frac{k}{k+1}}\mathbf{H}_{\text{LOS}} + \sqrt{\frac{1}{k+1}}\mathbf{H}_{\text{NLOS}}, \qquad (2)$$

where $\mathbf{H}_{\text{LOS}}$ and $\mathbf{H}_{\text{NLOS}}$ denote the line-of-sight (LOS) and the non-line-of-sight (NLOS) components of the

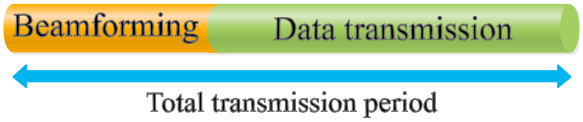


Fig. 2. Schematic of a packet transmission period.

channel, respectively, and the NLOS component is assumed to follow a complex Gaussian distribution. Also, k controls the relative strength of the LOS and the NLOS components. In (2), setting $k = 0$ represents an NLOS condition while $k \to \infty$ gives an LOS channel. We use this model because most cases in MMW systems have the LOS channel. Note that there are additional simulation results with different settings of k in [15] assuming single antenna users.

A. Initial Beamforming Procedure

Unlike a conventional beamforming procedure acquiring CSI, in MMW systems we suggest to perform codebook-based beamforming, which means selecting a precoding matrix $\mathbf{V}$ out of a predefined codebook $\mathbf{W}_\text{T}$ at the BS while selecting a combining matrix $\mathbf{U}$ out of a predefined codebook $\mathbf{W}_\text{R}$ at the receiver side, sending test signal and finally making decisions on transmit/receive beam patterns based on the users' feedback about their performance metrics. The IA will be finished as soon as a stable control link is established. The time structure for a packet transmission can be seen in Fig. 2, where part of the packet period is dedicated to design appropriate beams and the rest is used for data transmission. Thus, we need to find a balance between the beam design delay and the data transmission period by choosing an efficient approach.

Here, we use discrete fourier transform (DFT)-based codebooks [17] at both sides which are defined as

$$\mathbf{W}_\text{T} = \{w(m,u)\} = \{e^{-j2\pi(m-1)(u-1)/N_\text{vec}}\},\\ m = 1,2,\dots,M, u = 1,2,\dots,N_\text{vec}, \tag{3}$$

for the BS, while

$$\mathbf{W}_\text{R} = \{w(n,u)\} = \{e^{-j2\pi(n-1)(u-1)/N_\text{vec}}\},\\ n = 1,2,\dots,N, u = 1,2,\dots,N_\text{vec}, \tag{4}$$

for the users, where $N_\text{vec} \geq \max(M,N)$ is the number of codebook vectors.

B. Performance Metrics

The machine learning based schemes of [4]-[7] and [15] are generic, in the sense that they can be implemented for different metrics. For the simulations, however, we consider the service outage-constrained end-to-end throughput, the complexity and the average number of required iterations as the system performance metric. In some scenarios, it may be required to serve the users with some minimum required rates, otherwise *service outage* occurs. In the K-th iteration round of the algorithm, the service outage-constrained end-to-end throughput in bit-per-channel-use (bpcu) is defined as

$$\begin{aligned} &R(K) = (1-\alpha K)\sum_{i=1}^{\tau} r_i^K U(r_i^K, \log_2(1+\theta)),\\ &r_i^K = \log_2\left(1+\textbf{SINR}_i^K\right),\\ &U(r_i^K, \log_2(1+\theta)) = \begin{cases} 1 & r_i^K \geq \log_2(1+\theta)\\ 0 & r_i^K < \log_2(1+\theta). \end{cases} \end{aligned} \tag{5}$$

Here, r_i^k denotes the achievable rate of the user i at the end of the K-th iteration. Also, parameter α is the relative delay cost for running each iteration of the algorithm which fulfills $\alpha N_\text{it} < 1$ with N_it being the maximum possible number of iterations. Then, $\log_2(1+\theta)$ is the minimum per-user rate while θ represents the minimum required signal-to-interference-plus-noise ratio (SINR) of each user. Also,

$$\textbf{SINR}_i^K = \frac{\frac{P}{M} g_{i,i}^K}{BN_0 + \frac{P}{M}\sum_{i\neq j}^N g_{i,j}^K} \tag{6}$$

is the SINR at the receiver of user i in the iteration round K. Here, $g_{i,j}$ is the (i,j)-th element of the matrix $\mathbf{G}_K = |\mathbf{U}_K^H\mathbf{H}\mathbf{V}_K|^2$ which is referred to as the channel gain throughout the paper. Moreover, B is the system bandwidth and N_0 is the power spectral density of the noise. We set $BN_0 = 1$ to simplify the system so that the power P (in dB, $10\log_{10} P$) denotes the receiver side SNR as well.

The optimization problem of (5) is formulated as

$$\begin{aligned} \max_{K,\mathbf{U},\mathbf{V}} \quad & R(K)\\ \text{s.t.} \quad & \forall K \in \{1,2,3,...,N_\text{it}\}\\ & \forall \mathbf{V} \subseteq \mathbf{W}_\text{T}\\ & \forall \mathbf{U} \subseteq \mathbf{W}_\text{R}. \end{aligned} \tag{7}$$

As opposed to, e.g., [5, Eq. 3], [12, Eq. 1], [18, Eq. 43], [19, Eq. 3], [20, Eq. 5]and [21, Eq. 5], we consider the algorithm running delay in the performance analysis. As seen in the following, there is a trade-off between optimizing beamforming matrices and reducing the data transmission period. In this case the optimal solution may be achieved by running the algorithms for a limited number of iterations.

C. On the Effect of Power Amplifier Efficiency

The efficiency of the radio-frequency high Power Amplifier (PA) should be taken into consideration in the multi-antenna systems. Here, we consider the state-of-the-art PA efficiency model [22, Eq. 13], [23, Eq. 3]

$$\rho_\text{cons} = \frac{\rho_\text{max}^\mu}{\epsilon \times \rho_\text{out}^{\mu-1}} \tag{8}$$

where ρ_cons, ρ_out, ρ_max refer to as the consumed power, the output power and the maximum output power of the

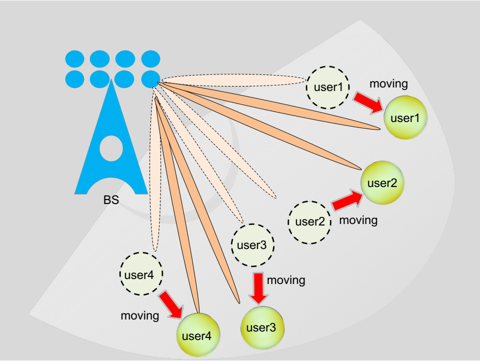


Fig. 3. Mobility model.

PA, respectively. Also, $\epsilon \in [0, 1]$ represents the power efficiency and $\mu \in [0, 1]$ is a parameter depending on the PA class. Setting $\epsilon = 1$, $P_{\text{max}} = \infty$ and $\mu = 0$ in (8) represents the special case (with an ideal PA).

D. On the Effect of User Mobility

Beamforming solutions for mobile users at high carrier frequencies are important in 5G wireless mobile communications. Here, we use the following mobility model to evaluate the performance of our proposed GA-based beamforming approach and compare the results with those of the considered state-of-the-art schemes. Consider Fig. 3 with $\tau = 4$ multiple-antenna users. Here, we have two cases during the users' mobility:

- **Case 1** Beam refinement with a random queen as initial guess (dash-line circles in Fig. 3)
- **Case 2** Beam refinement with using the queen in Case1 as initial guess (full-line circles in Fig. 3)

By mobility we exploit the spatial correlation by setting the queen of the previous time slot as one of the initial guesses of the next time slot. We assume that we know the moving speed v_m, the time duration of mobility t_m so that we can get an estimate of the user position in a circle whose radius is found by $v_m \times t_m$ in Case 2.

IV. Algorithm Description

In this study, we compare the performance of different IA beamforming methods as follows.

Extended GA-based Search [15]: The algorithm starts by making L possible beam selection sets at both transmitter and receiver, i.e., submatrices of each codebook. During each iteration, we choose the best set, named as the *Queen*, based on the performance metrics (for example, (5)). Next, we keep the Queen and regenerate $S < L$ similar sets around the Queen by making small changes to the Queen (in the simulations, we replace 10% of the Queen columns randomly). Finally, the other $L - S - 1$ beamforming matrices are selected randomly to avoid the algorithm from being trapped in a local minima. After N_{it} iterations (set by designer), the Queen is returned as the beam selection result in the current time slot. In this way, this is an extended version of our GA-based approach with beamforming at both the transmitter and the receiver, the basic principles of which can be found in [15].

Tabu Search [4]: The Tabu-search approach follows the basic idea as in the GA-based scheme [4] where we choose and update the Queen by iterations. The only difference is the evolution method of the Queen in successive iterations. With Tabu, we use the definition of *neighborhood* in [4]: One matrix $\mathbf{A}$ is defined as another matrix $\mathbf{B}$'s neighborhood if 1) $\mathbf{A}$ has only one different column compared with $\mathbf{B}$ or 2) the index difference between the two corresponding columns in $\mathbf{A}$ and $\mathbf{B}$ is equal to one. To make S beam selection sets, we change the queen from previous round to its neighbors.

Link-by-link Search [5]: In this strategy, the beam design of τ users is not optimized simultaneously. Instead, with a greedy approach, the beamforming solution is settled user-by-user by considering the interference from the other $\tau - 1$ links. The system performance improves in successive iterations until it converges to some (sub)optimal beamforming rules.

Two-level Search [6][7]: Being inspired by multi-stage beamforming techniques, e.g., [6] and [7], we design a two-level-codebook search scheme for our system. In the first level, the BS transmits messages over wider sectors using the codebook with $N_{\text{vec}}/2$ columns, while in the second level it searches the optimal solution within the best such sector by steering narrower beams with an N_{vec}-column codebook.

A. On the Implementation Complexity

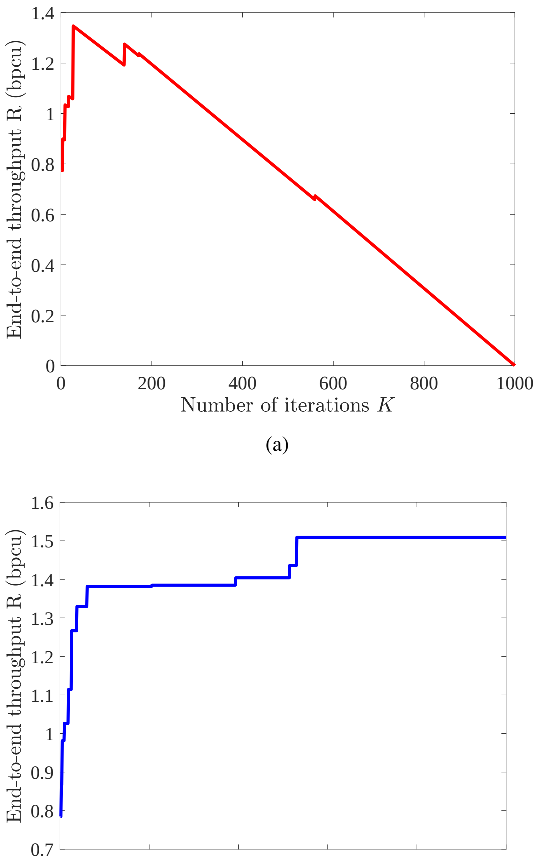
To compare different methods, it is necessary that we consider the implementation complexity of each algorithm. For this reason, we derive the per-iteration complexity of different algorithms based on the fact that the product of matrices of size $N \times M$ and $M \times M$ has the complexity $\mathcal{O}(NM^2)$ in MATLAB. In this way, the per-iteration complexity for the GA-based approach is given by

$$C_{\text{GA}} = L(2\mathcal{O}(N^2M) + \mathcal{O}(NM^2) + \mathcal{O}(NM)), \quad (9)$$

and $C_{\text{Tabu}} = C_{\text{GA}}$, $C_{\text{link-by-link}} = \tau \times C_{\text{GA}}$, $C_{\text{two-level}} = 2 \times C_{\text{GA}}$. L is the number of beam selection sets within each iteration.

V. Simulation Results

In the simulations, we use the channel model in (2) in the cases with $k = 0, 3$. We set $\mathbf{H}_{\text{LOS}} = \mathbf{1}_{N\times M}$ where $\mathbf{1}_{a\times b}$ refers to normalized all-ones complex matrix. Except for Fig. 4 which shows an example of the GA-based procedure, for each point in the curves the results are obtained by averaging over 10^4 different channel realizations. In all figures, we set $N_{\text{it}} = 1000$ since it is a sufficiently large number of iterations after which no performance improvement is observed. Also, in all



(a)

(b)

Fig. 4. Examples of the convergence process of the extended GA-based beamforming for systems with (subplot a) and without (subplot b) delay costs of the algorithm. $M = 32$, $\tau = 4$, $N = 12$, $P = -10$ dB, $k = 0$.

figures except for fig. 9, we use the normalized distance $d_{i,j} = d = 1$. Moreover, we set $L = 10$, $S = 5$ and $N_{\text{vec}} = 128$. In all figures, except for Fig. 7, we use the ideal PA, i.e., set $P_{\max} = \infty, \mu = 0, \epsilon = 1$ in (8). In Fig. 7 we study the effect of imperfect PAs. In Figs. 4-8, we consider the service outage-constrained end-to-end throughput (5) as the performance metric with $\theta = -4$ dB. Finally, Table I shows the average number of required iterations in each algorithm to reach the (sub)optimal solution.

On the convergence behavior: Figures 4a and 4b give examples of the GA performance in the cases with ($\alpha = 0.001$) and without costs of running the algorithm ($\alpha = 0$), respectively (see (5)). From Fig. 4a we observe that very few iterations are required to reach the maximum throughput. That is, considering the cost of running the algorithm, the maximum throughput is obtained by finding a suboptimal beamforming matrix and leaving the rest of the time slot for data transmission (see Fig. 2). On the other hand, as the number of iterations increases, the cost of running the algorithm reduces the end-to-end throughput converging to zero at $K = \frac{1}{\alpha}$ (see (5)).

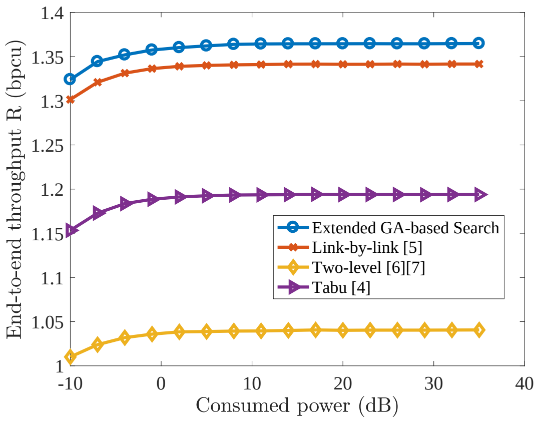


Fig. 5. Service outage-constrained end-to-end throughput of different methods. $M = 32$, $\tau = 4$, $N = 12$, $k = 0$, $\alpha = 0$.

If there is no running delay, on the other hand, the system performance improves with the number of iterations monotonically (Fig. 4b). However, the developed algorithm leads to (almost) the same performance as the exhaustive search-based scheme with very limited number of iterations. For example, with the parameter settings of Fig. 4b, our algorithm reaches more than 90% of the maximum achievable throughput with less than 100 iterations. Note that with the parameter settings of Fig. 4, exhaustive search implies testing in the order of 10^{30} possible beamforming matrices.

Finally, all considered schemes follow the same ladder-type convergence behavior as in Fig. 4. This is because with the considered algorithms the system performance is not necessarily improved in each iteration and may be trapped into local minima. However, considering a couple of random solution checks in each iteration helps to avoid the local minima as the number of iterations increases.

Comparison of different schemes: In Fig. 5, we compare the throughput (5) reached by different considered algorithms. It can be seen from the figure that for a broad range of SNRs the GA-based beamforming [15] leads to the best system throughput, followed by the link-by-link search [5], Tabu search [4] and two-level search [6], [7].

Moreover, using the same parameter settings of Fig. 5, in Fig. 6 we compare the cumulative distribution function (CDF) of per-user throughput (5) reached by different considered algorithms. From the figure we can see that the GA-based beamforming [15] leads to the best per-user throughput distribution, which means more user can be served by higher throughput, followed by the link-by-link search [5], Tabu search [4] and two-level search [6], [7].

Table I shows the average number of iterations $\bar{N}$ that are required in each scheme to reach a (sub)optimal solution. Here, the results are presented for $k = 0, M = 32, N = 4, 8, 12$. We can see that in all methods, except

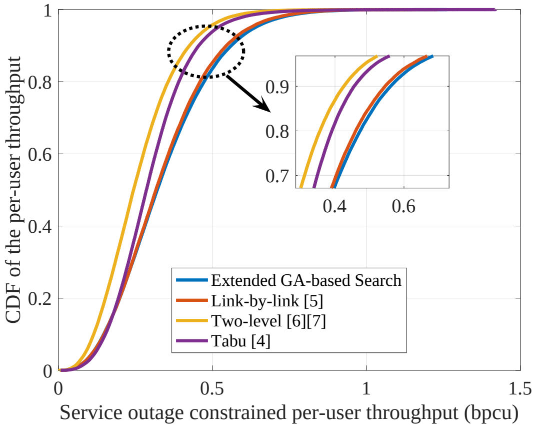


Fig. 6. CDF of per-user throughput with different methods. $M = 32$, $\tau = 4$, $N = 12$, $k = 0$, $\alpha = 0$.

TABLE I
AVERAGE NUMBER OF REQUIRED ITERATIONS $\bar{N}$ IN DIFFERENT SITUATIONS

M/N	GA	Tabu	link-by-link	two-level
32/12	502	498	307	501
32/8	500	501	288	498
32/4	488	502	261	500

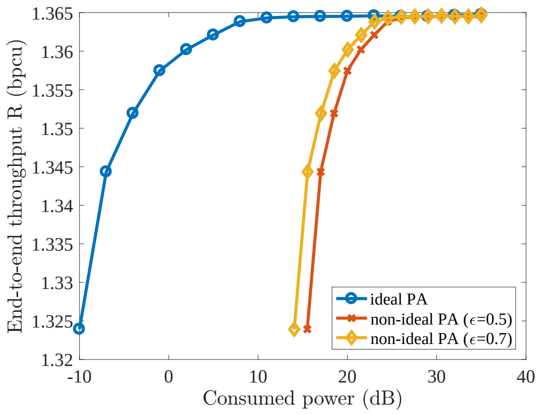


Fig. 7. The effect of power budget and PAs efficiency on the end-to-end throughput (5). $M = 32$, $\tau = 4$, $N = 12$, $k = 0$, $\alpha = 0$.

for the link-by-link approach, the required number of iterations is almost insensitive to the number of receive antennas.

On the effect of imperfect power amplifier: Figure 7 evaluates the effect of the power amplifier on the throughput (5). We can see that the inefficiency of the PA affects the performance remarkably but this effect decreases with the SNR. This is reasonable because the effective efficiency of the PAs $\epsilon^{\text{effective}} = \epsilon(\frac{p_{\text{out}}}{p_{\max}})^{\mu}$ increases with SNR.

On the effect of the number of receive antennas: Figure 8 shows the effect of number of receive antennas per user β on the throughput (5). As seen in the figure, the end-to-end throughput increases with the number of per-user antennas as expected, since multi-antenna techniques can improve the data rate remarkably. Moreover, the relative performance gain of the GA-based scheme, compared to other considered schemes, increases with the number of receive antennas, which is an interesting point when designing large-scale networks.

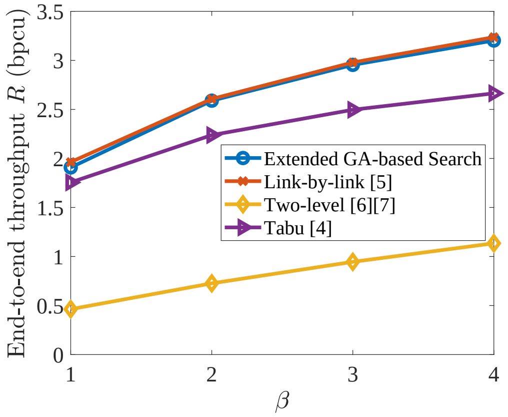


Fig. 8. Throughput (5) with different number of receive antennas at the user side β. $M = 32$, $\tau = 4$, $\beta = 1, 2, 3, 4$, $k = 3$, $\alpha = 0$, $P = 2$ dB.

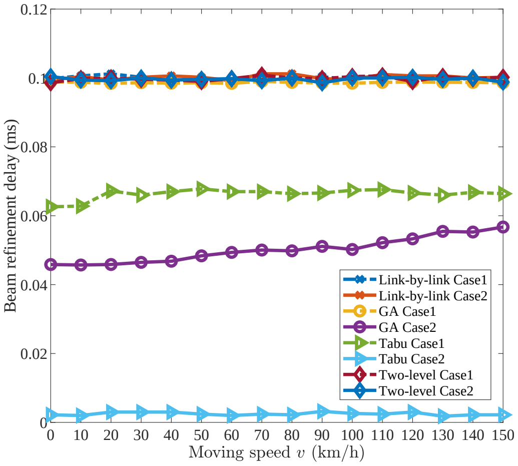


Fig. 9. Beam refinement delay with different moving speed v_m. $M = 32$, $\tau = 4$, $\beta = 2$, $k = 0$, $\alpha = 0$, $P = 32$ dB, moving time $t_m = 1$ ms, $\gamma = -3.5$.

On the effect of the user mobility: Figure 9 shows the effect of the users' mobility on the beam refinement delay for the considered algorithms. Inspired by [13], we evaluate the beam refinement delay (we assume that each iteration takes 10^{-4} overhead of t_m) of each algorithm in Case 1 and Case 2 to check how well these algorithms are suitable for the mobile users. The algorithm running delays in Case 1 and Case 2 of each method are all presented in the plot. Here, the results are presented with $M = 32$, $\tau = 4$, $\beta = 2$, $k = 0$, $\alpha = 0$, $P = 32$ dB, moving time $t_m = 1$ ms, $\gamma = -3.5$. As seen in the figure, both the GA-based algorithm and the Tabu-based algorithm can remarkably reduce the beam refinement delay for a broad range of users speeds, since they can use the beam refinement solution in Case 1 as the initial guess in Case 2 when the moving distance is not large. However, as the users speed increases the beam refinement delay increases slightly, intuitively because the spatial correlation between the positions in successive time slots decreases. Moreover, both the link-by-link search and the two-level-based search do not show noticeable performance gain.

VI. Conclusion

We extended our previously proposed genetic algorithm (GA)-based beam refinement scheme to include beamforming at both the transmitter and the receiver, and we compared the performance with alternative beam refinement algorithms in an MU-MIMO system, in terms of the service outage-constrained end-to-end throughput and the implementation complexity. Particularly, our extended genetic algorithm-based scheme can reach almost the same throughput as in the exhaustive search-based approach with relatively few iterations in delay-constrained systems. Also, compared to the considered state-of-the-art schemes, our scheme leads to the highest throughput/per-user throughput and the lowest per-iteration implementation complexity, and the relative performance gain increases with the number of receive antennas. Moreover, non-ideal power amplifiers affect the system performance remarkably, which should be carefully considered during the system design. Finally, the GA-based approach can exploit the spatial correlation and remarkably reduce the beam refinement delay for a broad range of users speeds, which means it is an appropriate approach for mobile users. As future work, we will investigate our proposed algorithm with more realistic parameter settings/scenarios, and compare the result with other structured beamforming methods.

Acknowledgment

The research leading to these results received funding from the European Commission H2020 programme under grant agreement $n°671650$ (5G PPP mmMAGIC project), and from the Swedish Governmental Agency for Innovation Systems (VINNOVA) within the VINN Excellence Center Chase.

References

[1] M. Cudak, A. Ghosh, T. Kovarik, R. Ratasuk, T. A. Thomas, F. W. Vook, and P. Moorut, "Moving towards mmwave-based beyond-4G (B-4G) technology," in *Proc. IEEE VTC'2013, Dresden, Germany*, Jun. 2013, pp. 1–5.

[2] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[3] S. Sun, G. R. MacCartney, M. K. Samimi, S. Nie, and T. S. Rappaport, "Millimeter wave multi-beam antenna combining for 5G cellular link improvement in new york city," in *Proc. IEEE ICC'2014, Sydney, Australia*, Jun. 2014, pp. 5468–5473.

[4] X. Gao, L. Dai, C. Yuen, and Z. Wang, "Turbo-like beamforming based on tabu search algorithm for millimeter-wave massive mimo systems," *IEEE Trans. Veh. Technol*, vol. 65, no. 7, pp. 5731–5737, Jul. 2016.

[5] J. Qiao, X. Shen, J. W. Mark, and Y. He, "MAC-layer concurrent beamforming protocol for indoor millimeter-wave networks," *IEEE Trans. Veh. Technol*, vol. 64, no. 1, pp. 327–338, Jan. 2015.

[6] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Multilevel millimeter wave beamforming for wireless backhaul," in *Proc. IEEE GLOBECOM'2011, Houston, Texas, USA*, Dec. 2011, pp. 253–257.

[7] L. Chen, Y. Yang, X. Chen, and W. Wang, "Multi-stage beamforming codebook for 60GHz wpan," in *Proc IEEE ICST'2011, Harbin, China*, Aug. 2011, pp. 361–365.

[8] J. P. Gilb, "IEEE standards 802.15. 3c–part 15.3: wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPANs) amendment 2: millimeter-wave-based alternative physical layer extension [s]," *IEEE Computer Society, New York*, Aug. 2009.

[9] C. Cordeiro *et al.*, "IEEE P802. 11 Wireless LANs, PHY/MAC Complete Proposal Specification (IEEE 802.11-10/0433r2)," May. 2010.

[10] H. Rate, "GHz PHY, MAC and PALs, Standard ECMA-387, ser. Available at: https://www.ecma-international.org/publications/files/ECMA-ST/ECMA-387.pdf," 2010.

[11] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmwave communications," in *Proc. IEEE ACSSC'2014, Pacific Grove, CA, USA*, Nov. 2014, pp. 1926–1930.

[12] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, "Initial access in millimeter wave cellular systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 7926–7940, Dec. 2016.

[13] M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5G mmwave cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 40–47, Nov. 2016.

[14] Y. Li, J. G. Andrews, F. Baccelli, T. D. Novlan, and J. Zhang, "On the initial access design in millimeter wave cellular networks," in *Proc. IEEE GLOBECOM'2016, Washington, USA*, Dec. 2016, pp. 1–6.

[15] H. Guo, B. Makki, and T. Svensson, "A genetic algorithm-based beamforming approach for delay-constrained networks," in *2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2017, pp. 1–7.

[16] B. Makki, T. Svensson, G. Cocco, T. de Cola, and S. Erl, "On the throughput of the return-link multi-beam satellite systems using genetic algorithm-based schedulers," in *Proc. IEEE ICC'2015, London, UK*, Jun. 2015, pp. 838–843.

[17] L. Wan, X. Zhong, Y. Zheng, and S. Mei, "Adaptive codebook for limited feedback MIMO system," in *Proc. IEEE WOCN'2009, Cairo, Egypt*, Apr. 2009, pp. 1–5.

[18] J. Choi, "Beam selection in mm-wave multiuser MIMO systems using compressive sensing," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2936–2947, Jun. 2015.

[19] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, March. 2014.

[20] B. Li, Z. Zhou, W. Zou, X. Sun, and G. Du, "On the efficient beam-forming training for 60GHz wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 504–515, Feb. 2013.

[21] H.-H. Lee and Y.-C. Ko, "Low complexity codebook-based beamforming for MIMO-OFDM systems in millimeter-wave WPAN," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3607–3612, Nov. 2011.

[22] B. Makki, T. Svensson, T. Eriksson, and M.-S. Alouini, "On the required number of antennas in a point-to-point large-but-finite MIMO system: Outage-limited scenario," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1968–1983, May. 2016.

[23] D. Persson, T. Eriksson, and E. G. Larsson, "Amplifier-aware multiple-input single-output capacity," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 913–919, Jan. 2014.