(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 May 2018 (03.05.2018)

W!P O ¦ PCT

(10) International Publication Number
WO 2018/078465 Al

(51) International Patent Classification:
H04B 7/04 (20 17.0 1) *H04B 7/0413* (2017.01)
H04L 25/03 (2006.01)

(21) International Application Number:
PCT/IB2017/055913

(22) International Filing Date:
27 September 2017 (27.09.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/412,561 25 October 2016 (25.10.2016) US

(71) Applicant: **KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY** [SA/SA]; 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA).

(72) Inventors: **LTAIEF, Hatem;** 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA). **REZKL Zouheir;** 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA). **ARFAOUI, Mohamed Amine;** 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA). **ALOUINI, Mohamed-Slim;** 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA). **KEYES, David E.;** 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available):* AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available):* ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.1 7(H))*
— *of inventorship (Rule 4.1 7(iv))*

Published:
— *with international search report (Art. 21(3))*

WO 2018/078465 A1

(54) **Title:** EFFICIENT SPHERE DETECTOR ALGORITHM FOR LARGE ANTENNA COMMUNICATION SYSTEMS USING GRAPHIC PROCESSOR UNIT (GPU) HARDWARE ACCELERATORS

(57) **Abstract:** A method of detecting a symbol transmitted over a communication channel in a multiple input-multiple output communication system. The method includes receiving a plurality of symbols transmitted over a communication channel of a multiple input-multiple output communication system. A sphere radius is imtialized based on attributes of the communication channel. A first matrix of possible transmitted signals is defined as well as a second matrix of received symbols. The matrix of possible transmitted signals is searched using a breadth-first search (BFS). Each level of the search tree is analyzed utilizing matrix multiplication to determine selected symbols satisfying the initialized sphere radius. A maximum likelihood solution is of the transmitted symbols is derived based on the selected symbols.

EFFICIENT SPHERE DETECTOR ALGORITHM FOR LARGE ANTENNA COMMUNICATION SYSTEMS USING GRAPHIC PROCESSOR UNIT (GPU) HARDWARE ACCELERATORS

TECHNICAL FIELD

[0001] The present disclosure is related generally to multiple-input, multiple output (MIMO) systems, and in particle to efficient decoding of received signal in MIMO systems.

BACKGROUND

[0002] A typical wireless communication system utilizes a transmitter unit coupled to a receiver unit in a single-input, single-output (SISO) configuration. More recently, multiple-input, multiple-output (MIMO) systems that significantly increases the data transmit capacities of wireless communication systems have been developed. In particular, MIMO is based on the principle that multipath propagation (i.e., spatial diversity) can be exploited to transmit multiple, independent information streams using co-located antennas and multi-dimensional signal processing. The transmitter includes plurality (multiple) transmit antennas that are utilized to send multiple streams of data. The receiver includes plurality (multiple) receiver antennas configured to receive the streams of data. The transmit streams take multiple paths between the transmit antennas and the receiver antennas. That is, the transmit streams can be described as going through a matrix channel that consists of all possible paths between the plurality of transmit antennas and the plurality of receiver antennas. The plurality of receive antennas generates a received signal vector that is decoded into an approximation of the original information.

[0003] A plurality of decoding schemes may be utilized to decode the original information or symbols from the received signal vector. Linear decoders are the least complex and therefore the easiest to implement while maintaining a desired latency (e.g., 10 milliseconds (ms) or less). However, linear decoders provide suboptimal error performance. In contrast with linear decoders, non-linear decoders such as the maximum

likelihood decoder (MLD), provide the best (i.e., lowest) error performance. The MLD essentially evaluates many hypotheses for all possible sequences of symbols that might have been transmitted based on all the information available at the receiver, and selects the best fit from all possible sequences. However, the complexity of MLD increases with the number of transmit antennas and increases exponentially based on the constellation of symbols, thus making MLD unworkable for most applications.

[0004] Sphere decoding is a type of non-linear decoder that provides a trade-off between the high complexity of MLD and the poor performance of linear detectors. In particular, sphere decoding searches for possible solutions to the transmitted vector within a sphere having a defined radius. The difficulty in sphere decoding is selecting the radius of the sphere. If the radius selected is too small, then very few possible solutions will be considered and the performance of the decoder might be poor, especially at low signal-to-noise ratio (SNR). Conversely, if the radius of the sphere selected is too large, then the complexity of the sphere decoder becomes essentially the same as the MLD decoder.

[0005] It would therefore be desirable to provide a method of tuning the sphere radius to optimize the performance of the sphere decoder (i.e., acceptable latency, low bit-error rate (BER)).

SUMMARY

[0006] According to some embodiments, a method of decoding a signal comprised of a plurality of symbols communicated via a multiple input-multiple output (MIMO) communication system includes receiving a signal at a receiver end of the communication channel. At the receiver side, the obtained baseband signal is a function of the transmitted signal altered byt eh MIMO channel and an additve Gaussian noise. A sphere radius is initialized based on attributes of the communication channel. A set of possible transmit signal symbols is defined based on the initialized sphere radius *r*. A maximum likelihood

solution of the transmitted signal is selected based on the selected set of possible transmit signal symbols.

[0007] According to another embodiment, a decoder circuit for decoding a multiple-input, multiple output (MIMO) signal comprises one or more processors configured to initialize a sphere radius r based on attributes of the communication channel. The one or more processors are further configured to define a first matrix Vk having size $(k, M_c Gk\text{-}i)$, wherein k is search tree level being evaluated, M_c defines a number of symbols in a constellation set, and $Gk\text{-}i$ defines a number of vectors at level $k\text{-}1$ determined to be located within the sphere radius r, wherein the first matrix Vk is comprised of vectors $sk\text{-}i$ located in the set $Lk\text{-}i$ representing the transmitted signal vectors that satisfy the sphere radius r. The one or more processors are further configured to define a second matrix yk based on the received signal y_k duplicated $M_c Gk\text{-}i$ times. The one or more processors are further configured to utilize matrix-matrix multiplication to calculate a third matrix Pk based on the equation $Pk = yk - RkVk$, wherein Rk is related to the QR decomposition of channel estimation H, wherein the third matrix Pk is evaluated with respect to the sphere radius r to derive the set of vectors Lk satisfying the sphere radius r for $k = 1 \ldots M$, wherein M is the number of transmitter. The one or more processors are further configured to select a maximum likelihood solution of the transmitted signal based on the matrix PM, wherein PM is the calculated third matrix Pk when $k = M$.

[0008] According to another embodiment, a method of decoding a signal comprised of a plurality of symbols communicated via a multiple input-multiple output (MIMO) communication system comprises defining a first matrix of possible transmit signal symbols and a second matrix of received symbols based on a breadth-first search (BFS) of search tree levels defined by potential transmitted symbols. Search tree levels are traversed utilizing matrix multiplication to determine the selected symbols satisfying the initialized sphere radius, and a maximum likelihood solution of the transmitted symbols is derived based on the selected symbols.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] Figure 1 is a block diagram of a massive multiple-input multiple-output (MIMO) system according to an embodiment of the present invention.

[0010] Figure 2 is a diagram illustrating a sphere decoder implemented as a breadth-first tree search as known in the prior art.

[0011] Figure 3 is a flowchart that illustrates operations performed by the sphere decoder according to an embodiment of the present invention.

[0012] Figure 4a is a line chart illustrating bit error rate (BER) of the sphere decoder and the linear receivers for 50 x 25 MIMO system and binary phase-shift keying (BPSK) modulation.

[0013] Figure 4b is a line chart illustrating execution time of the sphere decoder and the linear receivers for 50 x 25 MIMO system and binary phase-shift keying (BPSK) modulation.

[0014] Figure 5a is a line chart illustrating bit error rate (BER) of the sphere decoder for 50 x 25 MIMO system utilizing various constellations.

[0015] Figure 5b is a line chart illustrating execution time of the sphere for 50 x 25 MIMO system utilizing various constellations.

[0016] Figure 6a is a line chart illustrating bit error rate (BER) of the sphere decoder for 4-QAM modulation and for 50 x 25, 30 x 10, and 40 x 10 MIMO systems.

[0017] Figure 6b is a line chart illustrating execution time of the sphere for 4-QAM modulation and for 50 x 25, 30 x 10, and 40 x 10 MIMO systems.

DETAILED DESCRIPTION

[0018] The present invention provides a system and method of decoding massive multiple-input, multiple output (MIMO) wireless transmissions. In particular, the present invention utilizes a sphere decoding scheme wherein the radius of the sphere is fine-tuned to reduce the pool of possible candidates. In addition, one embodiment of the present invention utilizes a breadth-first tree traversal - as opposed to a depth first tree traversal. A benefit of this approach is that breadth-first tree traversal utilizes matrix-matrix

multiplication operations that may be processed in a graphics processing unit (GPU) environment that is compute-bound (as opposed to memory-bound). This is in contrast with depth-first traversal which requires vector-matrix multiplication and as a result requires memory-bound computations. By utilizing a GPU processing environment in combination with the parallelism gained via a breadth-first search tree, typically memory-bound computations are cast into compute-bound operations, the overall complexity of the MIMO decoder is reduced while maintaining real-time or near real-time processing.

[0019] Figure 1 is a block diagram of massive multiple-input multiple-output (MIMO) system 100 according to an embodiment of the present invention. Massive MIMO system 100 includes modulator 102, transmitter 104, channel 106, receiver 110, and demodulator 112.

[0020] In general, MIMO refers to a communication system using multiple transmit and receive antennas to exploit multi-path propagation. In the embodiment shown in Figure 1, a binary stream of data is being communicated. Modulator 102 modulates the stream of binary data utilizing one of a plurality of modulation techniques, including but not limited to binary phase shift keying (BPSK), quadrature amplitude modulation (QAM), which in turn may utilize one of QAM methods such as 4-QAM, 16- QAM or 64-QAM. The modulated data stream is provided to transmitter 104, which is comprised of a plurality of transmitters (e.g., large antenna array).

[0021] Transmitter 104 transmits the modulated stream of data via channel 106 to receiver 110, which is a comprised of a plurality of antennas (e.g., large antenna array). Channel 106 describes the propagation paths the signal takes and may be known or unknown. In one embodiment, the large transmitter array include M transmitters, while the large receiver array include N receivers. For an M x N MIMO configuration, the transmitter sends different signal streams on M antennas and the receiver receives N different signal streams, one per receiver antenna. With respect to channel 106, in some embodiments information may be known regarding the channel at the receiver side and how the channel interacts with the transmitted signal. Known channel properties are described as channel state information (CSI). In addition to channel properties, at least some noise is added to

the transmitted signal. Receiver 110 decodes the received signal and provides the decoded signal to demodulator 112. Decoding of the received signal may rely on a plurality of well-known decoders, including linear and non-linear decoders. Embodiments of the present invention propose a unique non-linear decoder for providing decoding operations at receiver 110. Performance of the decoder algorithm is measured in the bit-error rate (BER) of the decoder algorithm as well as computational complexity (e.g., time required to perform the algorithm). It is beneficial to provide good performance (e.g., low BER) at low complexity (e.g., fast computation time). Based on the modulation scheme utilized to modulate the binary stream of data originally, demodulator 112 demodulates the decoded signal to generate a binary stream output.

[0022] These operations can be described mathematically. The input-output relationship of an M x N MIMO system is described by the following equation:

$$\boldsymbol{y} = \boldsymbol{Hs} + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{y} = [y_i, y_2 \ldots, y_{N}]^T$ is the received vector, $\boldsymbol{H}$ is $N \times \boldsymbol{M}$ channel matrix, where each element $Hi_{,j}$ is an independent zero mean circularly symmetric complex Gaussian random variable with variance σ^2_c, and $\boldsymbol{w} = [w_i, w_2, \ldots w_N]^T$, where w_i is an independent zero mean circularly symmetric complex Gaussian random variable with variance σ^2. The transmitted vector (provided by transmitter array 104) is $\boldsymbol{s} = [s_i, s_2, \ldots s_M]$, where s_i is drawn from a finite complex constellation alphabet set of cardinal symbols $\boldsymbol{M_c}$, ($s \; E \; \boldsymbol{S} = \Omega^M$). In general, two types of detection algorithms are utilized to recover transmitted signals, linear decoders and non-linear decoders. Linear decoders operate by separating the received signal into streams by multiplying the received signal $\boldsymbol{y}$ by $\boldsymbol{Hi}_{\boldsymbol{v}}^{\boldsymbol{H}}$ to get the decoded signal $\hat{s} = Q(H_{inh}^H \cdot y)$, wherein Q(-) designates a mapping function to the original constellation. Conventional linear detectors include the maximum ratio combining (MRC), the zero forcing (ZF), and the minimum mean square error (MMSE). Non-linear decoders - such as the MLD calculates the ***a posteriori*** probability in terms of likelihood ratio for each possible transmitted vector $\boldsymbol{s}$ by browsing all the set $\boldsymbol{S}$. The ML estimate of the transmitted vector $\boldsymbol{s}$, $\hat{\boldsymbol{s}}_{ML}$ is given by the following equation:

$$\hat{\boldsymbol{s}}_{ML} = \underset{\boldsymbol{ses}}{\boldsymbol{argmin}} \; \|\boldsymbol{y} - \boldsymbol{Hs}\|^2 \tag{2}$$

The vector s that yields the smallest distance between the received vector y and the hypothesized message H vis selected as the most likely. Because the optimization problem is performed over the set $S = \Omega^M$, the algorithmic complexity of the ML decoder is $0(M_C^M)$, which becomes very high for a large number of transmitted antennas and/or for a large constellation size.

[0023] The sphere decoder is a variant of the ML decoder that reduces the complexity of the ML decoder. In general, the sphere decoder solves equation (2), provided above, by solving for all points which belong to a hypersphere of radius r around the received point y. This can be expressed as follows:

$$\|y - Hs\|^2 \leq r^2 \quad (3)$$

That is, instead of searching each possible transmitted vector s for the one that provides the minimum probability of error, only those points residing within the radius r of the sphere are searched. The search process using the sphere decoder algorithm is a combinatorial optimization problem which can be solved using the ascendant-tree-search algorithm. The two known tree-search algorithms are depth-first-search trees (DFS) and breadth-first-search trees (BFS). In general, the algorithmic complexity of the BFS and DFS algorithms are identical, because the same number of points are ultimately searched. However, as discussed in more detail below, in one embodiment the massive MIMO system 100 utilizes the BFS algorithm to take advantage of additional parallelism as compared with the DFS algorithm. For a fixed radius r, the algorithmic complexity of the sphere decoder is equal to $0(Mc^{\gamma M})$, where $\gamma = \gamma_{H,w,s}(r)$ is a real random variable between 0 and 1 and whom statistic is induced by those of the channel matrix H, the signal noise at the receiver w and the transmitted signal s.

[0024] In one embodiment, receiver 110 utilizes an efficient sphere detector (SD) algorithm (hereinafter "efficient SD algorithm") to reduce the complexity of the traditional SD algorithm. Modifications to the SD algorithm may be utilized alone or in conjunction with one another, and include modifications to selection of the sphere radius as well as implementation of a parallel sphere decoder algorithm that combines the efficient breadth-first-search tree (BFS) algorithm. In some embodiments, receiver 110 is comprised of one

or more processors that execute the efficient sphere detector algorithm to implement a decoder circuit. In some embodiments, at least one or more of the processors is a graphics processing unit (GPU). As discussed in more detail below, one of the benefits of the claimed invention is that the efficient SD algorithm implemented on one or more GPUs utilizes matrix-matrix multiplication that results in operations being compute-bound (i.e., does not require storage of results to memory as is the case with vector-matrix multiplication, which requires storage of a result to memory and subsequent access of the stored result to continue the calculation). Implementation of the efficient SD algorithm in a processor environment that allows for compute-bound operations is a benefit over prior art sphere decoders.

[0025] As illustrated above, the sphere radius utilized by the SD algorithm impacts the performance (in terms of BER) and the complexity (in terms of elapsed time) of the efficient SD detector. As the sphere radius increases, the sphere detector essentially becomes equivalent to the ML detector in which each possible transmitted vector s is searched, which provides high performance (high BER) and high complexity (slow computation time). An ideal choice of the criterion for selecting r is a radius such that the sphere contains the ML salutation as well as the minimum possible number of other candidates from S, so that the algorithm provides high performance (low BER) and low complexity (fast computation time).

Efficient Sphere Radius Selection

[0026] In one embodiment, the efficient SD algorithm converts the radius selection equation provided in Equation **3,** above, into an equivalent problem, wherein channel matrix $H = QR$, wherein QR is the decomposition of the matrix H and wherein Q 6 C^{NxN} is an orthogonal matrix and R6 C^{NxM} is an upper triangular matrix. In addition, $\bar{y} = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N]^T$ is the vector $Q^H y$ and since Q is orthogonal, Equation **3** can be simplified with $\|y - Hs\|^2 = \|\bar{y} - R \cdot s\|^2$. As result, the sphere decoding algorithm can be re-written as follows:

$$\hat{s}_{SD} \triangleq \underset{SEL_M}{argrain} \|\bar{y} - Rs\|^2 \quad \textbf{(4)}$$

$$L_M \triangleq \{s \in S, \|y - Rs\|^2 \leq r^2\}$$

[0027] Equation (4) can be solved by considering the set L_k for $1 \leq K \leq M$ defined by $L_k = \{s_k \in \Omega^k, \|\bar{y}_k - R_k s_k\|^2 \leq r^2\}$, where ¾ , is the lower right bloc of $\boldsymbol{R}$ and $\bar{y}_k$ is the lower bloc of $\bar{y}$. The enumeration procedure is a tree pruning algorithm, which enumerates all points satisfying Equation (4) recursively, starting with $\boldsymbol{k} = 1$.

[0028] In particular, an embodiment of the present invention evaluates the expectation of $\|\bar{y} - Rs\|^2$, which is a random variable such that its probability distribution is induced by those of $\boldsymbol{R}$, $\boldsymbol{s}$ and w, for the case where $\boldsymbol{s} = s_{ML}$. Consequently, assuming that $\boldsymbol{M} > \boldsymbol{N}$, $\|\bar{y} - Rs\|^2$ follows a gamma distribution with parameters $\boldsymbol{N}$ (representative of number of receivers) and variance σ^2, expressed as $j(n, a^2)$. As a result, the expected value of the random variable is expressed as follows:

$$E\|y - Rs\|^2 = N\sigma^2 \tag{5}$$

with the variance of the probability distribution can be expressed as:

$$var(\|\bar{y} - Rs\|^2) = N\sigma^4 \tag{6}$$

As a result, the radius $\boldsymbol{r}$ is selected to satisfy the equation:

$$r^2 = No^2 \tag{7}$$

This selection of the radius $\boldsymbol{r}$ ensures that most of the maximum likelihood (ML) predominant solution candidates will be encapsulated within the sphere defined by radius $\boldsymbol{r}$. In cases where the selected radius $\boldsymbol{r}$ results in an empty sphere, the radius of the sphere $\boldsymbol{r}$ is expanded. In one embodiment, the expansion is provided as follows:

$$\mathbf{r}^2 = \mathbf{r}^2 + var(\|\bar{y} - Rs\|^2) = \mathbf{r}^2 + N\sigma^4 \tag{8}$$

[0029] In this way, an embodiment of the present invention provides for efficient selection of a sphere radius $\boldsymbol{r}$ to be utilized in the efficient sphere detection (SD) algorithm, and for selectively increasing the size of the sphere in the event the initial radius $\boldsymbol{r}$ results in an empty set.

Parallel Breadth-First Search Sphere Decoder Algorithm

[0030] Figure 2 illustrates visually the tree search algorithm. The number of levels included in the tree search algorithm corresponds with the number transmitters $\boldsymbol{M}$. For example, in the embodiment shown in Figure 2, the number of transmitters $\boldsymbol{M}$ is equal to three, and as such the number of levels l_k is equal to three. The parallel breadth-first search

sphere decoder algorithm analyzes the levels of the search tree one by one (breadth first). At each stage l_k, we determine the set L_k defined in Equation (4), which corresponds with those values falling within the defined sphere of radius r^2. At the bottom of the search tree - corresponding with level IM - the minimum of the set L_M is searched, which corresponds to the maximum likelihood (ML) solution of the sphere decoder problem. For each level l_k, $k = 2, \ldots, M$, the matrix M_{k-i} is noted which contains all the vectors of the set L_{k-i}, G_{k-i} the cardinal of L_{k-i}. A matrix V_k is selected with size $(k, M_c G_{k-i})$, which contains all vectors s_{k-1} 6 L_{k-1}, duplicated M_c times and then for each one modified to include a possible constellation symbol from Ω. For each signal ¾ the quantity $\|\bar{y}_k$ —i ̨ $s_{ę}$ $||^2$ is evaluated to determine if it is smaller than the radius defined by r^2.

[0031] A benefit of this approach is that this criterion can be expressed for all signals ¾ simultaneously using a matrix formulation based on a matrix-matrix multiplication computational kernel. In one embodiment, the matrix P_k is defined as follows:

$$P_k = \bar{y}_k - R_k V_k \quad (9)$$

where y_k is the matrix which contains the vector y_k duplicated $M_c G_{k-i}$ times. The benefit of this approach stands in contrast with a depth first search algorithm, in which the criterion would be based on a matrix-vector multiplication kernel, which requires memory-bound operations.

[0032] The matrix 1 x $M_c G_{k-i}$ contains the square Euclidean norm of each vector $P_{\kappa,i}$ in matrix P_k. The previous criterion is reduced now to verifying if the weights in the matrix P_k are lower than r^2 or not (i.e., included under the sphere). For each evaluation, the matrix M_k is generated and the set L_k is derived. If the set L_k is empty, the radius r is increased and the algorithm is started again. For the case of $k=l$, the process is the same as considering L_o, which corresponds to the root node which is an empty set. Upon reaching the last level, IM, the minimum weight in matrix P_M is searched and from the matrix M_M the solution SML is derived.

[0033] According to one embodiment of the present invention, the following algorithm is employed for the case where $M \leq N$. The algorithm is reproduced below:

Algorithm 1 Parallel Standard BFS ¾ ∞ Algorithm

Inputs:
Received signal: y
Constellation order: Ω:
Channel estimation: H
Noise variance estimation: σ^2
QR Decomposition:
$H = QR$
Preliminary:
$\bar{y} = Q^H y$
Initialization:
$r^2 = M\sigma^2$
$M_0 = [\,]$
$G_0 = 1$
$Empty = 1$

1: **while** $Empty = 1$ **do**
2: **for** $k = 1 : 1 : M$ **do**
3: $Calculate\, V_k(M_{K-1})$
4: $\bar{\mathbf{y}}_k = \bar{y}_k \circ mathbbl(1, M_c G_{k-1})$
5: $\mathbf{P}_k = \bar{\mathbf{y}}_k - R_k V_k$
6: $Calculate\, P_k$
7: $size = 0$
8: **for** $i = 1 : 1 : M_c G_{k-1}$ **do**
9: **if** $P_k(i) \leq r^2$ **then**
10: $M_k(:, i) = V_k(:, i)$
11: $size + +$
12: **end if**
13: **end for**
14: **if** $size \neq 0$ **then**
15: $Empty = 0$
16: **else**
17: $Empty = 1$
18: $r^2 = r^2 + M\sigma^4$
19: $break$
20: **end if**
21: **end for**
22: **end while**
23: $\hat{s}_{ML} = M_M(:, min(P_M))$

[0034] In one embodiment, the sample algorithm illustrated above is implemented by a graphical processing unit (GPU) due to the high memory throughput in terms of bandwidth (byte/s) and high computation rate in terms of floating point operations per second (flop/s). However, most GPUs utilize memory located off-chip, which is accessed through a (relatively) slow PCIe link that operates at a bandwidth substantially lower than the internal GPU bandwidth. As a result, for parallel performance it is important to reduce the data off-loading between the CPU/device memory by reusing the freshly moved data on the GPU board as much as possible. A benefit of utilizing the breadth-first tree search algorithm outlined above is that operations involving multiple subsequent levels of the tree are cast into a single matrix-matrix multiplication kernel.

[0035] In addition, one of the challenges of implementing the parallel breadth first search tree resides in the programmability and/or productivity. Scientific codes can be accelerated on GPUs through compiler directives, CUDA programming model, or accelerated libraries (e.g., NVIDIA cBLAS). In one embodiment, the breadth-first search algorithm relies on the NVIDIA implementation of the matrix-matrix multiplication kernel from cuBLAS to deliver high performance computing.

[0036] Figure 3 is a flowchart illustrating steps performed by the efficient SD decoder according to an embodiment of the present invention. At step 300, the sphere radius r is initialized. In the embodiment shown in Figure 3, sphere radius r is defined as a function of the number of transmitters M being utilized in the MIMO configuration. In other embodiments, the sphere radius r is defined as a function of both the number of transmitters M being utilized and the noise variance estimation σ^2 (e.g., $r = M\sigma^2$).

[0037] At step 302, a matrix Vk is defined that has a size $(k, M_c Gk\text{-}i)$. The matrix Vk is comprised of all the vectors $sk\text{-}i$ that are included within the set $Lk\text{-}i$. That is, the matrix Vk is comprised of all the vectors of symbols from levels 1 through $\boldsymbol{k\text{-}l}$ that have been determined to be located within the selected radius r of the sphere. The number of rows in the matrix Vk is set to $\boldsymbol{k}$ instead of $\boldsymbol{k\text{-}l}$ because the vectors $sk\text{-}i$ further includes the next possible symbol selected from the constellation of symbols selected from the set Ω. The number of columns is defined as $M_c Gk\text{-}i$, wherein M_c defines the number of symbols in the constellation set Ω and $Gk\text{-}i$ represents the number of vectors at level $\boldsymbol{k\text{-}l}$ determined to be located within the radius $\boldsymbol{r}$ of the sphere. Because each vector located at level $\boldsymbol{k\text{-}l}$ may be followed by any one of the constellation symbols defined within the set Ω, the number of rows within the matrix is defined as $M_c Gk\text{-}i$. For example, if $Gk\text{-}i$ equals four and the number of symbols in the constellation set Ω equals three, then the number of columns is equal to the twelve. In this way, matrix Vk represents the symbols selected from levels one to $\boldsymbol{k\text{-}l}$ that are located within the radius of the sphere, as well as all possible symbols located at level $\boldsymbol{k}$ that need to be evaluated.

[0038] At step 304, the matrix $\boldsymbol{y_k}$ is generated which contains the vector $\boldsymbol{y_k}$ duplicated $M_c Gk\text{-}i$ times. The vector $\boldsymbol{y_k}$ represents the received signal, and this vector is duplicated $M_c Gk\text{-}i$ to provide a matrix of the same size as Vk defined at step 302.

[0039] At step 306, the matrix $\boldsymbol{P_k}$ is defined as equal to $\boldsymbol{y_k - RkVk}$. If the depth first algorithm was utilized to determine this criterion, then a matrix-vector multiplication vector would be utilized, which would require memory bound operations. In contrast, utilizing a breadth first algorithm allows this criterion to be determined using a matrix-matrix multiplication kernel, which is compute bound and therefore runs close to the

theoretical peak performance of the system. The ***1 x*** M_c*Gk-i* matrix contains the square Euclidean norm of each vector ***I¾ ·+*** in ***Pk.*** The previous criterion is reduced now to just verifying if the weights in the matrix ***Pk*** are lower than ***r²*** or not (i.e., located within the sphere).

[0040] At step 308, for each evaluation we generate the matrix *Mk* and we derive the set *Lk* that indicates those values that satisfy the radius criterion (i.e., fit under the sphere).

[0041] At step 310 the set *Lk* is evaluated to determine if the set is empty. If the set *Lk* is empty, then the radius ***r*** is increased at step 312 and the decoding process is restarted at step 302. If the set is not empty, then at step 314 the minimum weight is searched in matrix *PM* and the maximum likelihood solution *SML* is derived from the matrix *MM*- The ML solution *SML* represents the signal transmitted from the transmitters to the receiver as determined by the efficient SD decoder algorithm.

[0042] Figures 4a-4b, 5a-5b, and 6a-6b illustrate experimental results associated with embodiments of the present invention. Experiments were implemented using a host system equipped with two-sockets of ten-core Intel Ivy Bridge (20 cores total) running at 2:8GHz with 256GB of memory. There are three Tesla Kepler K40 GPU accelerators attached to the CPU board through PCIe Gen3 xl6, each with 12GB of GDDR5 memory. All experiments reported in this section consider only uncoded signals and we adopt the case of a perfect channel state information, i.e., the channel matrix is known by the decoder.

[0043] Figures 4a and 4b illustrate performance and execution times, respectively, of a plurality of decoder algorithms operating at various signal-to-noise ratios (SNRs). Results are based on decoding signals generated on a 50 (transmitter) x 25 (receiver) MIMO system utilizing BPSK modulation.

[0044] In the embodiment shown in Figure 4a, the efficient SD decoder (line 400) is compared with the standard SD decoder (line 402), the minimum mean square error (MMSE) decoder (line 404), the zero forcing (ZF) decoder (line 406), and the maximal-ratio combining (MRC) decoder (line 408). Performance is measured as a bit-error rate (BER) for various signal-to-noise ratios (SNRs). At very low SNRs (poor signal quality), the various decoders perform substantially the same, since the data are uncoded. However,

as SNR improves, the performance of efficient SD decoder (line 400) is superior to the other types of decoders, including SD decoder (line 402). At a SNR of 12 decibels (dB), the efficient SD decoder provides significant improvement over alternative systems.

[0045] In the embodiment shown in Figure 4b, the execution time required to decode signals received on the 50 x 25 MIMO system using BPSK modulation is shown. Once again, data is simulated for a plurality of decoder algorithms, including the efficient SD decoder (line 410) implemented according to an embodiment of the present invention, as well as standard SD decoder (line 412), and linear decoders, including MMSE decoder (line 414), ZF decoder (line 416), and the MRC decoder (line 418). In particular, while the linear decoders (e.g., MMSE, ZF and MRC decoders) provide the best (i.e., smallest) execution time, the execution time of efficient SD decoder 410 improves substantially with improved SNR. As the SNR approaches 21 dB, the execution time of efficient SD decoder 410 is approximately equal with the execution time of standard SD decoder 412, approximately 10^{-2} seconds (e.g., a hundredth of a second, which is approximately real-time).

[0046] Figures 5a and 5b illustrate performance and execution times, respectively, of the efficient sphere decoder (SD) algorithm utilized with respect to different constellations, including 64-QAM (line 500), 16-QAM (line 502), 4-QAM (line 504), and BPSK (line 506). In particular, Figure 5a illustrates that for a fixed number of transmitter and receiver antennas, the performance decreases when we increase the order of the modulation (BPSK = 2^1, 4-QAM = 2^2, 16-QAM = 2^4 and 64-QAM = 2^8). Similarly, Figure 5b illustrates faster execution times for lower order modulations (e.g., BPSK provides the fastest execution) as illustrated for 64-QAM (line 508), 16-QAM (line 510), 4-QAM (line 512), and BPSK (line 514).

[0047] Figures 6a and 6b illustrate performance and execution times, respectively, of the efficient sphere decoder (SD) algorithm utilized with respect to different MIMO configurations of transmitters and receivers as well as different computing platforms (e.g., GPU versus CPU). In particular, Figure 6a illustrates performance measured in BER of a 40 x 10 MIMO configuration (line 600), a 50 x 25 MIMO configuration (line 602) and a

30 x 10 MIMO configuration (line 604). Figure 6b illustrates difference in execution time for different MIMO configuration including whether execution is performed on a CPU or GPU. For example, Figure 6b illustrates a execution time for a 40 x 10 configuration implemented by a CPU (line 606), 50 x 25 configuration implemented by a CPU (line 608), 30 x 10 configuration implemented by a CPU (line 610), 40 x 10 configuration implemented by a GPU (line 612), 50 x 25 configuration implemented by a GPU (line 614), and a 30 x 10 configuration implemented by a GPU (line 616). In particular, Figure 6b illustrates the gain in execution time achieved by utilizing GPUs instead of CPUs, which as described above is attributable to the ability of GPUs to make use of compute-bound computations, rather than memory bound computations.

[0048] In addition, Figures 6a and 6b illustrates that the execution time and the complexity increase as the number of the transmitter antennas increase. For even larger MIMO systems, the efficient SD algorithm may adjust the number of subsequent levels to process simultaneously and/or add

[0049] In this way, the present invention provides a breadth-first-search (BFS) tree in the context of an efficient non-linear sphere decoder algorithm. In particular, the present invention utilizes the parallelism presented by breadth first searching to cast operations involving multiple subsequent levels of the tree into a single matrix-matrix multiplication kernel. Implementation of these operations on a GPU allows for compute-bound processing of the efficient non-linear sphere decoder, as opposed to more costly memory-bound processing.

[0050] While the invention has been described with reference to an exemplary embodiment(s), it will be understood by those skilled in the art that various changes may be made and equivalents may be substituted for elements thereof without departing from the scope of the invention. In addition, many modifications may be made to adapt a particular situation or material to the teachings of the invention without departing from the essential scope thereof. Therefore, it is intended that the invention not be limited to the particular embodiment(s) disclosed, but that the invention will include all embodiments falling within the scope of the appended claims.

CLAIMS:

1. A method of decoding a signal comprised of a plurality of symbols communicated via a multiple input-multiple output (MIMO) communication system, the method comprising:

receiving a signal at a receiver end of the communication channel of a multiple input-multiple output communication system, wherein the received signal is a function of a transmitted signal and the communication channel;

initializing a sphere radius *r* based on attributes of the communication channel; and

defining a set of possible transmit signal symbols based on the initialized sphere radius r; and

selecting a maximum likelihood solution of the transmitted signal based on the selected set of possible transmit signal symbols.

2. The method of claim 1, wherein attributes of the communication channel include number of transmitters utilized in the MIMO communication system, wherein the sphere radius *r* is initialized based, at least in part, on the number of transmitters.

3. The method of claim 2, wherein the sphere radius *r* is selected as a function of the number of transmitters and a noise variance estimation.

4. The method of claim 3, wherein the sphere radius *r* is selected according to the equation $r = M\sigma^2$, wherein *r* is the sphere radius, *M* is the number of the transmitters, and σ is the noise variance estimate.

5. The method of claim 1, wherein the sphere radius *r* is increased if no transmit signal symbols are located within the selected sphere radius.

6. The method of claim 5, wherein the sphere radius r is increased according to the equation: $r^2 = r^2 + M\sigma^4$, wherein r is the sphere radius, M is the number of the transmitters, and σ is the noise variance estimate.

7. The method of claim 1, wherein defining a set of possible transmit signal symbols based on the initialized sphere radius r includes utilizing breadth-first tree traversal to define the set of possible transmit signal symbols based on the sphere radius r.

8. The method of claim 7, wherein defining a set of possible transmit signal symbols based on the initialized sphere radius r includes defining a first matrix Vk having size $(k, McGk\text{-}i)$, for $k = 1 \ldots M$, wherein k represents a search tree level being evaluated, M represents a number of transmitters, M_c represents a number of symbols in a constellation set, and $Gk\text{-}i$ represents a number of vectors at level $k\text{-}1$ determined to be located within the sphere radius r, wherein first matrix Vk is comprised of vectors $sk\text{-}i$ located in the set $Lk\text{-}i$ representing the transmitted signal vectors that satisfy the sphere radius r, and further includes defining a second matrix yk based on the received signal y_k duplicated $M_c G_{k\text{-}i}$ times.

9. The method of claim 8, wherein matrix multiplication is utilized to calculate a third matrix Pk based on $yk - RkVk$, wherein Rk is related to the QR decomposition of channel estimation H, wherein the third matrix Pk is evaluated with respect to the sphere radius r.

10. A decoder circuit for decoding a multiple-input, multiple output (MIMO) signal, the decoder circuit comprising one or more processors configured to:

initialize a sphere radius r based on attributes of the communication channel;

define a first matrix Vk having size $(k, M_c Gt\text{-}i)$, wherein k is search tree level being evaluated, M_c defines a number of symbols in a constellation set, and $Gk\text{-}i$ defines a number of vectors at level $k\text{-}1$ determined to be located within the sphere radius r, wherein the first

matrix Vk is comprised of vectors $sk\text{-}i$ located in the set $Lk\text{-}i$ representing the transmitted signal vectors that satisfy the sphere radius r;

define a second matrix yk based on the received signal y_k duplicated $M_c Gt\text{-}i$ times;

utilize matrix-matrix multiplication to calculate a third matrix P_k based on the equation $Pk = yk - RkVk$, wherein Rk is related to the QR decomposition of channel estimation H, wherein the third matrix Pk is evaluated with respect to the sphere radius r to derive the set of vectors Lk satisfying the sphere radius r for $k = 1 \ldots M$, wherein M is the number of transmitter; and

select a maximum likelihood solution of the transmitted signal based on the matrix PM, wherein $P\backslash i$ is the calculated third matrix P_k when $k = M$.

11. The decoder circuit of claim 10, wherein attributes of the communication channel include number of transmitters utilized in the MIMO communication system, wherein the sphere radius r is initialized based, at least in part, on the number of transmitters.

12. The decoder circuit of claim 11, wherein the sphere radius r is selected as a function of the number of transmitters and a noise variance estimation.

13. The decoder circuit of claim 10, wherein the sphere radius r is increased according to the equation: $r^2 = r^2 + M\sigma^4$ if no transmit signal symbols are located within the selected sphere radius, wherein r is the sphere radius, M is the number of the transmitters, and σ is the noise variance estimate.

14. The decoder circuit of claim 10, wherein the matrix-matrix multiplication is a compute-bound computation performed by the processor.

15. The decoder circuit of claim 14, wherein the one or more processors are graphics processing units (GPUs).

16. A method of decoding a signal comprised of a plurality of symbols communicated via a multiple input-multiple output (MIMO) communication system, the method comprising:

defining a first matrix of possible transmit signal symbols and a second matrix of received symbols based on a breadth-first search (BFS) of search tree levels defined by potential transmitted symbols;

searching a level of the search tree utilizing matrix multiplication to determine select symbols satisfying the initialized sphere radius; and

deriving a maximum likelihood solution of the transmitted symbols based on the selected symbols.

17. The method of claim 16, wherein defining a first matrix of possible transmit signal symbols based on the initialized sphere radius r includes defining a first matrix Vk having size $(k, McGk\text{-}i)$, for $k = 1 \ldots M$, wherein k represents a search tree level being evaluated, M represents a number of transmitters, M_c represents a number of symbols in a constellation set, and $Gk\text{-}i$ represents a number of vectors at level $k\text{-}1$ determined to be located within a selected sphere radius r, wherein first matrix Vk is comprised of vectors $sk\text{-}i$ located in the set $Lk\text{-}i$ representing the transmitted signal vectors that satisfy the sphere radius r, and further includes defining a second matrix yk based on the received signal y_k duplicated $McGk\text{-}i$ times.

18. The method of claim 17, wherein matrix multiplication is utilized to calculate a third matrix P_k based on $y_k - RkVk$, wherein Rk is related to the QR decomposition of channel estimation H, wherein the third matrix Pk is evaluated with respect to the sphere radius r.

19. The method of claim 18, wherein deriving a maximum likelihood solution is selected based on the matrix PM, wherein PM is the calculated third matrix Pk when $k = M$.

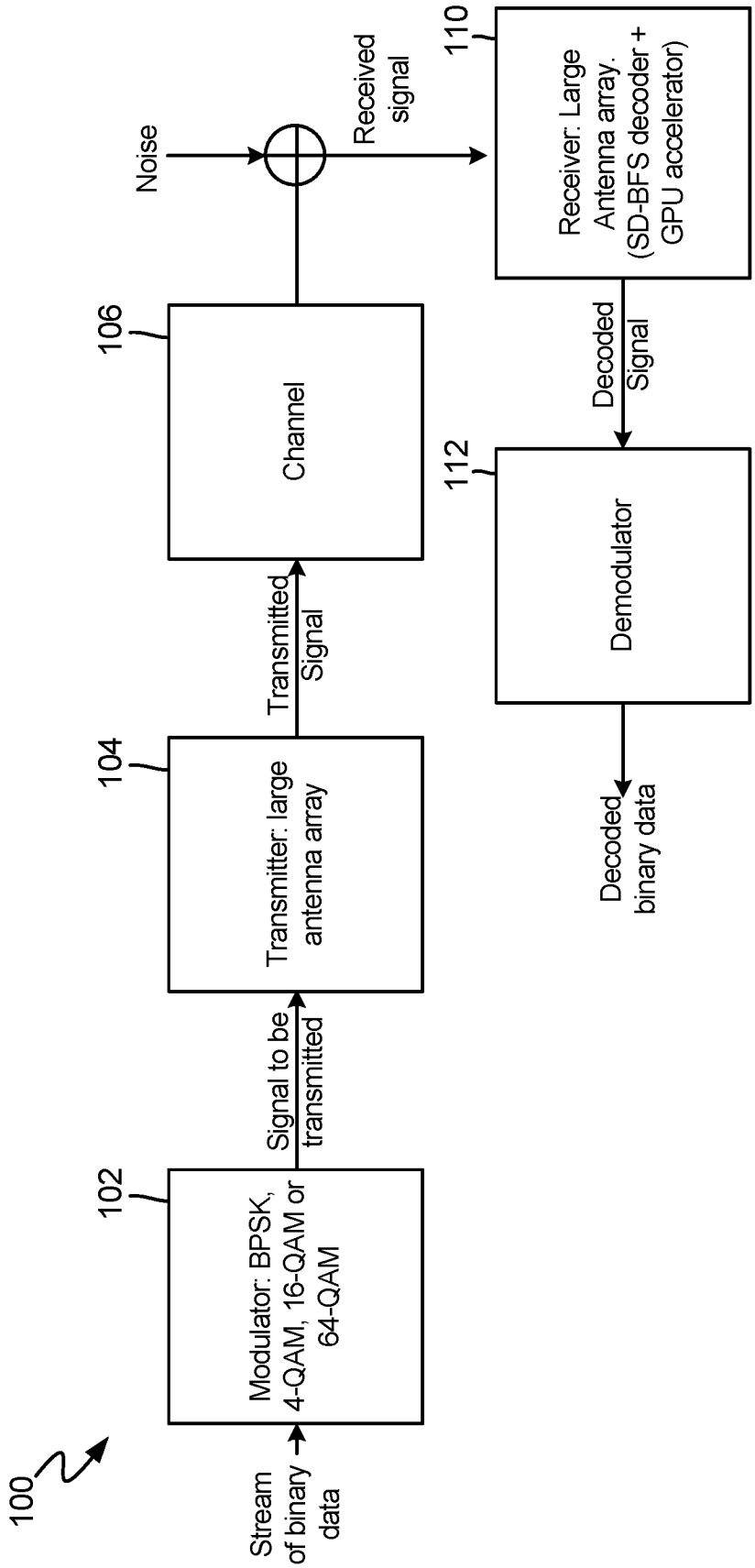
100
Stream of binary data
102
Modulator: BPSK, 4-QAM, 16-QAM or 64-QAM
Signal to be transmitted
104
Transmitter: large antenna array
Transmitted Signal
106
Channel
Noise
Received signal
110
Receiver: Large Antenna array. (SD-BFS decoder + GPU accelerator)
Decoded Signal
112
Demodulator
Decoded binary data

FIG. 1

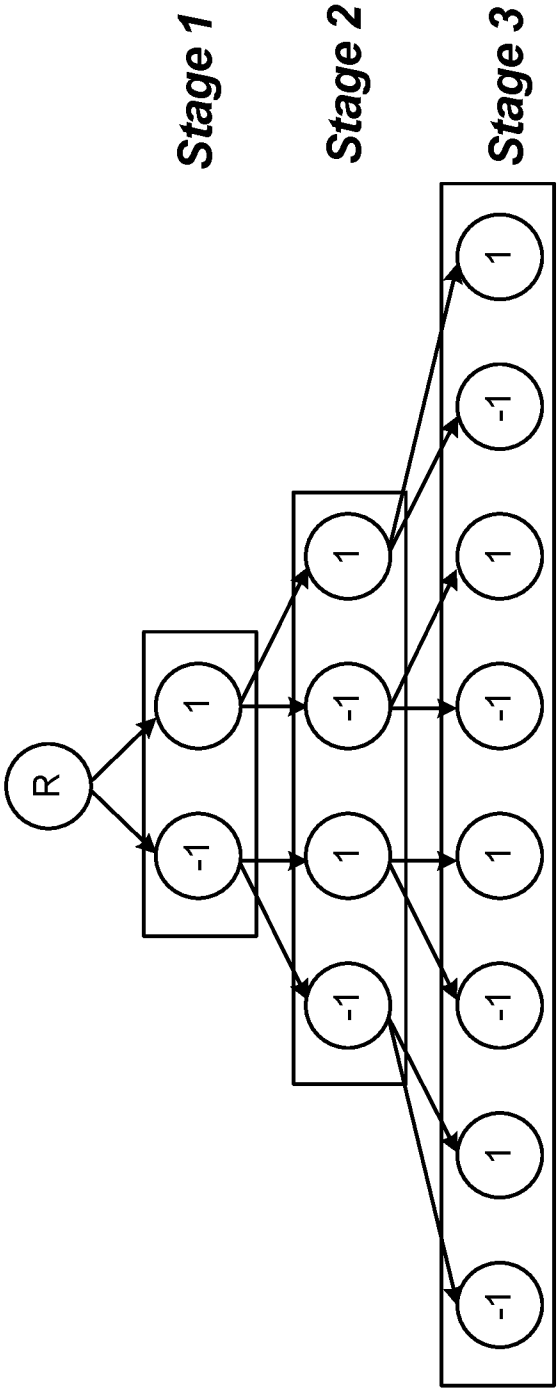

FIG. 2

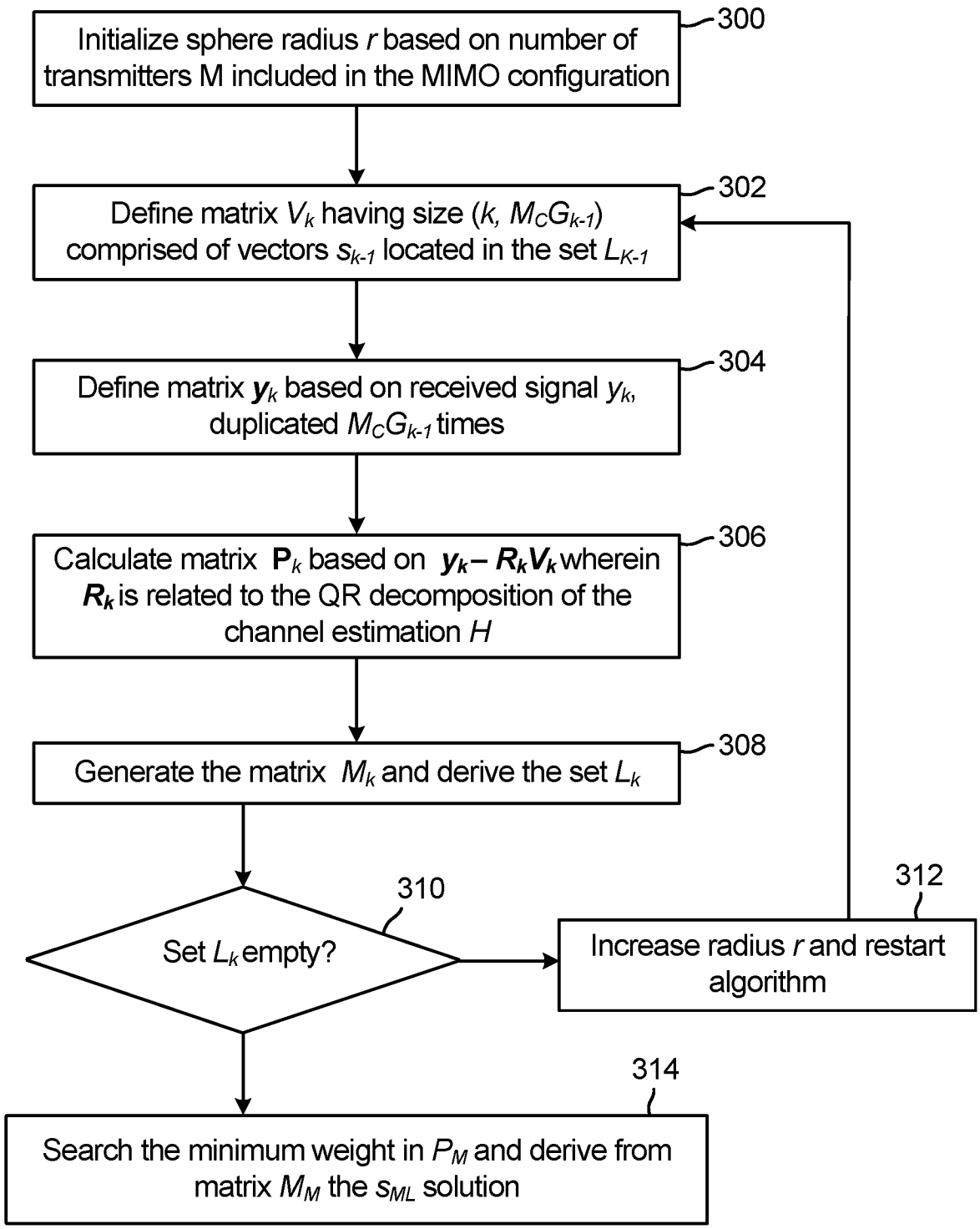
Initialize sphere radius r based on number of transmitters M included in the MIMO configuration
300
Define matrix V_k having size (k, $M_C G_{k-1}$) comprised of vectors s_{k-1} located in the set L_{K-1}
302
Define matrix $\boldsymbol{y}_k$ based on received signal y_k, duplicated $M_C G_{k-1}$ times
304
Calculate matrix $\mathbf{P}_k$ based on $\boldsymbol{y_k} - \boldsymbol{R_k V_k}$ wherein $\boldsymbol{R_k}$ is related to the QR decomposition of the channel estimation H
306
Generate the matrix M_k and derive the set L_k
308
Set L_k empty?
310
Increase radius r and restart algorithm
312
Search the minimum weight in P_M and derive from matrix M_M the s_{ML} solution
314

FIG. 3

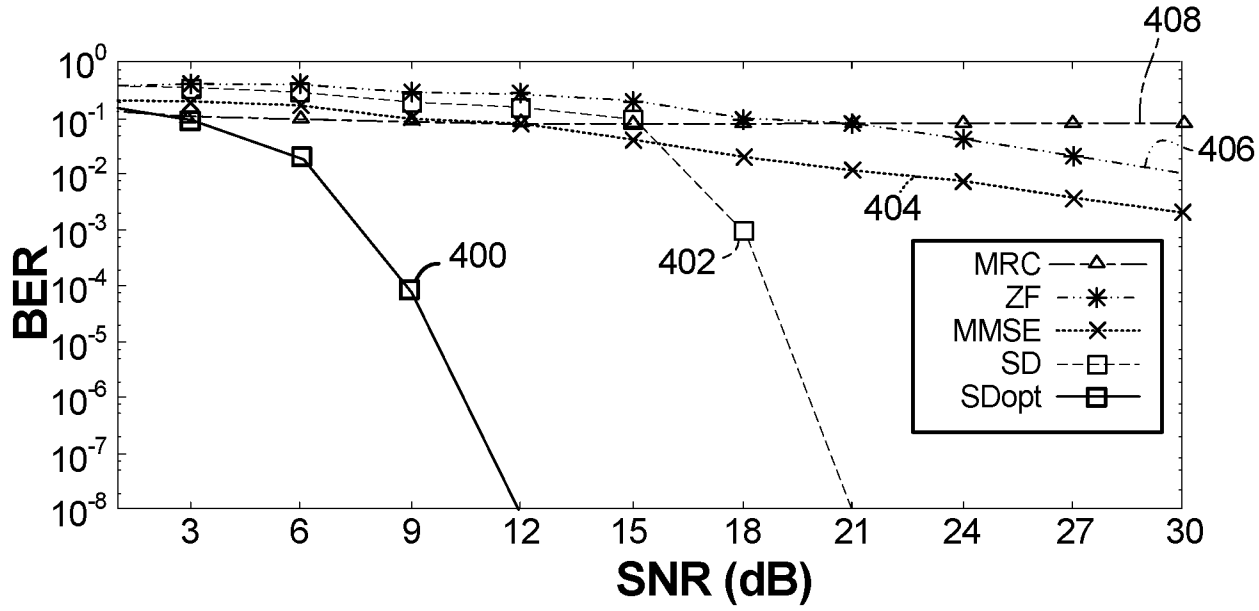
BER
SNR (dB)
MRC
ZF
MMSE
SD
SDopt
400
402
404
406
408

FIG. 4A

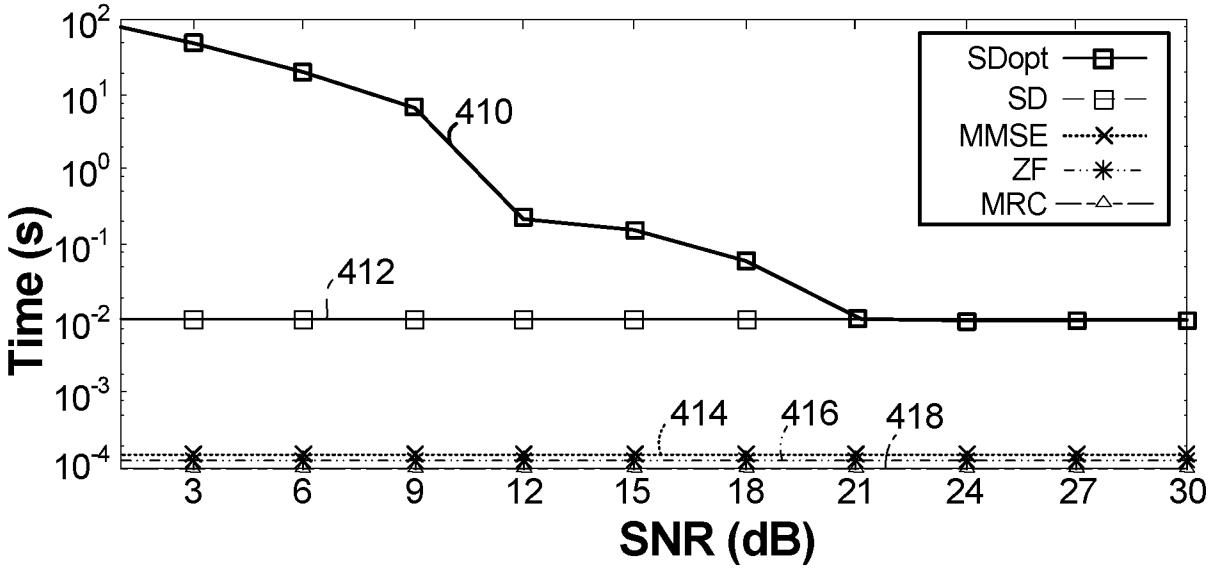
Time (s)
SNR (dB)
SDopt
SD
MMSE
ZF
MRC
410
412
414
416
418

FIG. 4B

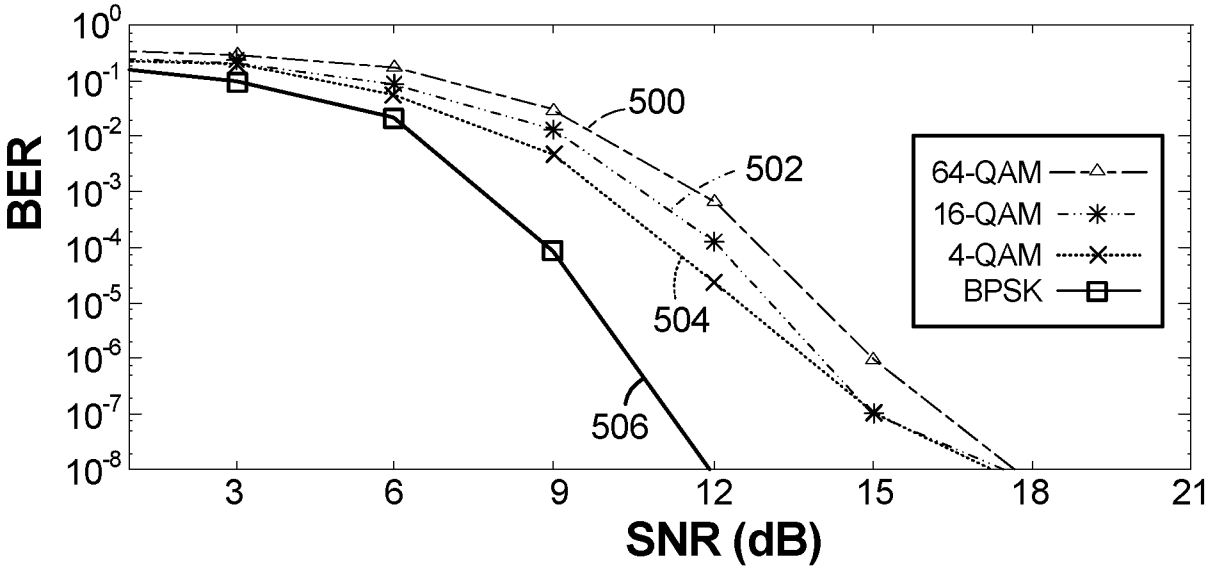
BER
SNR (dB)
500
502
504
506
64-QAM
16-QAM
4-QAM
BPSK

FIG. 5A

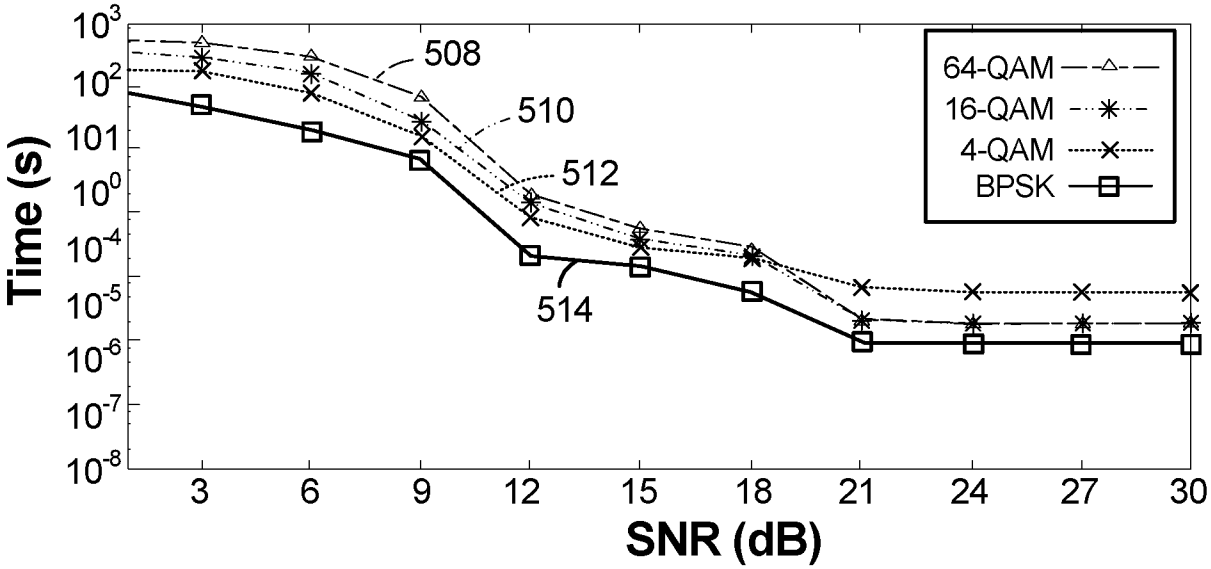
Time (s)
SNR (dB)
508
510
512
514
64-QAM
16-QAM
4-QAM
BPSK

FIG. 5B

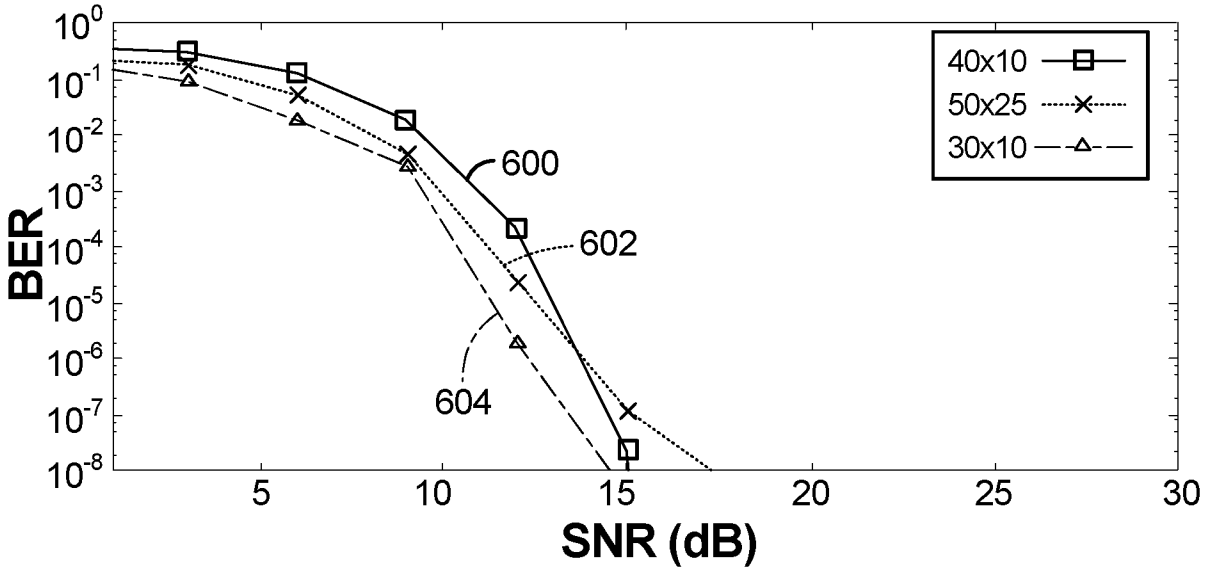
BER
SNR (dB)
40x10
50x25
30x10
600
602
604

FIG. 6A

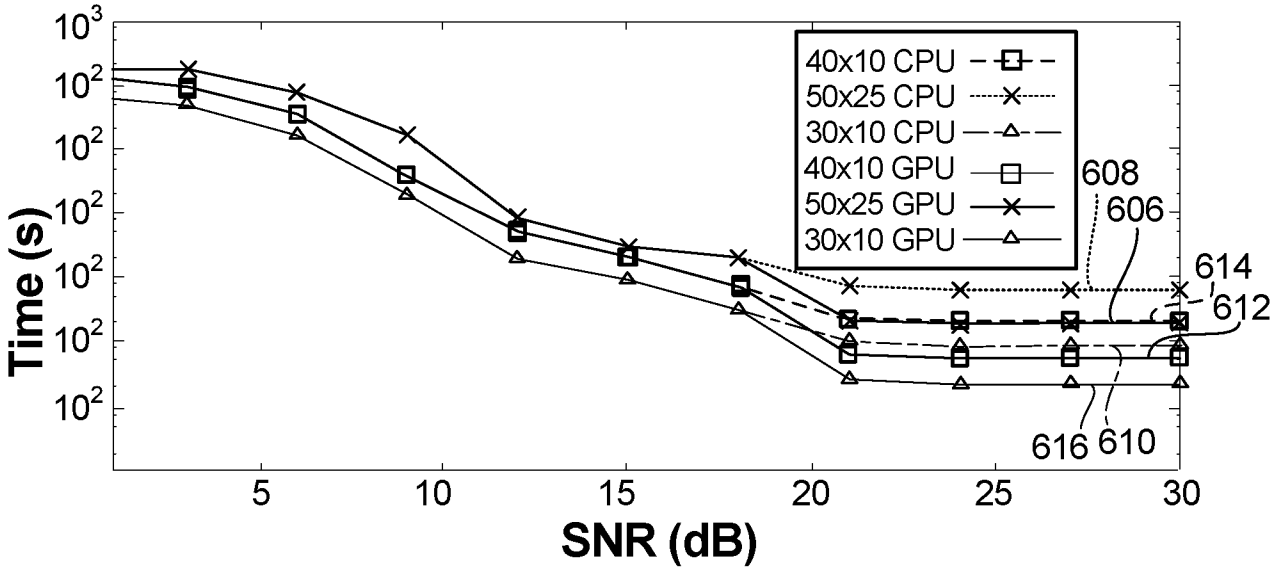
Time (s)
SNR (dB)
40x10 CPU
50x25 CPU
30x10 CPU
40x10 GPU
50x25 GPU
30x10 GPU
608
606
614
612
616
610

FIG. 6B

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2017/055913

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04B7/04 H04L25/03 H04B7/0413
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04B H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ARFAOUI MOHAMED-AMINE ET AL: "Efficient Sphere Detector Algorithm for Massive MIMO using GPU Hardware Accelerator", PROCEDIA COMPUTER SCIENCE, ELSEVIER, AMSTERDAM, NL, vol. 80, 1 June 2016 (2016-06-01), pages 2169-2180, XP029565756, ISSN: 1877-0509, DOI: 10.1016/J.PROCS.2016.05.377 the whole document ----	1-19

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" documentwhich may throw doubts on priority claim(s) orwhich is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 December 2017	20/12/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lutz, Tobias