Int. J. Electron. Commun. (AEÜ) 70 (2016) 1443–1448

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

journal homepage: www.elsevier.com/locate/aeue

Regular paper

Performance and efficiency analysis of THP algorithms: A practical perspective

Luechao Yuan [a,b,*], Xiaolin Chen [b], Daniel Günther [b], Chuan Tang [a], Gerd Ascheid [b], Anupam Chattopadhyay [c], Zuocheng Xing [a]

[a] College of Computer, National University of Defense Technology, 410073 Changsha, Hunan, PR China
[b] Institute for Communication Technologies and Embedded Systems, RWTH Aachen University, 52056 Aachen, Germany
[c] School of Computer Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 6 July 2015
Accepted 10 August 2016

Keywords:
Tomlinson–Harashima precoding
Complexity
Efficiency
VLSI

ABSTRACT

As an attractive interference cancellation (IC) technique, Tomlinson–Harashima precoding (THP) has been investigated thoroughly in theory. Several high performance THP variants have been proposed, e.g., sorted QR decomposition (SQRD), Cholesky decomposition, vertical Bell Laboratories space time (V-BLAST) and lattice reduction aided THPs. From a practical perspective, however, limited hardware implementations have been reported in the literature so far. To bridge the progress gap between the theory and the practice, we present a comprehensive analysis of these THP variants in terms of performance and implementation efficiency in this paper. We first evaluate their bit-error rate (BER) performance under perfect and imperfect channel state information (CSI) scenarios. Subsequently, the emphasis is put on their implementation efficiency, including the computational complexity, the numerical precision requirement and the parallelism potential. Our analysis shows a wide trade-off space exists between the performance and the implementation efficiency in different THP variants, which is especially valuable for hardware designers to implement cost-efficient architectures biased towards practical systems.

1. Introduction

Wireless communication systems are experiencing massive growth in terms of traffic and devices volumes currently. However, the capacity limitations confine many potential user demands and application perspectives. In fact, the interference plays a crucial role in current wireless systems which induces a typical upper bound to the system performance. To mitigate the mutual interference caused by parallel transmission to different users/antennas, many interference cancellation (IC) methods have been proposed in the literature [1]. They can be classified into two main types, i.e., pre-IC at the transmitter side and post-IC at the receiver side. Considering the uplink/downlink duality [2,3], without loss of generality, we focus on the downlink scenario, i.e., transmitter side processing techniques, in this paper.

Transmitter side processing strategies can be generally divided into two categories, i.e., linear pre-equalization and non-linear precoding. Due to the effect of noise enhancement, linear methods have a poor power efficiency. Alternatively, this disadvantage can be overcome by non-linear precoding strategy, e.g. Tomlinson–Harashima precoding (THP) [4,5]. In contrast to its dual to decision-feedback equalization (DFE) at the receiver, THP avoids error propagation since the transmitter have the exact knowledge of data signals to be sent. Moreover, THP can be considered as the one-dimension approximation of dirty paper coding (DPC) [6,7], which is the most practical candidate to achieve the capacity promised by DPC [8] with feasible computational complexity.

It has been demonstrated that the processing order has a great impact on the system performance [9]. The order plays a key role in the sense of maximizing the signal-to-noise ratio (SNR) of the sub-channels [10] or in the sense of minimizing the error-variance between the data signal and the received signal [11]. A large volume of works exists on permuting the processing order of the users according to the channel matrix $\boldsymbol{H}$ in the literature. Briefly, these sorting methods can be classified into three categories according to their reordering criteria: column norms of $\boldsymbol{H}$, column norms of $\boldsymbol{H}^{-1}$ and diagonal entries of $\boldsymbol{H}\boldsymbol{H}^{H}$, which typically correspond to SQRD [12], V-BLAST [13] and Cholesky decomposition [11], respectively. Moreover, lattice reduction [14] aided THP is able to improve the system performance further.

* Corresponding author at: Institute for Communication Technologies and Embedded Systems, RWTH Aachen University, 52056 Aachen, Germany.
E-mail address: luechao.yuan@ice.rwth-aachen.de (L. Yuan).

http://dx.doi.org/10.1016/j.aeue.2016.08.008
1434-8411/

Although these algorithms have been proposed for a decade, the amount of published works of their implementations has been limited. As far as we know, two hardware implementations of THP algorithms without reordering have been reported in [15,16] respectively. For more sophisticated THP algorithms which involve reordering, few implementation has been published so far.

To bridge the mismatch between the theory and the practice, we analysis the performance and the implementation efficiency of these algorithms in this paper. The BER performance, which is a critical criterion in practical systems, is evaluated first under both perfect and noisy CSI scenarios. To make a fair and realistic comparison, the evaluation is performed based on a unified multi-user MIMO OFDM platform. Afterwards, the emphasis of this work is focused on the implementation efficiency analysis of THP algorithms with reordering. Our analysis is carried out in three directions, i.e., computational complexity, numerical precision and parallelism potential, which are tightly related to the area, throughput, energy consumption and computation delay of their VLSI implementations. The experimental results show a wide trade-off space spreading over the performance and the implementation efficiency of these sophisticated THP algorithms, which is especially valuable for hardware designers to find a cost-efficient solution to satisfy the system requirements.

1.1. Outline

The rest of this paper is organized as follows. Firstly, the MIMO system model and the channel model are described in Section 2. And the THP variants are introduced in Section 3. Afterwards, the performance of the aforementioned THP algorithms are simulated and analyzed in Section 4. Importantly, the implementation efficiency analysis of these algorithms is presented in Section 5 and the trade-off space is discussed in Section 6. Finally, this paper is concluded in Section 7.

1.2. Notation

Some symbols and operators used in this paper are defined here. Lower letters stand for scalars (e.g., σ), and bold lower letters represent vectors (e.g., $\boldsymbol{r}$). Matrices are denoted by bold upper-case symbols (e.g., $\boldsymbol{H}$), and $h_{i,j}$ stands for the element located in the ith row and the jth column of $\boldsymbol{H}$. The notation $\mathrm{E}[\bullet]$ denotes expectation and $(\bullet)^{-1}$ denotes inversion.

2. System model

2.1. Multi-user MIMO downlink model

As a typical implementation case of THP, we assume a multi-user MIMO downlink system. A central access point (AP) equipped with N_T transmit antennas serves $K(\leqslant N_\mathrm{T})$ scattered non-cooperating receivers. Each receiver is equipped with one receive antenna. The system model in the equivalent complex baseband for each data-carrying OFDM tone is depicted in Fig. 1. The system inputs $u_i, i = 1,\ldots,K$ are randomly drawn from a normalized M-ary square QAM signal constellation $\mathcal{S} = \{(a + \mathrm{j}b)/\alpha | a, b \in \pm 1, \pm 3, \ldots, \pm(\sqrt{M} - 1)\}$.[1] The channel inputs $x_i, i = 1,\ldots,N_\mathrm{T}$ are generated according to u_i by applying precoding at the transmitter, and then transmitted by the N_T antennas simultaneously. The final effective channel gain observed between the ith receiver and the jth transmit antenna is denoted by $h_{i,j}$, which represents a weighted sum of several propagation paths over time and space. n_i is the additive white zero-mean complex Gaussian noise at the input of the receiver i.

[1] The normalization factor α is equal to $\sqrt{2}, \sqrt{10}$ and $\sqrt{42}$ for 4-, 16- and 64-QAM, respectively.

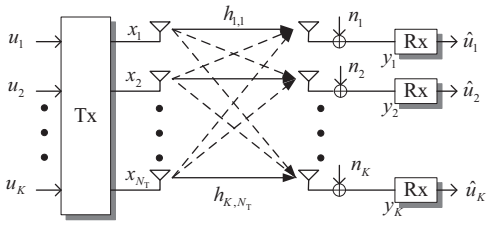

Fig. 1. MIMO downlink system model with central transmitter and decentralized receivers. Dashed arrows represent the interferences.

2.2. Channel model

To evaluate the precoding performance under realistic conditions, we adopt the TGac channel model B [17] as the simulating channel. The TGac channel models a set of indoor wireless channels proposed by the IEEE 802.11 working group. They assume the radio waves arrive in clusters [18], each of which is associated with a set of taps. For example, the ith channel tap for the kth user $\boldsymbol{h}_k^i$ can be modeled as a vector of zero-mean complex Gaussian distributed random variables, i.e.,

$$\boldsymbol{h}_k^i \sim \mathcal{N}_\mathbb{C}(0, \boldsymbol{C}_k) \tag{1}$$

where $\boldsymbol{C}_k$ is the spatial correlation matrix of the cluster.

In the tth time instance, the tap vector is simulated by

$$\boldsymbol{h}_k^i[t] = \boldsymbol{C}_k^{1/2} \boldsymbol{h}_{w,k}^i[t] \tag{2}$$

where $(\bullet)^{1/2}$ represents the Cholesky decomposition and $\boldsymbol{h}_{w,k}^i[t]$ is a vector of complex Gaussian process which has taken the time-dispersive fading (Doppler effect) into account. Finally all the taps are collected by the tapped delay line (TDL) model [19] to model the frequency-dispersive fading effect. The CSI at the transmitter is obtained by sending the preamble sequence first and then collecting the estimated CSI relayed by the receivers. The channel estimation and the following processing on the obtained channel matrices are performed for each of the data-carrying OFDM tone. A detailed error model including the channel estimation error and the feedback delay error can be found in our previous work [20].

3. Tomlinson–Harashima precoding

On a per-tone basis, Fig. 2 shows the block diagram of THP which consists of two processing phases: the channel matrix preprocessing phase and the interference cancellation (IC) phase. The preprocessing phase provides permutation filter ($\boldsymbol{P}$), feedback filter ($\boldsymbol{B}$) and feedforward filter ($\boldsymbol{W}$) for the IC phase. The transmitted symbols for different users are first reordered by $\boldsymbol{P}$. Subsequently, the interference caused by the up-layer users is canceled when precoding the low-layer users using $\boldsymbol{B}$. Finally, the reverse interference from the low-layer users to the up-layer users is suppressed by the feedforward filter $\boldsymbol{W}$. Consequently, it results to several clean and parallel subchannels for all the users. The scaling factor β is used to satisfy the total transmit power constraint.

The modulo operation is used to restrict the signal power increased by the non-linear feedback stage, which is defined as

$$M(x) = x - \left\lfloor \frac{\mathrm{Re}(x)}{\lambda} + \frac{1}{2} \right\rfloor \lambda - \mathrm{j} \left\lfloor \frac{\mathrm{IM}(x)}{\lambda} + \frac{1}{2} \right\rfloor \lambda \tag{3}$$

where $\lfloor \bullet \rfloor$calculates the integer no greater than the argument and the constant λ is determined by the employed constellation, e.g.,

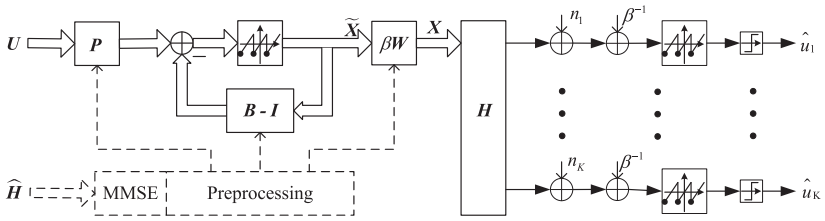

Fig. 2. Block diagram of Tomlinson–Harashima precoding framework for decentralized receivers. Dashed part stands for the preprocessing phase of THP. Since the superior performance when taking the additive noise into account, we only consider the augmented channel matrix (MMSE) instead of $\boldsymbol{h}$ in this paper.

$\lambda = 2\sqrt{2}$ for QPSK. The output signal $\tilde{x}_i$ is asymptomatically uniformly distributed within $[-\lambda/2, \lambda/2)$ for both real and imaginary parts with infinitely large matrices, which will cause a transmitting power increase compared to the discrete signaling set $\mathcal{S}$ by a factor of $M/(M-1)$.

At the receiver end, the scaling effect at the transmitter is compensated by β^{-1}, which is actually performed by an automatic gain control mechanism in practice. Then the same modulo operation is applied to mitigate the modulo effect at the transmitter. Finally, the output of the modulo operator is quantized to get the estimated $\hat{u}_i$ as the recovered signal.

3.1. THP preprocessing algorithms

As we can observe from Fig. 2, the performance of THP depends heavily on the filters generated by the preprocessing phase. There are two optimizing approaches for THP preprocessing, i.e., maximizing the equivalent channel SNR [10] and minimizing the mean squared error (MSE) between the data signals and the received signals [11]. Two heuristic sorting strategies, SQRD [12] and V-BLAST [13], can be exploited to approximate the optimal SNR, which perform a greedy reordering according to the column norm of $\boldsymbol{H}$ and $\boldsymbol{H}^{-1}$ respectively. Alternatively, derived from the view of MMSE, an ordered Cholesky decomposition[2] based on the diagonal entries of $\boldsymbol{H}\,\boldsymbol{H}^H$ is able to achieve optimum performance [11].

Moreover, as a low-complexity but full diversity achieving technique [21], lattice basis reduction (LR) [14] is able to improve the system performance further. The idea is to find a set of basis which has a smaller orthogonality defect to represent the lattice spanned by the columns of the channel matrix $\boldsymbol{H}$. The precoding process is performed based on the new and reduced basis, which results in better performance since the basis is close to orthogonal. The transformation between the two sets of basis is reversed at the very beginning of THP so that we can get the desired results in the end. To reduce the computational complexity, we choose the reverse Siegel (RSiegel) criterion based SQRD algorithm [22] in this paper.

4. Performance analysis

4.1. Simulation setup

In this section, we present the numerical BER results of the THP algorithms over the TGac based multi-user MIMO OFDM platform. Part of the parameters used to generate the channel taps and to build the system are listed in Table 1, please refer [17,23] for more detail.

To simulate the realistic channel effect, the signals to be sent to channel are filtered by all taps using the tapped delay line model. Assuming a static channel within one OFDM symbol duration, all signals in this OFDM frame are filtered by the same set of tap coefficients. This set of coefficients is indexed by $k_t^{u_i}$ for the ith user at the tth frame. The fading channel is simulated by increasing the set of indexes by a constant C, i.e., $k_{t+1}^{u_i} = k_t^{u_i} + C$, when filtering the next OFDM frame. As shown in Fig. 3, to simulate the CSI feedback delay, we first send a sounding preamble frame to measure the channel; and then the CSI feedback delay D follows, during which time the transmitter receives the CSI fed back from the receivers; thereafter, the data frames are transmitted finally.

Table 1
Summary of the setup parameters of the MIMO OFDM simulation platform.

Carrier frequency	5.25 GHz	Environment speed	1.2 km/h
802.11 Case	B	Doppler spread	≈6 Hz
Bandwidth	40 MHz	FFT size	128
Number of clusters	2	Number of taps	9
Coherence times simulated	1000	Channel samples of each tap	445709

Moreover, the used SNR is defined as the SNR received at the receiver

$$\mathrm{SNR} = \frac{N_T \sigma_x^2}{\sigma_n^2} \tag{4}$$

where $\sigma_x^2 = \mathrm{E}[xx^H]$ is the average transmitted power, by which the power increase caused by the modulo operation has been taken into account. The BER performance is used as the benchmarks for the following comparisons. Note that all the results are mean of 150,000 channel realizations, and 20 OFDM symbols are transmitted under each channel realization. As the typical target BER range is $10^{-2} \sim 10^{-3}$, these system settings are sufficient enough.

4.2. Numerical results

Fig. 4 shows the averaged system BER results of the aforementioned THP algorithms without the channel estimation error and the CSI delay error under 16-QAM modulation. It is the perfect CSI case that acts as the comparison baseline for the following imperfect CSI case. One can see that the BER of MMSE is inferior compared to the other three algorithms. It is also observed that, at low SNR region (<40 dB), RSiegel and SQRD perform better than V-BLAST does; however, the BER curve of V-BLAST decreases more significantly than the curves of RSiegel and SQRD do at high SNR region. A closer observation illustrates that the BER curves of RSiegel and SQRD are almost overlapped when SNR<30 dB. The reason is that the average time $\bar{\rho}$ of lattice reduction in RSiegel increases with increasing SNR. As shown in Fig. 5, the effect of lattice reduction is negligible below 20 dB so that the performance is determined by SQRD. However, at high SNR region, this effect becomes prominent and thereby improves the system performance significantly.

Fig. 6 depicts the system BER results with the channel estimation error and the CSI delay error. Compare to Fig. 4, significant

[2] This method is named as MMSE in the rest of the paper.

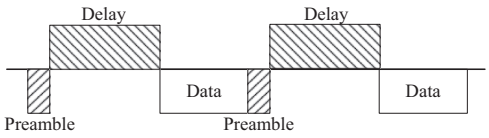

Fig. 3. The simulated CSI feedback scheme.

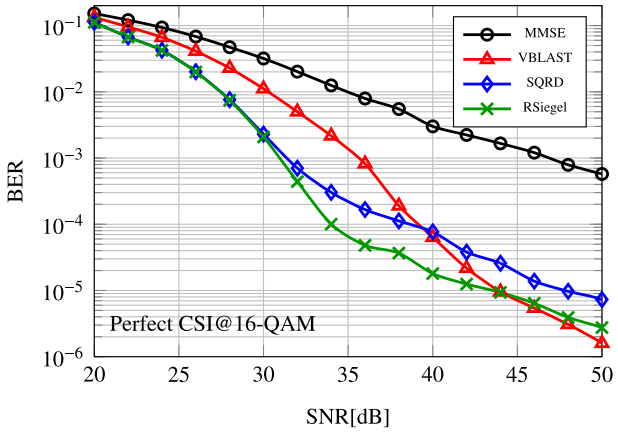

Fig. 4. Averaged uncoded BER performance of various preprocessing algorithms without channel estimation error and feedback delay error of a MIMO OFDM system with $N_R = N_T = 4$, 16-QAM.

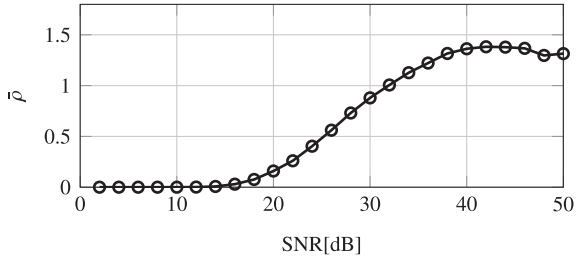

Fig. 5. Average times of lattice reduction $\bar{\rho}$ in the RSiegel algorithm of a system with $N_R = N_T = 4$.

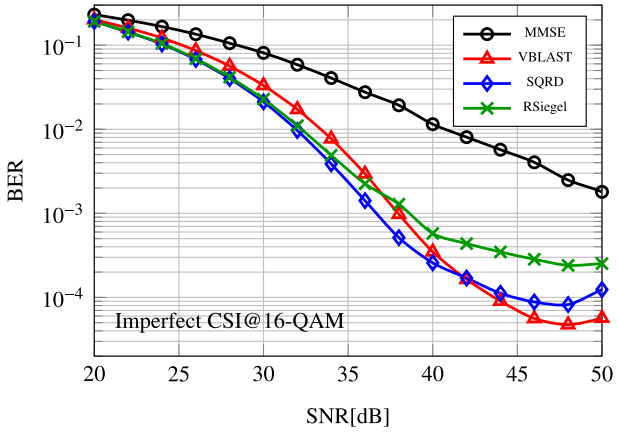

Fig. 6. Averaged uncoded BER performance of various preprocessing algorithms with channel estimation error and feedback delay $D = 1$ of a MIMO OFDM system with $N_R = N_T = 4$, 16-QAM.

performance degradation can be observed due to the sensitivity of THP towards imperfect CSI. Moreover, the BER curve of RSiegel goes higher than SQRD at high SNR region, which performs reversely compared to the results under perfect CSI case. This fact implies that RSiegel algorithm is more sensitive to imperfect CSI than the other algorithms. In contrast, V-BLAST shows better stability against imperfect CSI and is still able to achieve the lowest BER result at high SNR region.

It is worthy to mention that a 10^{-5} BER is not necessary for the uncoded performance requirement in a practical system. Actually, the rest errors can be corrected by using some forward error correction (FEC) codes. Considering the complexity of FEC codes ranged from linear block codes to irregular low-density parity-check (LDPC) or turbo codes [24], system designers can chose different combination of precoding scheme and FEC based on the realistic working condition to satisfy system requirements and save cost. A detailed efficiency analysis of precoding schemes is presented in the following section.

5. Implementation efficiency analysis

In this section, we compare the implementation efficiency of the aforementioned THP algorithms in terms of the numerical precision, the computational complexity and the parallelism potential. Numerical precision and computational complexity are tightly related to the cost efficiency of their implementations, e.g., area, throughput and power consumption. Moreover, parallelism potential of these algorithms implies their control overhead and computation delay. Therefore, the implementation efficiency exploration is of great importance at the early design stage. Considering the sensitivity of THP towards CSI imperfection, the comparisons are based on floating-point arithmetic to achieve high accuracy. Additionally, due to the complexity imbalance of the preprocessing phase and the IC phase of THP, we put emphasis on the preprocessing algorithms.

5.1. Numerical precision

A floating-point number consists of three parts, i.e., the sign bit S, the exponent bits E and the mantissa bits M. The real value of a floating-point represented number a is calculated as $a = S \cdot M \cdot 2^E$, where E represents a signed integer and M ranges within $[1, 2)$. To identify the minimal numerical precision of the algorithms without degrading the system performance, a customized floating-point operation library is built, where we can change the width of the exponent bits and the mantissa bits arbitrarily. A set of $E(M)$ bits with a sufficient large $M(E)$ bits is simulated.

As an exemplary case, Fig. 7 illustrates the BER performance of the algorithms with increasing mantissa bits under 16-QAM modulation. We can observe that the performance improvement caused by one more mantissa bit will saturate at some point. These saturate points determine the minimal numerical precisions, where the performance degradation is negligible compared to their full precision cases. Generally, the better the performance of one algorithm can achieve, the more mantissa bits are required. For example, due to its inferior performance, MMSE requires the least precision compared to the other three algorithms; in contrast, the V-BLAST algorithm requires more bits for its superior performance.

The complete results are summarized in Table 2. Interestingly, we find that RSiegel has one mantissa bit advantage over SQRD and V-BLAST to achieve the same BER. Therefore, it is the most numerical stable algorithm among the listed algorithms. A closer comparison between RSiegel and SQRD implies that lattice reduction is able to improve the numerical stability. The diverse bitwidth requirement of these algorithms under different modulations enable the designers to develop a customized processor with

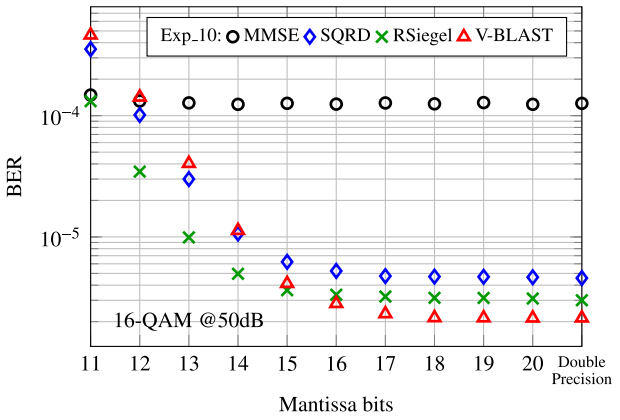

Fig. 7. Average uncoded BER performance of the preprocessing algorithms with various mantissa bits of a system with $N_R = N_T = 4$.

Table 2
Floating-point precision requirement of a system with $N_R = N_T = 4$ for **V**-BLAST, **S**QRD, **R**Siegel and **C**holesky preprocessing algorithms.

	QPSK				16-QAM				64-QAM			
	V	S	R	C	V	S	R	C	V	S	R	C
Mantissa	11	11	10	9	17	16	15	11	17	17	17	11
Exponent	5	5	5	5	5	5	5	6	6	5	5	6

configurable bitwidth to save power consumption without compromising performance.

5.2. Computational complexity

Table 3 lists the real operations required by various preprocessing schemes to calculate the three filters and the scale factor β for each OFDM tone.[3] According to the properties of the implemented floating-point operation cores listed in [26], by comparing the product of the number of stages and the area (slices), the implementation efforts of a division and a square-root are equivalent to $8\times$ and $4\times$ floating-point multiplications, respectively. In the worst case, one lattice reduction calculation involves $4\rho(12N-2)$ real multiplications and $4\rho(12N-2)$ real additions. Additionally, the calculation of matrix inversion used in linear MMSE pre-equalization is also considered as a reference, which is implemented based on the modified Gram-Schmidt QR decomposition [27].

According to Table 3, although the computational complexity of the IC phase of THP is slight higher than the linear pre-equalization, it is not always the case at the preprocessing phase. We find that the computational complexity of regularized inversion is much higher than SQRD, MMSE and RSiegel. This fact is an advantage of the non-linear precoding over its linear counterparts under fast fading scenarios, where preprocessing has to be performed frequently to trace the rapidly changing channel. Due to the expensive division operation to perform normalization, MMSE is of higher computational complexity than SQRD. In contrast, for RSiegel, since $\bar{\rho} = 1.6$ is the worst case, its actual computation load could be lower and even approach SQRD at low SNR region. The algorithm that requires the highest computational complexity is V-BLAST for its multiple calculations of pseudo inversion.

5.3. Parallelism potential

In the previous subsection, the complexity comparison ignores the control overhead of these algorithms, which is also a key concern in their hardware implementation. Actually, inversion calculation is more computationally efficient than the non-linear algorithms in practice. Non-linear calculation involves nondeterministic branches/iterations that cannot be predicted before running, which increases the control overhead and limits the potential parallelism in their implementations. Moreover, the serial nature of the IC phase of THP exacerbates this unfriendly scenario towards hardware further. These facts are two important reasons why there is so limited publications on their implementation. In this subsection, we explore the maximum parallelism that can be utilized to implement these algorithms.

Within the preprocessing phase, although the sorting operations have to be executed in serial, there is still much instruction level and data level parallelism exists in each serial stage, i.e., the Gram-Schmidt process in SQRD, the matrix update in MMSE and the matrix inversion in V-BLAST. Nevertheless, the maximum parallelism in these algorithms decreases from N_T to 1 due to the decreasing dimension among the sequential sorting operations.

Traditionally, the preprocessing phase and the IC phase are performed separately and sequentially. For SQRD and MMSE, in fact, the computations of preprocessing and IC can be partially overlapped. This insight follows from the fact that it is not necessary for the IC phase to know the full reorder information and the complete feedback filter coefficients to process the signal for the first user. The IC phase is able to start whenever necessary coefficients are available before the end of the preprocessing phase. However, similar strategy cannot be applied to V-BLAST. Due to the reverse calculation order, the necessary coefficient to process the first user is only available at the end of V-BLAST preprocessing. For lattice reduction, unfortunately, since the basis reduction process has to be performed column by column in a nondeterministic order, its parallelism is very limited.

6. Discussion

In summary, according to our analysis, we find three interesting facts related to the implementation of THP, namely, 1, the lattice reduction aided algorithm is of better numerical stability but it is more sensitive to imperfect CSI; 2, the computational complexity of some non-linear precoding algorithms (SQRD and RSiegel) is not always greater than their linear pre-equalization counterparts at the preprocessing stage, which is an advantage under fast fading scenarios; 3, the parallelism potential of these THP algorithms diverse significantly, which implies their implementation efforts diverse significantly as well. Overall, SQRD and MMSE are more implementation efficient than V-BLAST and RSiegel.

Moreover, the trade-off space of different THP algorithms is summarized in Table 4. Some general guidelines can be drawn from this table. For example, due to its better BER performance and lower implementation effort, SQRD is more suitable to be employed under fast fading scenarios where the preprocessing has to be performed frequently. Under slow fading scenarios, however, the computation load of preprocessing phase is negligible compared to the IC phase and the transmitter is more likely to know close to perfect CSI. Therefore, it is advisable to employ RSiegel and V-BLAST to achieve high performance in low and high SNR region respectively.

[3] Although it is not considered in this table, it is worthwhile to mention that the channel matrices of adjacent tones might be highly correlated in an OFDM system. This property could be utilized to reduced the computational complexity of preprocessing via the interpolation-based approach [25].

Table 3
Real operations required by various preprocessing schemes of a system with $N_R = N_T = N$ antennas for each OFDM tone.

Type	Mul	Add	Div	SQRT	Sum[1]	N = 4
SQRD	$12N^3 - 4N^2$	$12N^3 - 5N^2$	$N^2 - N$	–	$24N^3 - N^2$	1492
MMSE	$14N^3 - N^2$	$\frac{37}{3}N^3 - 4N^2$	$\frac{2}{3}N^3 - N^2$	1	$\frac{95}{3}N^3 - 13N^2$	1972
RSiegel	$12N^3 - \frac{7}{2}N^2 + 4\bar{\rho}(12N - 2)$	$12N^3 - 5N^2 + 4\bar{\rho}(12N - 2)$	$N^2 - N$	–	$24N^3 - N^2 + 8\bar{\rho}(12N - 2)$	2080 ($\bar{\rho} = 1.6$)[2]
Inverse	$\frac{52}{3}N^3 + N^2$	$\frac{49}{3}N^3 - \frac{11}{2}N^2$	$\frac{2}{3}N^3 + N^2$	$N + 1$	$39N^3 - \frac{15}{2}N^2$	2772
V-BLAST	$\frac{77}{6}N^4 + \frac{32}{3}N^3$	$\frac{151}{12}N^4 + \frac{38}{3}N^3$	$\frac{1}{6}N^4 + 2N^3$	$\frac{1}{2}N^2 + \frac{1}{2}N$	$\frac{107}{4}N^4 + \frac{118}{3}N^3$	10188

[1] Sum of equivalent operations = M+ A + 8xD + 4xSQRT.
[2] $\bar{\rho}$ represents the average time of lattice reduction.

Table 4
Trade-off space of different THP algorithms.

	Performance				Implementation efficiency		
	Low SNR	High SNR		Robustness	Precision	Complexity	Parallelism
		Perfect CSI	Noisy CSI				
SQRD	✓	–	–	–	×	✓	✓
MMSE	×	×	–	–	✓	–	✓
V-BLAST	–	✓	✓	✓	×	×	–
RSiegel	✓	✓	–	×	✓	–	×

"✓", "–" and "×" designate "good", "acceptable" and "poor" respectively.

7. Conclusion

In this paper, we make a comprehensive evaluation of several well-known THP variants in terms of performance and implementation efficiency. From a practical perspective, we find that a wide trade-off space exits to implement these algorithms, which is especially valuable for hardware designers to design cost-efficient VLSI solutions for different communication scenarios in realistic systems.

References

[1] Miridakis N, Vergados DD. A survey on the successive interference cancellation performance for single-antenna and multiple-antenna OFDM systems. IEEE Commun Surv Tutorials 2013;15(1):312–35.
[2] Viswanath P, Tse DNC. Sum capacity of the vector Gaussian broadcast channel and uplink–downlink duality. IEEE Trans Inf Theory 2003;49(8):1912–21.
[3] Schubert M, Boche H. A unifying theory for uplink and downlink multiuser beamforming. IEEE International Zurich Seminar on Access, Transmission, Networking. p. 27-1–6. http://dx.doi.org/10.1109/IZSBC.2002.991770.
[4] Tomlinson M. New automatic equaliser employing modulo arithmetic. Electron Lett 1971;7(5):138–9.
[5] Harashima H, Miyakawa H. Matched-transmission technique for channels with intersymbol interference. IEEE Trans Commun 1972;20(4):774–80.
[6] Costa MH. Writing on dirty paper (corresp.). IEEE Trans Inf Theory 1983;29(3):439–41.
[7] Erez U, ten Brink S. A close-to-capacity dirty paper coding scheme. IEEE Trans Inf Theory 2005;51(10):3417–32.
[8] Erez U, Shamai S, Zamir R. Capacity and lattice strategies for cancelling known interference. International Symposium on Information Theory and Its Application, Honolulu, HI, USA. p. 681–4.
[9] Liu J, Krzymien WA. Improved Tomlinson–Harashima precoding for the downlink of multiple antenna multi-user systems [mobile radio applications]. IEEE Wireless Communications and Networking Conference (WCNC), vol. 1. p. 466–72.
[10] Windpassinger C, Detection and precoding for multiple input multiple output channels, Shaker; 2004.
[11] Kusume K, Joham M, Utschick W, Bauch G. Efficient Tomlinson–Harashima precoding for spatial multiplexing on flat MIMO channel. IEEE International Conference on Communications (ICC), vol. 3. p. 2021–5.
[12] Wübben D, Rinas J, Böhnke R, Kühn V, Kammeyer K. Efficient algorithm for detecting layered space-time codes. The 4th International ITG Conference on Source and Channel Coding (SCC). p. 1–7.
[13] Foschini GJ. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. Bell Labs Tech J 1996;1(2):41–59.
[14] Windpassinger C, Fischer RF, Huber JB. Lattice-reduction-aided broadcast precoding. IEEE Trans Commun 2004;52(12):2057–60.
[15] Shimazaki K, Yoshizawa S, Hatakawa Y, Matsumoto T, Konishi S, Miyanaga Y. A VLSI design of an arrayed pipelined Tomlinson-Harashima precoder for MU-MIMO systems. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). p. 1–4.
[16] Ludwig F, Budweg A, Paul S. FPGA implementation of ZF-THP for MU-MISO-OFDM systems. The 17th International OFDM Workshop (InOWo'12). p. 1–5.
[17] Breit G, Sampath H et al. IEEE 802.11-09/0569r0, TGac Channel Model Addendum Support Material; 2009.
[18] Saleh AA, Valenzuela R. A statistical model for indoor multipath propagation. IEEE J Selected Areas Commun 1987;5(2):128–37.
[19] Iskander CD, A matlab-based object-oriented approach to multipath fading channel simulation, Hi-Tek Multisystems 21; 2008.
[20] Yuan L, Wang G, Ascheid G, Cang L, Xing Z. A flexible low-complexity robust THP approach for MISO downlinks with imperfect CSI. IEEE/CIC International Conference on Communications in China (ICCC).
[21] Stierstorfer C, Fischer RF. Lattice-reduction-aided Tomlinson–Harashima precoding for point-to-multipoint transmission. Int J Electron Commun (AEÜ) 2006;60(4):328–30.
[22] Fischer RF. Complexity-performance trade-off of algorithms for combined lattice reduction and QR decomposition. Int J Electron Commun (AEÜ) 2012;66(11):871–9.
[23] Schumacher L, Dijkstra B. Description of a MATLAB implementation of the indoor MIMO WLAN channel model proposed by the IEEE 802.11 TGn Channel Model Special Committee; 2004.
[24] Lin S, Costello DJ. Error control coding. 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall Inc; 2004.
[25] Borgmann M, Bolcskei H. Interpolation-based efficient matrix inversion for MIMO-OFDM receivers. IEEE Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, vol. 2. IEEE; 2004. p. 1941–7.
[26] Govindu G, Scrofano R, Prasanna VK. A library of parameterizable floating-point cores for FPGAs and their application to scientific computing. International Conference on Engineering Reconfigurable Systems and Algorithms. p. 137–48.
[27] Singh CK, Prasad SH, Balsara PT. VLSI architecture for matrix inversion using modified Gram-Schmidt based QR decomposition. The 20th IEEE International Conference on VLSI Design. p. 836–41. http://dx.doi.org/10.1109/VLSID.2007.177.