

Trade-offs of Performance and Single Chip Implementation of Indoor Wireless Multi-Access Receivers

Ning Zhang, Ada Poon, David Tse and Robert Brodersen

Berkeley Wireless Research Center, Dept. of EECS, University of California Berkeley
and

Sergio Verdu

Princeton University

Abstract—The performance and computational complexity of five multi-access receivers are compared. A methodology is then presented for making area and power estimates of these algorithms for both software programmable DSP and dedicated direct mapped architectures. With this methodology and by using experimental data from previous designs, the feasibility of implementation of the multi-access receivers can be determined

I. INTRODUCTION

CMOS integrated circuit technology has improved to the point that complex receiver systems for wideband systems can now be implemented for portable devices. This increase in complexity will allow receivers to support new algorithms for combating multipath and multiuser interference, as well as yielding increased capacity through use of multi-element arrays. However, the ability for algorithm developers to rise to this challenge, with ever more algorithmic complexity, means that not all algorithms will be possible to integrate now or even in the future, since it is unlikely that the improvement in technology will surpass the historical gains of 1 ½ times every 2 years. This complexity limit comes from the energy and cost (silicon area) constraints of typical portable applications that require wireless connectivity.

A methodology is therefore needed to determine the feasibility of algorithms for implementation and to give guidance to algorithm developers about the present and future capabilities of IC technology. This methodology should evaluate algorithms in terms of not only the traditional measures of performance issues such as throughput and BER, but the energy and area requirements as well.

The design of the baseband digital signal processing section of receivers that implement these algorithms will be investigated, using two basic architectural approaches; a software solution based on a programmable digital signal processor; and a dedicated, hard wired approach that provides the lower bound on area and power. These estimates of power and area will be based on parameters determined from experimental test chips.

To narrow the focus we investigate algorithms appropriate for an indoor wireless environment, where the goal is to provide multimedia access from a basestation to

multiple portable terminals. One of the challenges in design of the algorithms for such a link is to provide enough diversity against multipath fading. For the indoor channel, the delay spreads are usually on the order of 10-100 ns. In order to support a data rate ranging from 1-10 Mbits/sec per user and 10's of users, a combination of time, frequency and/or spatial diversity is required.

We will consider two basic approaches. The first is based on wideband direct-sequence CDMA, and exploits frequency diversity to resolve the multipath components at the receiver and uses multiuser detection to suppress the interference from signals intended for other users. The other approach we will investigate is to employ a narrowband solution by splitting the channel bandwidth into subbands with bandwidths small compared to the channel coherence bandwidth. The flat fading due to the destructive addition of the various paths in each of the narrow band can be combated through the use of spatial diversity using multiple antennas at the transmitter and at the receiver, which can additionally be used to increase the capacity.

II. RECEIVER ALGORITHMS

A significant thrust of communications research over the decades has been done to address the issues involved in multipath mitigation. Among them, time, frequency and spatial diversity [1] are the most common lines of development. Basically, time diversity is to transmit the same piece of information simultaneously in different time slots separated by at least the coherence time. Frequency diversity is achieved by spreading the same information content over a larger spectrum. Similar to multi-carrier transmission, but now the same information is transmitted over the individual carriers, where the carriers are separated by at least the coherence bandwidth of the channel. As spectral spreading improves the temporal resolution, another method to achieve diversity is to resolve the multipath components and hence obtain several delayed replicas of the same transmitted signal. Finally, spatial diversity can be achieved by using multiple antennae. Though these diversity techniques have been extensively studied [2,3], except for the most elementary

techniques, they were thought to be too complex for implementation in energy and cost constrained portable devices.

Recently, it has been shown that the mutual information achieved in a multi-antenna system under flat, independent Rayleigh fading between transmit and receive antenna pairs grows linearly with the number of transmit or receive antennae whichever is smaller [4,5]. This result sets the stage for exploiting an extra degree of freedom in addition to frequency and the time, but with an even greater cost in implementation.

Since temporal spreading incurs time delay, we will investigate five algorithms which exploit the other two diversity techniques, which have increasing levels of performance at the expense of increased complexity. The comparison of the implementation costs of these receivers will be made using state-of-the-art CMOS fabrication technology. The relationship between the computational architecture and the feasibility of their implementation will be presented.

Four of the algorithms compared use wideband synchronous DS-CDMA which employ a single antenna with different multiuser detection schemes and a fifth more complex algorithm which is based on a "narrowband" FDMA strategy using multiple transmit and receive antennae. A symbol rate of 0.8 Msymbols/sec per user with a spreading of 31 for the wideband systems, yields a chip period of 40 ns. An indoor slowly fading environment is assumed which has a Doppler frequency of less than 2 Hz and with a delay spread of 100 ns which yields two resolvable multipaths for the wideband systems. All the analyses are based on 28 users with equal transmit power, 15 dB signal-to-noise ratio and 10^{-5} bit-error-rate. Time multiplexed pilot symbols are used in channel estimation if necessary.

A. Algorithms and Performance Comparisons

A brief introduction of the algorithms is given below. The comparison in terms of computational complexity of the various algorithms is based on the data recovery and the decision-directed channel tracking parts in the downlink.

Algorithm 1 – CDMA and matched filter detection

The received signal over the symbol of interest after the A/D converter is

$$\bar{r} = \sum_{k=1}^K b_k (\alpha_1 \bar{s}_{k1} + \alpha_2 \bar{s}_{k2}) + \bar{n}$$

where \bar{s}_{k1} and \bar{s}_{k2} are the signature sequence and its delayed replica respectively for user k , b_k is the transmitted signal from user k , K is the number of active users and \bar{n} is the Gaussian noise with covariance matrix $\sigma^2 \mathbf{I}$. The term inside the bracket is termed as the effective signature sequence for user k , \bar{s}_k .

In the pilot period, the effective signature sequence of the desired user is estimated using the RAKE architecture,

say \hat{s}_k . While in the data period, the demodulated signal is simply the correlation of the received signal with the estimated effective signature sequence,

$$\hat{b}_k = \hat{s}_k^H \bar{r}$$

Because of the interference from the other users, it can only support BPSK and hence has a lower spectral efficiency which is the trade-off for the simplicity in implementation (seen from Table I).

Algorithm 2 – CDMA and trained adaptive MMSE detection

The MMSE estimate of the transmitted signal is [6]

$$\hat{b}_k = \left(\left(\tilde{S}^H \tilde{S} + \sigma^2 \mathbf{I} \right)^{-1} \tilde{S}^H \bar{r} \right)_k$$

where $\tilde{S} = [\tilde{s}_1 \cdots \tilde{s}_K]$.

The update equation for the linear detector in user k is

$$\bar{d}_k[n] = \bar{d}_k[n-1] - \mu \left(\bar{d}_k[n-1]^H \bar{r}[n] - \hat{b}_k[n] \right)^* \bar{r}[n]$$

and the demodulated signal is

$$\hat{b}_k[n+1] = \bar{d}_k[n]^H \bar{r}[n+1]$$

Algorithm 3 – CDMA and blind adaptive MMSE detection [6]

As in Algorithm 1, the effective signature sequence is first estimated, then in the data period, the linear detector in user k is updated as

$$d_k[n] = d_k[n-1] - \mu \left(r[n] d_k[n-1] - \langle r[n], \hat{s}_k \rangle \hat{s}_k \right)$$

and the demodulated signal is

$$\hat{b}_k[n] = d_k[n]^H \bar{r}[n]$$

Both trained and blind adaptive MMSE can support QPSK but the former needs a training sequence which has more stringent requirements. For example, the same pilot symbols can be used for all the users but different training sequences have to be used for different users for the algorithm to converge. Also, depending on the length of spreading, the training sequence usually lasts longer than the pilot symbols and hence reduces the overall spectral efficiency. On the other hand, the computational complexity of trained adaptive MMSE is only about half of that of blind adaptive MMSE.

Algorithm 4 – CDMA and exact decorrelating detection [6]

The decorrelated signal is

$$\hat{b}_k = \left(\left(\tilde{S}^H \tilde{S} \right)^{-1} \tilde{S}^H \bar{r} \right)_k$$

Since equal power among active users is assumed, subspace optimization can be used to perform the matrix inversion iteratively from the received signals with knowledge on the effective signature sequence of the desired user only. As the MMSE algorithms, QPSK is supported, but since the size of the matrix inversion is up to the number of active users and the length of spreading codes, the computational complexity is 2 to 3 orders of magnitude higher than that of the MMSE algorithms. Therefore, evaluation of algorithms in terms of traditional performance metrics is not enough, power consumption and complexity in implementation should also

be taken into account. It is interesting to estimate the complexity that will be possible in the future if Moore's law for improvement of digital technology continues as it has in the recent past. With this assumption it will take approximately 20 years before the technology improves to the level that this algorithm can be implemented with the power and area that algorithm 3 requires today. Clearly, further algorithmic optimization is thus required if a single chip solution is required [6].

Algorithm 5 - FDMA using multiple antennae

A multi-antenna system with t transmit and r receive antennae in a frequency non-selective and slowly fading channel is assumed, so that the sampled baseband-equivalent channel model is given by

$$\vec{Y} = \mathbf{H}\vec{X} + \vec{Z},$$

where \mathbf{H} is the complex channel matrix with the (i,j) -th element being the random fading between the j -th transmit and i -th receiver antennae and \vec{Z} is the complex, zero-mean Gaussian noise with covariance matrix being $\sigma^2\mathbf{I}$. The symbol transmitted at the j -th antenna is X_j , and the symbol received at the i -th antenna is Y_i . After singular value decomposition (SVD) of the matrix H , it becomes

$$\vec{Y} = \mathbf{U} \Sigma \mathbf{V}^H \vec{X} + \vec{Z}.$$

If \vec{X} is any vector in C^n , then \mathbf{V}^H will be absorbed into \vec{X} . As a result, it is impossible to estimate \mathbf{V} from the received signal at the receiver side. But it is possible to estimate the $\min(r,t)$ dominant components of \mathbf{U} up to a unitary transformation using signal subspace estimation if the data on different transmit antennae are independent. Further if uniform power allocation is used, the $\min(r,t)$ dominant components of both \mathbf{U} and Σ can be estimated with considerable level of confidence. The receiver will send back the sub-sampled versions of the demodulated signals to the transmitter for estimation of \mathbf{V} . The transmitter then projects the transmitted information on the eigen-modes of the channel with more information on the strong modes and less on the weak modes.

In our study, we analyze the case of 4 transmit and 4 receive antennae system using M-PSK (where M is chosen depending on the strength of the eigen-modes estimated). From Table I, increasing the number of antennae at the receiver and the transmitter boosts the data rate at the expense of a modest increase in the digital computational complexity.

B. Computational Complexity

First order, high level estimates of the relative power and area of the implementation of the above algorithms can be obtained simply by counting the number and type of the arithmetic operations. Additionally the data accesses required to retrieve and store the operands must be determined. A simplified form of such a breakdown for the algorithms being investigated here is given in Table I. The relationship of these

operation counts to the resultant area and power is strongly dependent on the architecture used for implementation as will be shown in the following sections. In particular the cost of the data storage and access heavily depends on the memory architecture.

TABLE I
PERFORMANCE AND COMPUTATIONAL COMPLEXITY
COMPARISONS

	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4	Algorithm 5
Performance with 28 users (symbol rate of 0.8Msym/sec per user):					
Data Rate per User	0.8 Mbps	1.6 Mbps	1.6 Mbps	1.6 Mbps	4.8 Mbps
Computational Complexity (operations per sample):					
MULT	124	248	496	228656	736
ALU	124	252	502	237708	800
MEM	248	620	1240	642754	2120

III. IMPLEMENTATION

Given an algorithmic specification, a designer is faced with selecting the computational architecture, as well as determining various parameters such as word length, clock period and supply voltage. The multi-dimensional space offers a large range of possible trade-offs and thus fast high level estimation mechanisms is needed, especially when the goal is to achieve an area efficient, low power integrated circuit implementation.

The results from [9,10] have made it apparent that the most dramatic power reduction stems from optimization at the highest levels of design. In particular, case studies indicate that high-level decisions regarding selection and optimization of algorithms and architectures can improve design performance metrics by many orders of magnitude, while gate and circuit level optimizations typically offer a factor of two or less improvement. This suggests that a top-down approach should be taken to low power design. Specifically, optimization efforts should begin at the algorithm/architecture level.

A. CMOS Power Considerations

Energy consumption in a properly designed CMOS chip is dominated by the dynamic power associated with the charging and discharging of the parasitic capacitances, resulting in an energy per operation cost of $E_{op} = C_{eff}V_{dd}^2$. C_{eff} is the effective capacitance being switched and is different for each operation, where V_{dd} is the supply voltage. Clearly, the most energy efficient solution can be obtained with the minimum supply voltage, but this comes at the expense of increased delay time and thus slower circuits. A supply voltage around 1 V minimizes the energy-delay-products of the critical circuits in a state-of-the-art 0.25 μm technology. Thus, 1 V is chosen for a reasonable energy versus delay trade-off for this technology.

In order to operate at such low voltages and to still be able to accomplish the real time processing rates required for the algorithms, it is necessary to maximally parallelize the

algorithms. This can be accomplished in a number of ways depending on the degree of flexibility that is required. One approach, which has maximum flexibility, is to use a number of software programmed, digital signal processors (DSP's). The overhead of splitting the algorithm into a number of processors can become prohibitive if a large number of processors are required (e.g. >10), but it is useful for comparison purposes to obtain the lower bound of a completely software solution by assuming this overhead is negligible.

Another approach is to use a direct mapping strategy, which maximizes the hardware parallelism by directly implementing the data flow graph of the algorithm in the silicon. While this method is the most energy efficient, because it removes the overhead of instruction fetch and centralized memories, it is the least flexible.

The direct mapping strategy can also exploit another technique for low power design by avoiding wasteful activity associated with over accurate computations. The number of bits used in the arithmetic strongly affects all key parameters of a design, including speed, area and power. It is desirable to minimize the number of bits for power and area optimizations while maintaining sufficient algorithmic accuracy and performance. Simulations on receiver performance with finite word length effects have been performed. The simulations assumed that a 10-bit A/D converter generates the input data and the word lengths for the computation within each algorithm was chosen such that the performance penalty for fixed-point arithmetic is small compared with floating point results [14].

At the algorithm level, functional pipelining, algebraic transformations (e.g., operation and strength reductions) and loop transformations can also be used to increase speed and allow lower supply voltages [7]. For instance, feedback loops often become the bottleneck in an adaptive algorithm. Adding delays in the loop can shorten a critical path and increase concurrency, thus providing more opportunity to maintain throughput at reduced voltage pipelining. However, the modified algorithm needs to be analyzed to insure there is no performance degradation.

B. Software Implementations

Data from a state-of-the-art low power DSP [12] is chosen for the estimation of a software implementation. It is a 1V, 63MHz, 16-bit fixed-point DSP and has a single cycle multiply accumulator. Its on-chip memory includes a 6K x 16b SRAM and 48K x 16b ROM.

C codes describing the algorithms were compiled using performance optimizations. It was found that the energy and cycle based code generations produced very similar code and energy. The running times of programs track each other closely even for processors representing distinct styles [11]. Thus, the shortest sequence requires the lowest energy. This guideline is further justified by the results of instruction level analysis [11, 13], where it is found that the overhead

associated with instruction level control and centralized memory storage dominates the energy consumption.

Performance and power calculations are based on benchmarking of compiled assembly code and adding the contributions of all instructions using instruction level power consumption data from [13]. Table II shows the comparisons of software implementations of the algorithms being considered for a fixed sample rate of 0.8 MHz per user.

It was found that arithmetic instructions only take about 30-50% of the total execution time and power, while memory accesses and control take the majority of both. This indicates that considerable advantages might be achieved in a direct mapped custom design, which has duplicated functional units with dedicated data sources and vastly reduced control. In a direct mapped implementation, each algorithmic function has a corresponding dedicated hardware unit thus providing a high level of optimization of the hardware for a particular algorithm. Moreover, by using optimally sized register files, a direct mapped implementation can avoid the area overhead due to large on-chip memories.

The results from Table II indicate that except possibly for algorithm 1 (simple matched filter) a software solution is not feasible for these data rates.

TABLE II
SOFTWARE IMPLEMENTATION COMPARISONS

	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 5
Parallel Processors	5	11	23	87
Power (mW)	68	152	303	1149
Area (mm ²)	115	253	530	2000

C. Direct Mapped Implementations

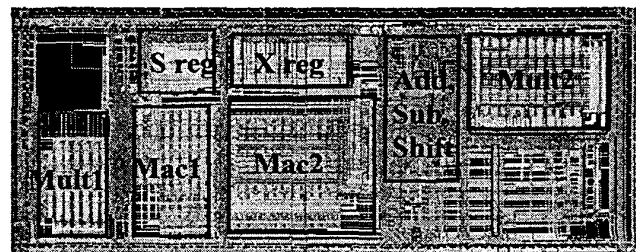
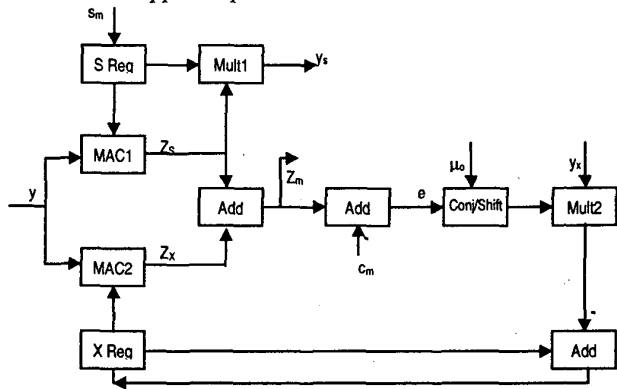


Fig. 1. Direct mapped implementation of an adaptive pilot correlator using algorithm 3.

A direct mapped implementation maps each function of the data flow graph of the algorithm into a dedicated hardware block. For instance, Fig. 1. shows the block diagram and the layout view of the direct mapped implementation of an adaptive pilot correlator using algorithm 3.

For a direct mapped architecture, the power dissipation can be estimated by a weighted sum of the number of operations in the algorithm. A library based approach is adopted which uses a high level model of the average energy consumed by a hardware macro module (which includes clock load and intra-block interconnect) to determine the weighting factor. To estimate the energy consumption at this level the statistics of the applied data are ignored and a random data model is employed. The selection of a macro cell library enables us to somewhat simplify the estimation process since the effect of intra-block routing capacitance is already taken into account during calculation of execution units and register contribution. The area can be estimated by the total area of the required hardware blocks plus certain routing overhead.

A test chip with four adaptive pilot correlators using algorithm 3 has been fabricated in a CMOS 0.25 μm technology. Table III shows the comparisons of hardware implementations of the algorithms being considered, where power and area estimates are based on the library modules used for the test chip. All the implementations operate with a clock rate of 25 MHz from a 1 V supply.

Since algorithm level power estimations are to be used for guiding design decisions and not for making absolute claims about the power consumption, it is only critical that these estimations give accurate relative information. Since algorithm level decisions can result in orders of magnitude difference, this approach provides meaningful power predictions to guide high level decisions.

TABLE III
HARDWARE IMPLEMENTATION COMPARISONS

	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 5
Word length	8	12	12	16
Complex mac/mult units	1	2	4	9
Add/sub/shift units	0	6	10	57
Registers	68	147	225	200
Power (mW)	0.4	1.6	3.1	8.0
Area (mm ²)	0.6	2	3	10

IV. CONCLUSION

Often different algorithms are available for the same wireless communication application, but they achieve quite different system performance and have quite different complexities. The methodology proposed advocates a top-down approach to design optimization. At the algorithm/architecture level, system designers can explore and evaluate the performance of algorithms as well as their

suitability for implementation in highly integrated CMOS implementations.

There are several orders of magnitude difference in power consumption and area between software and highly optimized hardware implementations of the same algorithm, which suggests that computational architecture needs to be considered together with algorithm choice for criteria such as energy efficiency, computational capability and algorithmic flexibility. A heterogeneous hardware architecture which trades off these various criteria will more optimally address the requirement of future wireless systems.

Estimates at the algorithm level for hardware implementation can be used for relative evaluation of different designs. Accurately predicting power and performance based solely on algorithmic criteria is a difficult problem. By taking a pre-characterized library-based approach, we are able to give estimates a firm basis.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of DARPA and the industrial members of the Berkeley Wireless Research Center, Cadence, Ericsson, Hewlett Packard, Intel, Lucent, ST Microelectronics, and Texas Instruments.

REFERENCES

- [1] John G. Proakis, *Digital Communications*, McGraw Hill, New York, 3rd Ed., 1995.
- [2] R. D. Gitlin, J. Salz and J. H. Winters, "The Impact of Antenna Diversity on the Capacity of Wireless Communication Systems," *IEEE Trans. Comm.*, Vol. 42, No. 4, pp. 1740-1751, 1994.
- [3] Siavash M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," *IEEE J. Select. Areas Comm.*, Vol. 16, No. 8, pp. 1451-1458, October 1998.
- [4] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels," AT&T-Bell Labs Internal Tech. Memo., June 1995.
- [5] Gerard J. Foschini, "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-Element Antennas," *Bell Labs Technical Journal*, Autumn 1996.
- [6] Sergio Verdu, *Multuser Detection*, Cambridge University Press, Cambridge, UK, 1998.
- [7] A. P. Chandrakasan, M. Potkonjak, R. Mehra, J. Rabaey, and R. W. Brodersen, "Optimizing Power Using Transformations," *Transactions on CAD*, January 1995.
- [8] "Low Power Design Methodologies," edited by J. M. Rabaey and M. Pedram, Kluwer Academic Publishers.
- [9] P. Landman, R. Mehra, and J. Rabaey, "An Integrated CAD Environment for Low-Power Design," *IEEE Design and Test*, summer, 1996.
- [10] A. Raghunathan, N. K. Jha, and S. Dey, "High-Level Power Analysis and Optimization," Kluwer Academic Publishers.
- [11] V. Tiwari, S. Malik, A. Wolfe, and M. Lee, "Instruction Level Power Analysis and Optimization of Software," *Journal of VLSI Signal Processing*, pp. 1-18, 1996.
- [12] W. Lee, et al., "A 1-V Programmable DSP for Wireless Communications," *IEEE Journal of Solid-State Circuits*, November, pp. 1766-1777, 1997.
- [13] C. Turner, "Calculation of TMS320LC54x Power Dissipation," Technical Application Report SPRA164, Texas Instruments, 1997.
- [14] N. Zhang, C. Teuscher, H. Lee, and R. Brodersen, "Architecture Implementation Issues in a Wideband Receiver Using Multuser Detection," *Proceeding of the Thirty-Sixth Annual Allerton Conference*, September 1998.