A 70GOPS, 34mW Multi-Carrier MIMO Chip in 3.5mm^2

Dejan Markovic, Robert W. Brodersen, Borivoje Nikolic

Berkeley Wireless Research Center, University of California, Berkeley, USA

Abstract

An ASIC realization of the MIMO baseband processing for a multi-antenna WLAN is described. The chip implements a 4×4 adaptive singular value decomposition (SVD) algorithm with combined power and area minimization achieving a power efficiency of 2.1GOPS/mW in just 3.5mm^2 in a 90nm CMOS. The computational throughput of 70GOPS is implemented with 0.5M gates at a 100MHz clock and 385mV supply, dissipating 34mW of power. With optimal channel conditions the algorithm implemented can deliver up to 250Mbps over 16MHz band.

Introduction

To satisfy a growing need for higher capacity and extended range, OFDM-based WLAN devices [1] are moving to using MIMO algorithms as being defined in the 802.11n standard. Ideally, a complete MIMO channel decomposition would be performed independently in each of the narrowband sub-carriers, which would produce a computational increase that outstrips the improvements provided by scaling of technology alone. To address this challenge we present a methodology for power *and* area optimal design of multi-dimensional algorithms operating over many parallel sub-carriers. We demonstrate the integration of 16 parallel 4×4 SVD decompositions which requires 70GOPS (12 bit equivalent add operations at 100MHz) in 3.5mm^2 with 34mW of power in a standard-V$_T$ 90nm CMOS technology.

Architecture for Adaptive SVD

Decoding of a MIMO channel requires a matrix inversion which can be done in block-form using an SVD, shown in Fig. 1. The channel matrix H is decomposed into matrices U, Σ, and V, where U and V are unitary and Σ is a diagonal matrix. With partial channel knowledge V at the transmitter Tx and projection of the received vector $\underline{y}$ onto space of U^+, we can effectively orthogonalize the channel between $\underline{x}$' and $\underline{y}$' and utilize spatial channels to send independent data streams across Tx antennas.

We implement an adaptive SVD algorithm described in [2] that reduces matrix operations to vector operations at the expense of extra square rooting and division operations. The algorithm performs LMS-based estimation of eigen-pairs ($\underline{u}_i$, λ_i), and deflation for successive rank reduction as shown in Fig. 2. Square rooting and division are implemented using an iterative Newton-Rhapson method achieving single-iteration convergence under slow channel variations. The step-size of the UΣ LMS is adaptively adjusted as $0.05/\lambda_i$. The narrow-band algorithm of [2] is extended to multiple carriers to use 16MHz of bandwidth. By using data-stream interleaving over antennas/vectors and sub-carriers, Fig. 3a, the area and routing complexity are reduced. Vectoring and re-organization in time domain allows for folding over antennas for further area reduction, Fig. 3b. Memory inside the *Alg** block (~64Kb) is directly realized in pipeline registers.

Power/Area Optimization

The minimum in area and energy is achieved by using the architectural/circuit techniques that yield largest area or energy savings for a given decrease of throughput, starting from a direct-mapped architecture at a nominal supply. This process is repeated until all the techniques are balanced, [3]. The clock speed required to implement the algorithm in a direct-mapped parallel architecture is significantly lower than what is available in current technology. The excess speed is used to reduce power and area through concurrent architecture and circuit-level optimizations. Since 1MHz wide sub-channels require 1MS/s to process the data, a direct-mapped architecture would need 1MHz baseline clock if the *Alg** operation can be realized with one cycle of latency. The critical loop of the algorithm has 6 real multiplies, 11 adds, and 2 mux operations. While this is feasible within 1μs, even at a reduced V_{DD}, area minimization dictates a streamed architecture with folding, resulting in a 64MHz clock rate (1MHz × 16 sub-carriers × 4 antennas).

High energy efficiency is reached through aggressive voltage scaling and gate sizing. The voltage is scaled down to 0.4V, without compromising static VTC characteristic of logic gates, Fig. 4a. Designing for 64MHz at 0.4V in the SS corner translates into 512MHz timing constraint for logic synthesis under the worst case model (0.9V, 125C). Due to limitation of synthesis tool, we balance the tradeoffs with respect to logic sizing (W) and supply voltage (V_{DD}) [3] sequentially. First, we constrain the logic synthesis with a 20% slack at the nominal supply and perform sizing optimization, as illustrated in Fig. 4b. The supply voltage is then scaled down to 0.4V to balance the sensitivities to 0.8, resulting in the design optimized for target speed of 64MHz.

Architectural optimization is done early in the design, using a timed data-flow model (Simulink), based on energy-area-delay estimates of building blocks such as add or multiply. To ensure top-level optimality, the building blocks are characterized in latency vs. cycle time space, such that the appropriate latency is assigned to each block. This modularity reduces loop retiming to simple feed-forward problem and provides initial hardware estimates for power and area at the Simulink level. Word-length reduction is also performed using an automatic Simulink-based optimizer that minimizes hardware area subject to quantization error at the output, [4], resulting in a 30% reduction in area and energy compared to a 16-bit design.

Techniques for energy and area minimization are illustrated in Fig. 5 and Table I. Starting from a design with optimal V_{DD} and W, interleaving and folding reduce the area without an energy increase as shown in Fig. 5. Both techniques introduce pipeline registers around feedback loops, but also speed-up the clock to maintain throughput, thus coming back to the original point on Energy-Delay line of pipeline logic blocks. The area of the logic blocks is shared by sub-carriers and also over antennas, leading to a 36× reduction in chip area due to interleaving and folding combined, Table I.

Measured Results

Simulink environment is also used for chip testing. The design is programmed onto a Xilinx Virtex-II Pro FPGA, which feeds data into the chip and verifies the chip outputs in real-time. An adaptive 4×4 eigen-mode decomposition is shown in Fig. 6. After the reset, the chip is trained with a stream of identity matrices. In blind tracking mode, PSK modulated data is sent over the antennas with constellations varying according to the estimated SNR, achieving up to 250Mbps over 16 sub-carriers.

Measured 100MHz operation, corresponding to TT corner, is obtained at 385mV to 425mV over 9 dies, with a 2× variation in leakage power. The leakage and clocking power are 12% and 30% of the total power, respectively. Area and power of functional blocks at 100MHz are given in Table II. Die features are summarized in Fig. 7.

1-4244-0006-6/06/$20.00 (c) 2006 IEEE

Acknowledgments

The authors would like to acknowledge funding support from C2S2 under MARCO contract 2003-CT-888 and BWRC member companies, STMicroelectronics for chip fabrication, P. Droz and H. Chen for help with chip testing.

References

[1] J. Thomson *et al.*, “An integrated 802.11a baseband and MAC processor,” in *Proc. ISSCC*, Feb 2002, pp. 126-127.

[2] A. Poon, D. Tse, and R.W. Brodersen, “An adaptive multiple-antenna transceiver for slowly flat-fading channels,” *IEEE Trans. Comm*, vol. 51, no. 13, pp. 1820-1827, Nov 2003.

[3] D. Markovic, V. Stojanovic, B. Nikolic, M.A. Horowitz, and R.W. Brodersen, "Methods for True Energy-Performance Optimization," *IEEE JSSC*, vol. 39, no. 8, pp. 1282-1293, Aug 2004.

[4] C. Shi and R.W. Brodersen, "Automated Fixed-point Data-type Optimization Tool for Signal Processing and Communication Systems," in *Proc. IEEE DAC*, pp. 478-483, June 2004.

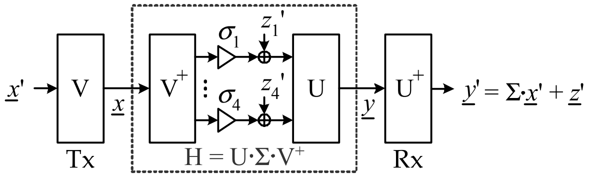

Fig. 1. Decoupling of MIMO channel through SVD.

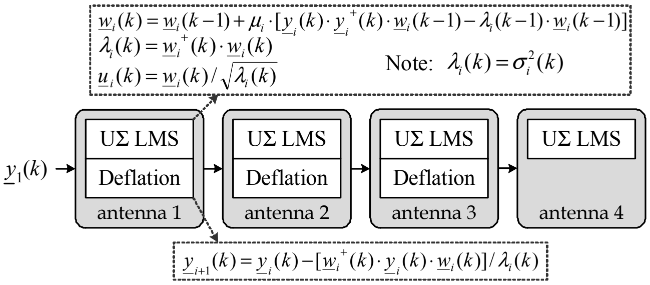
$\underline{w}_i(k) = \underline{w}_i(k-1) + \mu_i \cdot [\underline{y}_i(k) \cdot \underline{y}_i^+(k) \cdot \underline{w}_i(k-1) - \lambda_i(k-1) \cdot \underline{w}_i(k-1)]$
$\lambda_i(k) = \underline{w}_i^+(k) \cdot \underline{w}_i(k)$
$\underline{u}_i(k) = \underline{w}_i(k) / \sqrt{\lambda_i(k)}$
Note: $\lambda_i(k) = \sigma_i^2(k)$
$\underline{y}_1(k)$
UΣ LMS
Deflation
antenna 1
UΣ LMS
Deflation
antenna 2
UΣ LMS
Deflation
antenna 3
UΣ LMS
antenna 4
$\underline{y}_{i+1}(k) = \underline{y}_i(k) - [\underline{w}_i^+(k) \cdot \underline{y}_i(k) \cdot \underline{w}_i(k)] / \lambda_i(k)$

Fig. 2. Adaptive 4×4 eigen-mode decomposition algorithm.

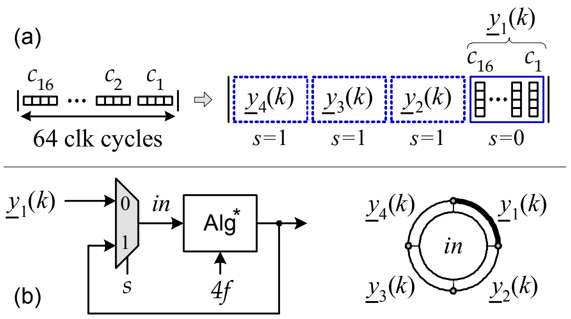

Fig. 3. (a) Vectoring and time-serial ordering of interleaved data. (b) Folding of Alg operation (UΣ LMS, Deflation).

TABLE I. SUMMARY OF MAIN OPTIMIZATION TECHNIQUES

Opt technique	Area reduction	Energy reduction
Word-length opt	30%	30%
Gate sizing (W)	20%	40%
V_{DD} scaling	N/A	7×
Interleaving/folding	13.8× / 2.6×	–2%

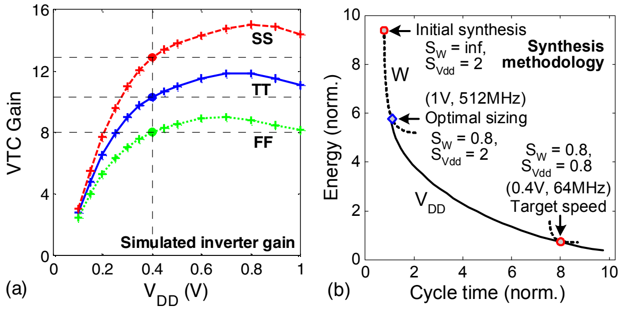

Fig. 4. (a) Inverter VTC does not degrade at V_{DD}=0.4V. (b) Energy minimization via gate sizing and V_{DD} reduction.

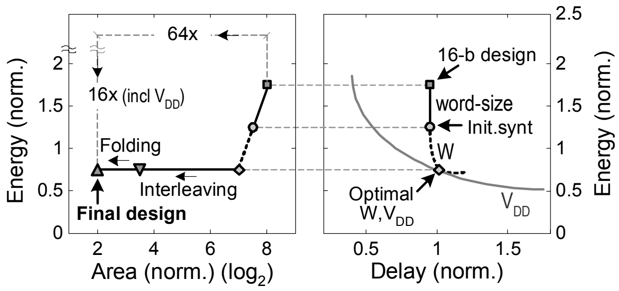

Fig. 5. Area-Energy-Delay tradeoff for constant throughput.

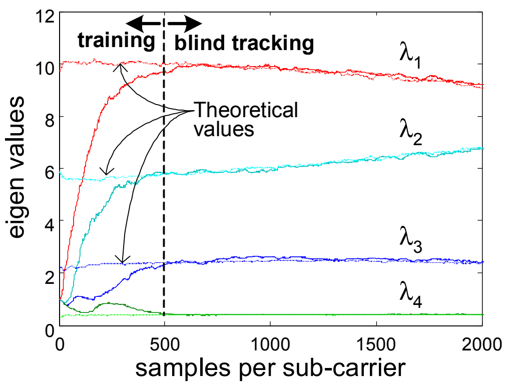

Fig. 6. Measured tracking of eigen-modes.

TABLE II. AREA AND POWER OF FUNCTIONAL BLOCKS

100MHz	Power (mW)	Area (mm^2)	GOPS
UΣ LMS	20	2.31	42.6
Deflation	14	1.19	27.4

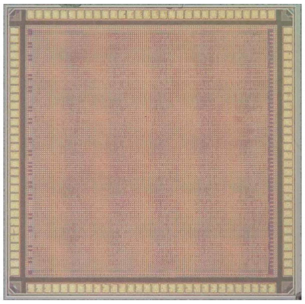

TABLE III: CHIP FEATURES

Technology	90nm CMOS
Core area	1.9 × 1.9 mm
Die area	2.3 × 2.3 mm
Pad count	120
IO/core V_{DD}	1V / 0.4V
Cell count	420,304
Frequency	100 MHz
P (act/leak)	30mW / 4mW
Efficiency	2.1GOPS/mW

Fig. 7. Die photo of 4x4 SVD, summary of chip features.