LOW-POWER DESIGN METHODOLOGY FOR DSP SYSTEMS USING MULTIRATE APPROACH

An-Yeu Wu† *K. J. Ray Liu* *Zhongying Zhang*
Kazuo Nakajima *Arun Raghupathy*

† AT&T Bell Laboratories, Murray Hill, NJ 07974, USA
Electrical Engineering Department, University of Maryland at College Park, MD 20742, USA

ABSTRACT

We present a low-power design methodology based on the multirate approach for DSP systems. Since the data rate in the resulting multirate implementation is M-times slower (where M is a positive integer) than the original data rate while maintaining the same throughput rate, we can apply this feature to either the low-power implementation, or the speed-up of the DSP systems. This design methodology provides VLSI designers a systematic way to design low-power DSP systems at the algorithmic/architectural level. The proposed low-power multirate design scheme is verified by the implementation of two FIR VLSI chips with different architectures: One is the normal pipelined design and the other is the multirate design with downsampling rate equal to two. The experimental results show that the multirate FIR chip consumes only 21% power of the normal FIR chip given the same data throughput rate.

1. INTRODUCTION

Due to the limited power-supply capability of current battery technology, the state-of-the-art personal communication services (PCS) devices call for low-power VLSI design at all aspects (algorithmic, architectural, circuit, logic, and device levels) to minimize the total power consumption, while maintaining the system performance such as data throughput rate [1]. In this paper, we present a systematic approach for the low-power design of a general linear time-invariant (LTI) FIR/IIR system based on the multirate approach. In general, the direct implementation of the system transfer function $H(z)$ (see Fig. 1(a)) has the constraint that the speed of the processing elements must be as fast as the input data rate. As a result, it cannot compensate the speed penalty under low supply voltage [1]. On the other hand, the multirate system in Fig. 1(b) requires only low-speed processing elements at one-third of the original clock rate to maintain the same throughput. Therefore, the processing elements can be operated at a lower supply voltage to reduce the power dissipation and the data throughput rate is not degraded by the lowered voltage. As a result, the multirate implementation provides a direct and efficient way to compensate the speed penalty in low-power designs at the algorithmic/architectural level [2]. Based on this design concept, we present a design methodology for the design of low-power DSP systems. The users can simply follow the design steps to convert a speed-demanding system function into an equivalent multirate transfer function. Since the data processing rate in the multirate implementation is M-times slower than the input data rate, we can apply this feature to either the low-power implementation of the FIR/IIR system, or to the processing speed-up of the system.

*THIS WORK WAS SUPPORTED IN PART BY THE ONR GRANT N00014-93-10566 AND THE NSF GRANT MIP9457397.

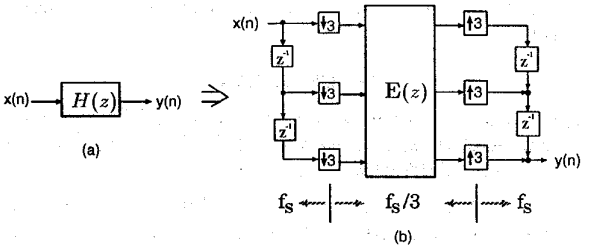

Figure 1. (a) An LTI FIR/IIR system. (b) Its equivalent multirate implementation, where f_s is the data sampling rate.

We also verify the effectiveness of the proposed multirate low-power design by the implementation of two FIR VLSI chips with different architectures. One is the normal pipelined design and the other is the multirate design with downsampling rate equal to two. We implement both chips using the same CAD synthesis tool and the same VLSI technology. The only difference lies in the architectural design. Therefore, the effectiveness of the algorithm-based low-power design can be observed. The selected FIR system is a Quadrature Mirror Filter (QMF) which is widely used in the applications of image compression and subband coding [3]. The simulation results show that by trading 50% more silicon area, we can save up to 71% of the total power consumption without sacrificing the data throughput rate. This observation is later verified by the testing of the two FIR chips.

2. MULTIRATE DESIGN METHODOLOGY

In what follows, we present the design methodology to derive the multirate LTI system of Fig. 1. Without loss of generality, we assume that $M = 3$ in our derivation. The design procedure can be easily extended for an arbitrary M.

The Design Procedure

Given an LTI FIR/IIR system $H(z)$ with order N and decimation factor M, the design procedure is as follows (see Fig. 2).

Step(a): Insert $M - 1$ unit delays after the transfer function $H(z)$.

Step(b): Replace the delay element with its equivalent "delay chain perfect reconstruction system."

0-7803-3073-0/96/$5.00 ©1996 IEEE

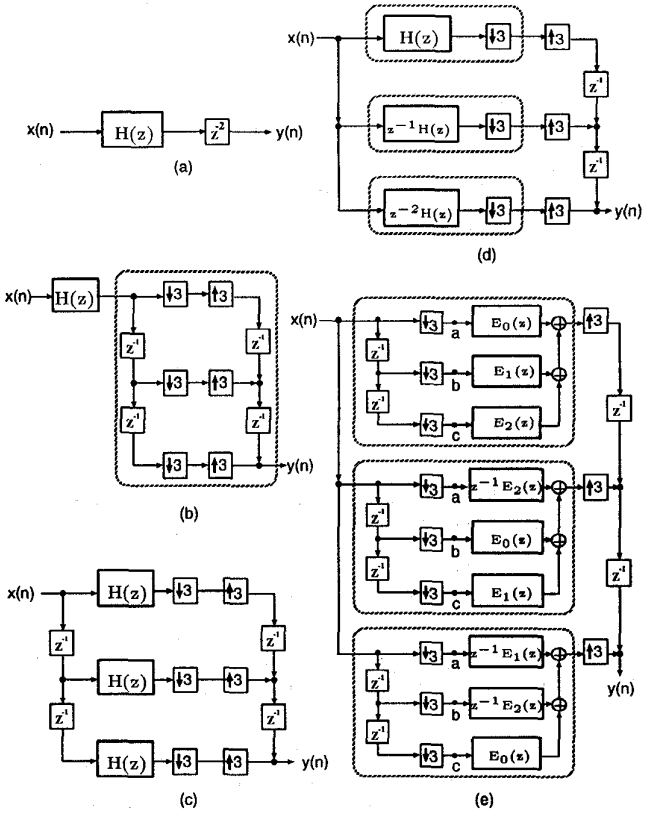

Figure 2. Design procedure for the multirate LTI system.

Step(c): Move $H(z)$ to the right till reaching the decimation operators.

Step(d): Merge the delay elements with the transfer functions. Group the resulting new transfer functions ($H(z)$, $z^{-1}H(z)$, and $z^{-2}H(z)$) with their associated decimation operators.

Step(e): Replace each decimation circuit in the circle shown in Fig. 2(d) with its *polyphase implementation* [4]. Then we have Fig. 2(e), where $E_i(z)$, $i = 0, 1, \ldots, M-1$, are the *Type I polyphase components* of $H(z)$.

Step(f): Note that the data inputs at points designated by **a** are the same, and so are those at points **b** and **c**. After merging the common data paths in Fig. 2(e), we obtain Fig. 2(f) in which

$$\mathbf{E}(z) \triangleq \begin{bmatrix} E_0(z) & E_1(z) & E_2(z) \\ z^{-1}E_2(z) & E_0(z) & E_1(z) \\ z^{-1}E_1(z) & z^{-1}E_2(z) & E_0(z) \end{bmatrix}. \quad (1)$$

The general form of $\mathbf{E}(z)$ with an arbitrary decimation factor M can be shown to be

$$\mathbf{E}(z) = \begin{bmatrix} E_0(z) & E_1(z) & \cdots & E_{M-1}(z) \\ z^{-1}E_{M-1}(z) & E_0(z) & \cdots & E_{M-2}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z^{-1}E_1(z) & z^{-1}E_2(z) & \cdots & E_0(z) \end{bmatrix}, \quad (2)$$

which is also known as the *pseudocirculant matrix* in the context of alias-free QMF filter banks [4].

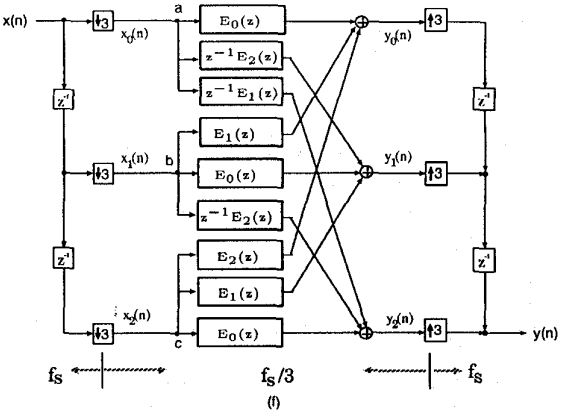

Figure 2. (cont.)

The design procedure described in Fig. 2 provide a systematic way to design a low-power FIR/IIR system. In the implementation of the FIR system, each $E_i(z)$ in Fig. 2(f) represents a subfilter of order N/M, it can be shown that the total hardware complexity to realize the multirate FIR system is MN multipliers and $(MN + M^2)$ adders. Basically, we pay a linear increase of hardware overhead in exchange for the advantage of an M-times slower processing speed.

In the multirate IIR system design, we first find out the polyphase components $E_i'(z)$'s, $i = 0, 1, \ldots, M-1$, of the given IIR function $H'(z)$. After replacing each $E_i(z)$ in Fig. 2 with its corresponding $E_i'(z)$, for $i = 0, 1, \ldots, M-1$, we can apply the aforementioned design methodology to convert $H'(z)$ into its equivalent multirate transfer function. Note that the complexity of each $E_i'(z)$ can be as high as that of the original transfer function $H'(z)$. Hence, we may pay up to $O(M^2)$ hardware overhead for the implementation of the multirate IIR filter.

2.1. Diagonalization of Pseudocirculant Matrix

Although the multirate implementation of Fig. 2(f) can be readily applied to low-power design, the global communication of this structure is not desirable in the VLSI implementation. Therefore, we want to diagonalize the pseudocirculant matrix of Eq. (2) so as to eliminate global communication in the multirate implementation.

There are two ways to diagonalize $\mathbf{E}(z)$ of Eq. (2). One is to use the DFT approach [5]. The DFT approach requires complex-number operations for the filtering operations. In addition, it still has global communications in the DFT/IDFT networks. The second diagonalization approach is based on polynomial convolution techniques [6]. As an example, for the case of $M = 2$, Eq. (2) can be rewritten as

$$\begin{bmatrix} E_0(z) & E_1(z) \\ z^{-1}E_1(z) & E_0(z) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} E_0(z) & 0 & 0 \\ 0 & E_0(z)+E_1(z) & 0 \\ 0 & 0 & -E_1(z) \end{bmatrix} \underbrace{\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ z^{-1} & -1 \end{bmatrix}}_{\mathbf{A}}. \quad (3)$$

The corresponding structure is depicted in Fig. 3. This diagonalization approach involves only real-number operations to process the decimated sequences. However, as M increases, the derivation becomes complicated and the resulting architecture is highly irregular (as opposed to the

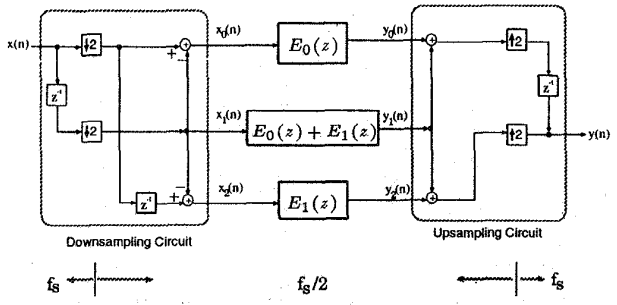

Figure 3. Multirate FIR/IIR architecture with $M = 2$.

DFT approach) [6]. In the verification part, we will use the multirate FIR structure in Fig. 3 for our low-power FIR chip design.

2.2. Power Estimation for The Multirate FIR Architecture

Before we proceed to the chip design, we consider the power dissipation of the multirate design. For the normal FIR architecture, it requires N multipliers and N adders. For the low-power multirate FIR architecture depicted in Fig. 3, $3N/2$ multipliers and $3N/2$ adders are required. Since all operators are running at a 2-times slower clock rate, V'_{dd} can be as low as $3.1V$ [2]. Provided that the capacitance due to the multipliers is dominant in the circuit and is roughly proportional to the number of multipliers and adders, the power consumption of the multirate FIR design can be estimated as

$$(\frac{\frac{3N}{2}}{N}C_{eff})(\frac{3.1V}{5V})^2(\frac{1}{2}f) \approx 0.29P_0, \qquad (4)$$

where P_0 denotes the power consumption of the original system. Although the multirate architecture requires about 50% hardware overhead, it consumes only 29% power of the original pipelined design. Basically, we trade hardware complexity for low-power consumption.

Another attractive application of the multirate design is in the very high-speed filtering. If we do not lower down the supply voltage to save chip power consumption, the multirate FIR structure of Fig. 3 can process data at a rate which is twice as fast as the maximum speed of the processing elements. We will also verify this in the next section.

3. VERIFICATION OF THE MULTIRATE LOW-POWER DESIGN

We now verify the power saving of the multirate low-power FIR structure. The selected FIR design is a Quadrature Mirror Filter (QMF) which is widely used in the image compression and subband coding [3].

3.1. Selection of System Parameters

First, we consider the design and implementation issues of the QMF. The implementation of the QMF filter requires a large number of multipliers. In order to save the chip area, we choose the QMF design with power-of-two (POT) coefficients [7] for our chip implementation. For the purpose of further lowering the hardware complexity, we modify the QMF design of [7] by truncating some boundary tap coefficients and by dropping some relatively small components in each coefficient. Shown below are the QMF coefficients

Filter type	Filter length	PSNR (dB)	Coeffi-cient type	Fixed-point adders
Filter 32D in [3]	32	44.8	Float	N/A
Filter in [7]	32	38.8	POT	84
Filter of Eq. (5)	22	37.0	POT	36

Table 1. PSNR results for different QMF's.

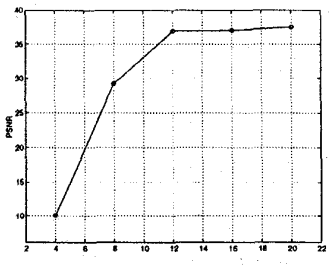

Figure 4. PSNR results for the modified QMF as a function of system wordlength B.

used in our FIR chip design:

$$\begin{aligned}
&h(10) = h(11) = 2^{-2} + 2^{-4} + 2^{-6}, && h(5) = h(16) = 2^{-7} + 2^{-8},\\
&h(9) = h(12) = 2^{-4} + 2^{-5}, && h(4) = h(17) = -2^{-6} - 2^{-7},\\
&h(8) = h(13) = -2^{-4} - 2^{-7}, && h(3) = h(18) = -2^{-8},\\
&h(7) = h(14) = -2^{-5}, && h(2) = h(19) = 2^{-6},\\
&h(6) = h(15) = 2^{-5} + 2^{-7}, && h(1) = h(20) = 0,\\
& && h(0) = h(21) = -2^{-7}.
\end{aligned} \qquad (5)$$

To verify the performance of this modified QMF, we carried out simulations by passing the LENA image through the subband coding structure (see [7, Fig. 5]). Table 1 lists the peak SNR (PSNR) results. Compared with the original design of [7], our modified design suffers from little degradation in PSNR but with much less hardware complexity.

Next we want to determine the system wordlength to be used in our chip design. Since the wordlength would directly affect the resulting chip area as well as the total number of switching events in the logic circuits, it is important to determine the minimum wordlength without degrading the PSNR performance. We conducted computer simulations by feeding the LENA image into the subband coding structure under fixed-point arithmetic. The results are shown in Fig. 4. Since the PSNR curve saturates around $B = 12$, we use the wordlength of 12 bits in our design.

3.2. Chip Design

Having decided the system specifications of our design, we use PARTHENON [8] to design/synthesize both the normal and multirate FIR filters. The resulting chip layouts are shown in Fig. 5. In the multirate design, the upper right module realizes the upsampling and downsampling circuits of Fig. 3. The other three modules realize the three $N/2$-tap FIR filters, $E_0(z)$, $E_1(z)$, and $E_0(z) + E_1(z)$ of Fig. 3, at the operating frequency of $f_s/2$. Their output signals are sent back to the up/downsampling module to reconstruct the filtering output $y(n)$ running at f_s. The chip area of the multirate design is about 50% more than that of the normal design as we expected. Therefore, our estimation of the effective capacitance in Eq. (4) is very accurate.

In order to see the effect of supply voltage on the speed of the FIR design, we conducted SPICE simulation for the critical path of the FIR structure. The simulation results depicted in Fig. 6 show that the propagation delay is approximately doubled as the supply voltage reduces form 5V

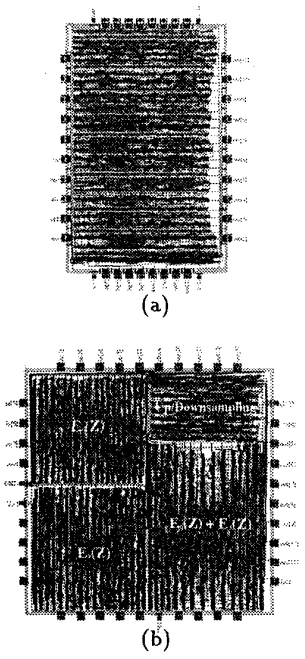

Figure 5. (a) Final layout of the normal FIR filter (dimension = $4400 \times 6600\lambda^2$). (b) Final layout of the multirate FIR filter (dimension = $6500 \times 6600\lambda^2$).

to 3.1V. This is consistent with the results presented in [1]. Since the delay in the critical path generally determines the maximum clock rate of the chip, we can predict that the performance of the filtering operations will be degraded by 50% under the 3.1V supply voltage. Nevertheless, the data throughput rate of the multirate FIR will not be affected by such a speed penalty since the slowed-down devices are in the $f_s/2$ region (see Fig. 3). The I/O data rate will remain at f_s which is the same as the normal FIR design operated at 5V.

3.3. Testing Results

These two chip designs have been fabricated by MOSIS using 2μ double metal CMOS technology. The chips were tested under the same test environments: the same input data sequence and the same test equipment (HP 82000 IC evaluation system).

The measurement of the power dissipation and maximum

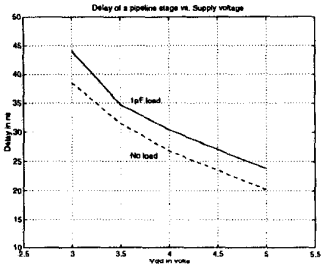

Figure 6. Timing analysis for one pipelined stage in the FIR design.

	Voltage (V)	Current (mA)	Max speed (MHz)	Power (mW)
Normal FIR	5.0	98.5	20.0	492.5
Multirate FIR	5.0	141.8	31.3	709.0
	4.5	106.4	26.0	478.8
	4.0	76.9	24.4	307.6
	3.5	42.4	21.7	148.4
	3.2	32.6	20.0	104.3

Table 2. Testing results of the normal and multirate FIR chips.

speed of both chips are listed in Table 2. As we can see, at the same 20 MHz data rate, the multirate FIR chip can operate at a lower supply voltage of 3.2V and consumes only 21% power of the normal FIR chip. These results agree with our arguments for the supply voltage and power consumption of the low-power design. Under the normal 5V supply voltage, the multirate FIR chip can operate at 31.3 MHz, which is 56% faster than the normal FIR chip. This speed is slower than our theoretical estimation. The main reason is that the maximum operating frequency of the 2 μ technology by MOSIS is only 35 MHz. Hence, the maximum speed of the multirate FIR chip cannot reach 40 MHz. However, it is very close to the speed limit of the 2 μ technology.

4. CONCLUSIONS

In this paper, we presented an algorithm-based low-power design methodology for LTI systems using the multirate approach. To verify the effectiveness of the multirate architecture in the application of low-power design, we have also implemented two FIR filters (normal and multirate designs) onto VLSI chips so as to compare the real power savings. The proposed methodology not only provides a systematic way to derive low-power LTI systems at architectural level, it also can be applied to very high-speed filtering, in which only low-speed/inexpensive operators are required.

REFERENCES

[1] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, April 1992.

[2] A.-Y. Wu and K. J. R. Liu, "Algorithm-based low-power transform coding architectures," in *Proc. IEEE Int'l Conf. Acoust. Speech, Signal Processing*, (Detroit), May 1995, pp. 3267–3270.

[3] J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks," *Proc. IEEE Int'l Conf. Acoust. Speech, Signal Processing*, 1980, pp. 291–294.

[4] P. P. Vaidyanathan, *Multirate systems and filter banks*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[5] H. K. Kwan and M. T. Tsim, "High speed 1-D FIR digital filtering architectures using polynomial convolution," in *Proc. IEEE Int'l Conf. Acoust. Speech, Signal Processing*, (Dallas, TX), April 1987, pp. 1863–1866.

[6] Z.-J. Mou and P. Duhamel, "Short-length FIR filters and their use in fast nonrecursive filtering," *IEEE Trans. Signal Processing*, vol. 39, pp. 1322–1332, June 1991.

[7] C.-K. Chen and J.-H. Lee, "Design of linear-phase quadrature mirror filters with power-of-two coefficients," *IEEE Trans. Circuits Syst. II*, vol. 41, pp. 445–456, July 1994.

[8] Y. Nakamura, K. Oguri, and A. Nagoya, "Synthesis from pure behavioral descriptions," in *High-level VLSI Synthesis* (R. Camposano and W. Wolf, eds.), Kluwer Academic Publishers, 1991, pp. 205–229.