Design and Implementation of Filters for Low Power Applications

Anantha Chandrakasan Alan Oppenheim (co-investigators)
In Collaboration with: Vadim Gutnik, Nitin Kasturi, Jeff Ludwig, Hamid Nawab, Duke Xanthopoulos
Department of EECS, Massachusetts Institute of Technology, Cambridge

ABSTRACT

This report summarizes our research on low-power signal processing under the ARL Federated Laboratory program. Several techniques are presented to implement DSP filters with the lowest possible power consumption. Power in CMOS circuits is consumed when capacitance is switched and can be minimized through logic optimization, exploitation of spatial and temporal signal correlations, and operation reduction. Supply voltage scaling exploiting concurrency and data dependencies is also key to power minimization. A system level approach to low-power design can result in orders of magnitude power reduction.

1. Introduction

The weight and size of batteries, which are major factors in wireless devices (e.g. remote sensors), are primarily determined by the power dissipation of electronic circuits. Until recently, there has not been a major focus on a design methodology which directly addresses power reduction, with the focus rather on ever faster clock rates and logic speeds. The emphasis of this work is on minimizing the power consumption in DSP applications, which have two important attributes that make then different from general purpose computation: throughput constrained computing and correlation in the processed data.

The throughput constraint implies that once the required sample rate is met by the hardware, there is no advantage in making the computation faster (unlike general purpose computation where is goal is to compute as fast as possible). This enables a variety of voltage scaling techniques that lower power while maintaining a constant throughput. For example, one approach (termed architectural voltage scaling) involves using parallel structures to scale the clock rates and supply voltages [1]. An order of magnitude power reduction over conventional time-multiplexed architectures is possible at the cost of increased silicon area. A new type of voltage scaling strategy is introduced here which involves dynamically adjusting the supply voltage based on time varying computational requirements. The second attribute of signal processing systems that can be exploited for low-power is the correlation present in the data being processed. An important goal of low-power architecture selection is to preserve naturally occurring signal correlations and minimize switching activity.

2. Power Estimation

For CMOS circuits, it is well known that power is primarily dissipated in charging and discharging parasitic capacitors and is given by $\alpha_{0\to1} C_L V_{dd}^2 f_{clk}$, where C_L is the load capacitance, V_{dd} is the power supply voltage, f_{clk} is the clock frequency, and $\alpha_{0\to1}$ is the node transition activity factor. For example, Figure 1 shows the activity factor, $\alpha_{0\to1}$, for a simple 2-input NOR gate as a function of the input signal probabilities, p_a and p_b.

Estimating the switching activity turns out to be a very challenging problem since it depends not only on the statistics of the data being processed, but also on low-level circuit and technology parameters (such as gate delay and non-linear capacitances) which affect the glitching component of power. Accurate power estimators do exist at the circuit level (e.g., SPICE or PowerMill), but they tend to be very slow and are not suitable for architectural design space exploration. We have developed an architectural power estimation tool with accuracy close to low-level circuit simulators while being orders of magnitude faster.

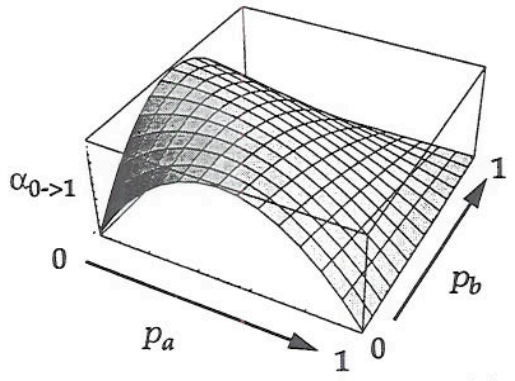

Figure 1. Dependence of power on signal statistics.

The tool works with a structural Verilog circuit description and estimates power by monitoring transitions based on user specified input patterns. A unique feature of this tool compared to other switched capacitance simulators is that it takes into account capacitance vs. voltage non-linearities for the gate and junction components. The tool uses a hybrid approach in which some modules are evaluated using simulation (e.g., a multiplier whose energy consumption is strongly dependent on input patterns) and others using high-level black box models (e.g., an SRAM array where the energy per access is independent of data patterns). The tool also allows run time adaptation of supply voltage.

The power estimates and simulation speed for a 300,000 transistor video decompression DSP are show in Table 1.

Table 1. Power Estimates for a video decoder.

Tool	Estimated Average Power	Execution Time
PowerMill (low precision mode)	16.137 mW	10,925 Sec
Our RTL Estimator	15.8 mW	475 Sec

3. Minimizing Switched Capacitance

Minimizing the switched capacitance ($\Sigma\alpha_{0->1} C_L$) involves optimizing at all levels of abstraction ranging from low-level transistor sizing and logic design to operation sequencing and choice of number representation. This section focuses on architectural and algorithmic techniques we have developed to minimize the switched capacitance.

3.1 Delay Line Structures

An important problem in many signal processing and communications applications (e.g., matched filters) involves minimizing the power dissipation of delay line or shift register structures. The transition activity of shift register structures can be reduced using parallelism. For example, the computational transitions of a shift register of length L clocked at f can be reduced by using n length L/n shift-registers in parallel clocked at f/n, as shown in Figure 2 (shown for the case with n=2). Ideally, this gives a power reduction of $1/n$. However, the overhead associated with parallelism (i.e., the multiplexors, producing the excess clock phases, and routing) limits the amount of power savings. Figure 3 shows the power dissipation of various length shift registers each normalized to the case with no parallelism.

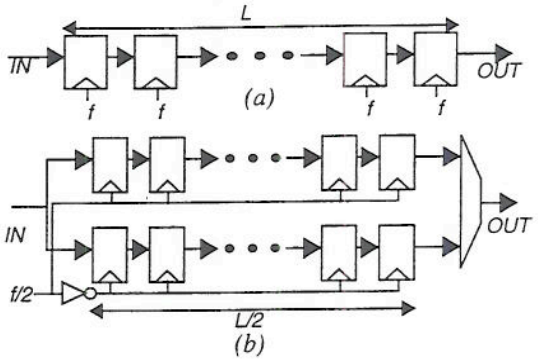

Figure 2. Shift register configuration (a) serial (b) parallel

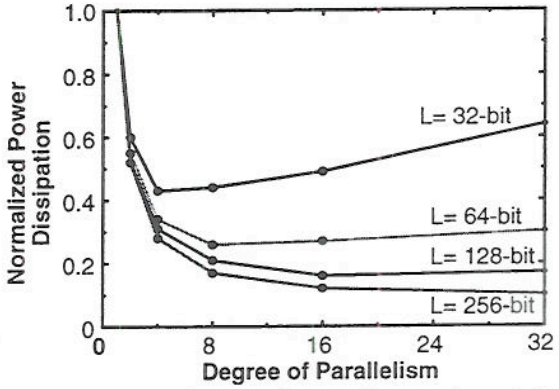

Figure 3. Power Consumption vs. degree of parallelism.

Significantly lower power reduction numbers are observed when many taps from the delay line have to be extracted. For example, for a 520-bit 32 tap digital matched filter, only a factor of 2 power reduction is achieved over a standard implementation due to the high overhead of extracting the taps.

3.2 Operation Sequencing

The switching activity of filter structures can be minimized by optimizing the sequencing of operations. Figure 4 shows two different approaches for implementing an FIR filter. On the left is a conventional approach in which one output sample is produced for every new input sample and enough buffering is provided to hold data samples on the order of the filter length. Each coefficient is accessed once to compute each output sample. For the three tap filter shown, the three coefficients (a_0,a_1, a_2) are each accessed once per sample period.

Another approach is to unroll the computation and re-order the operations. The right side of Figure 4 shows the case with N=3. In this case, the three multiplications corresponding to a_0 are performed, followed by the three corresponding to a_1 and finally the ones corresponding to a_2. Since each

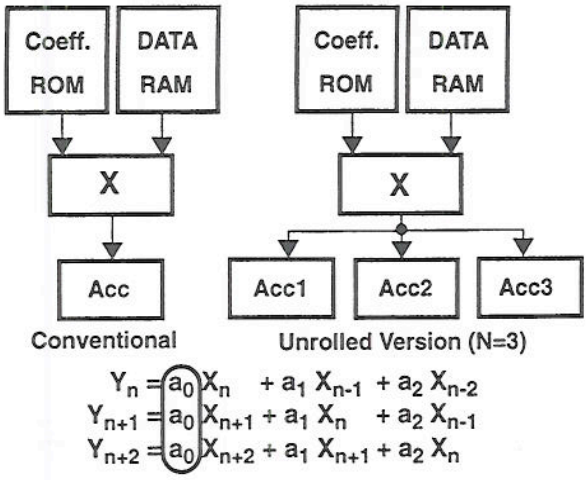

Figure 4. Loop unrolling to reduce switching activity.

coefficient is accessed only once to compute three output samples, the energy dissipation of ROM access and bus transfer is reduced. Also, the energy per multiplication is reduced since one input to the multiplier has a reduced switching activity. The penalty for this approach is the increased number of accumulators and slightly larger data RAM.

3.3 Variable Length Filtering

In many contexts it is important to structure algorithms and systems to allow for the possibility of trading off between the accuracy or optimality of the result and the utilization of resources such as power, time, bandwidth, etc. The formal approach to tradeoff quality for resources has been termed "approximate signal processing" [2]. A well established example in communications involves the use of scalable video coding algorithms to trade quality for bitrate as the available channel bandwidth varies. Scalable architectures are required to minimize energy consumption of processors when the quality requirements on computational results change.

Approximate signal processing concepts can be applied to the design of adaptive digital filters [3]. Adaptive filtering algorithms have generally been used to dynamically change the values of the filter coefficients, while maintaining a fixed filter order. In contrast, the approximate processing approach involves the dynamic adjustment of the filter order. The basic idea is to exploit the fact that in many frequency-selective filter applications, the out of band signal ("noise") can vary with time and therefore the filter order can be dynamically adjusted. The idea is to keep the stopband energy in the filter output below a specified threshold while using as small a filter order as possible.

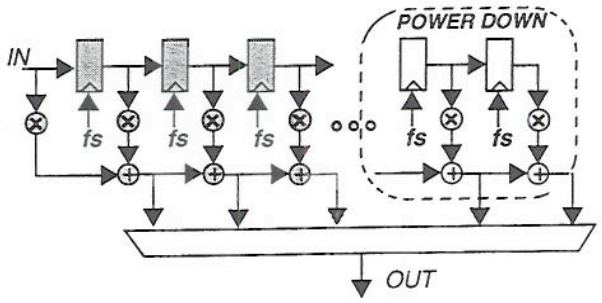

Figure 5. Conceptual diagram of a variable length tapped delay line.

Figure 5 shows a conceptual block diagram of a variable length tapped delay line. Since power consumption is proportional to filter order, this approach achieves power reduction with respect to a fixed-order filter whose output is similarly guaranteed to have stopband energy below the specified threshold. Power reduction is achieved by dynamically minimizing the order of the digital filter.

The key challenge in such a filtering system is to determine the filter order as a function of time. Let a signal, $x[n]$, consist of a passband component, $x_p[n]$, and a stopband component, $x_s[n]$ (i.e., $x[n] = x_p[n] + x_s[n]$). If it were possible to *cost-effectively* measure the strength of the stopband component, $x_s[n]$, from observation of $x[n]$, the amount of stopband attenuation needed at any particular time can be determined. When the energy in $x_s[n]$ increases, it is desirable to increase the stopband attenuation of the filter. This can be accomplished by using a higher-order filter. Conversely, the filter order may be lowered when the energy in $x_s[n]$ decreases. We have developed a practical low-overhead approach, using feedback techniques, for dynamically estimating the energy fluctuations in the stopband component, $x_s[n]$, and using them to adjust the order of a frequency-selective filter.

The amount of power reduction is a strong function of signal statistics. Figure 7 illustrates the nature of the adaptation performed for a frequency division multiplexing example where the out of band signal varies. One of the curves shows the evolution of the filter order while the other curve shows the energy profile of the stopband signal. Clearly, the variations in filter order roughly follow the energy variations of the stopband signal. In particular, the most power savings is achieved during the silent regions of the stopband signal.

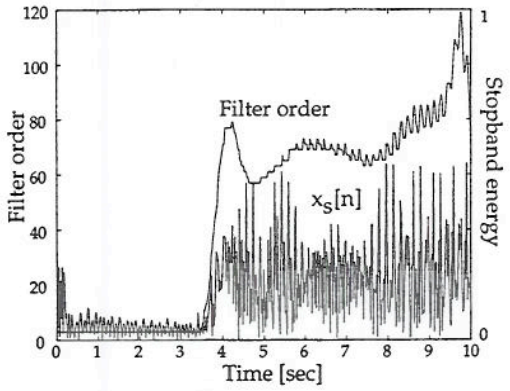

Figure 6. Filter order vs. time.

4. Optimizing Power Supply Voltage

From the previous section, it is clear that there are applications in DSP where the computational workload per input sample varies with time. An approach to reduce the energy consumption of such systems beyond shut down involves the dynamic adjustment of supply voltage based on computational workload [4]. The basic idea is to lower power supply voltage and clock frequency when the workload is less than peak instead of working at a fixed supply and idling for some fraction of the time.

Figure 7 shows a plot of power vs. normalized workload ($V_{DD(max)}$=2V and V_T=0.4). The straight line represents a fixed power supply system. If the workload is less than peak workload, then the processor would compute as fast as possible at a fixed voltage and idle for a fraction of the sample period. The lower curve represents the case for variable power supply system. If the workload for a given sample period is less than peak, then the delay of the processing element can be increased by a factor of 1/workload without loss in throughput, allowing the processor to operate at a lower power supply voltage. A test chip has been fabricated and characterized to validate the variable voltage concept.

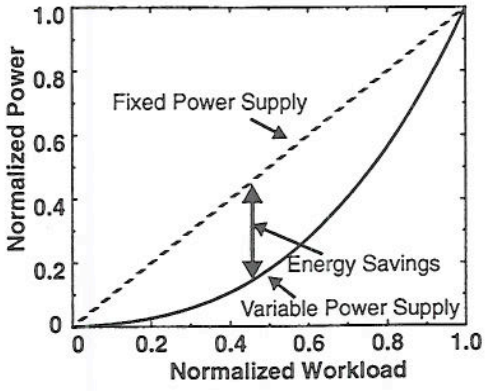

Figure 7. Power reduction using variable supply.

5. Conclusion

Optimizing at the architecture and algorithm level often has the greatest impact on the power dissipation. The number of operations performed per sample and capacitance switched per operation can be minimized by exploiting signal statistics. When the number of operations performed per sample varies dynamically, a variable supply voltage can be employed to lower energy dissipation.

Acknowledgments

Some of the work presented in this report is funded in part by DARPA under the Low Power Electronics program and the RASSP program.

References

[1] A. Chandrakasan, R. Brodersen, *Low Power Digital CMOS Design*, Kluwer Academic Publishers, July 1995.

[2] H. Nawab, A. Oppenheim, A. Chandrakasan, J. Ludwig, J. Winograd, "Approximate Signal Processing", to appear in the *Journal of VLSI Signal Processing*, December 1996.

[3] J. Ludwig, H. Nawab, A. Chandraksan, "Low-Power Digital Filtering Using Approximate Processing," *IEEE Journal of Solid-State Circuits*, pp. 395-400, March 1996.

[4] V. Gutnik, A. Chandrakasan, "An Efficient Controller for Variable Supply-voltage Low Power Processing," *IEEE VLSI Circuits Symposium*, pp. 158-159, June 1996.