Low-Power Digital Systems Based on Adiabatic-Switching Principles

William C. Athas, Lars "J." Svensson, *Member, IEEE,* Jeffrey G. Koller,
Nestoras Tzartzanis, and Eric Ying-Chin Chou, *Student Member, IEEE*

Abstract— **Adiabatic switching is an approach to low-power digital circuits that differs fundamentally from other practical low-power techniques. When adiabatic switching is used, the signal energies stored on circuit capacitances may be recycled instead of dissipated as heat. We describe the fundamental adiabatic amplifier circuit and analyze its performance. The dissipation of the adiabatic amplifier is compared to that of conventional switching circuits, both for the case of a fixed voltage swing and the case when the voltage swing can be scaled to reduce power dissipation. We show how combinational and sequential adiabatic-switching logic circuits may be constructed and describe the timing restrictions required for adiabatic operation. Small chip-building experiments have been performed to validate the techniques and to analyse the associated circuit overhead.**

Index Terms— **Adiabatic amplification, adiabatic charging, adiabatic switching, low-power CMOS, reversible computation, switching energy reduction with preserved signal energies.**

I. Introduction

THE IMPORTANCE of reducing power dissipation in digital systems is increasing as the range and sophistication of applications in portable and embedded computing continues to increase. System-level issues such as battery life, weight, and size are directly affected by power dissipation. Inroads into reducing power dissipation of the digital systems only serves to improve the performance and capabilities of these systems.

CMOS has prevailed as the technology of choice for implementing low-power digital systems. One of the most important reasons is the reduction in switching energy per device caused by the continually shrinking feature sizes. Another important reason is the almost ideal switch characteristics of the MOS transistor, which translates into a negligible static power dissipation compared to the switching (transient) power dissipation.

Basic energy and charge conservation principles can be used to derive the switching energy and power dissipation for static, fully restoring CMOS logic. Throughout this article, we will frequently refer to such CMOS logic as "conventional" logic. Consider the generic CMOS gate shown in Fig. 1. A load capacitance C_L, representing the input capacitance of the next logic stage and any parasitic capacitances, is connected to the dc supply voltage V_{dd} through a pull-up block composed of pFET's and to ground through a pull-down block of nFET's. With the pull-down network tied and the pull-up network cut, the output is discharged to ground. When the inputs change so that the pull-up network is tied and the pull-down network cut, charge will flow out of V_{dd} and into the load capacitance until the output reaches V_{dd}.

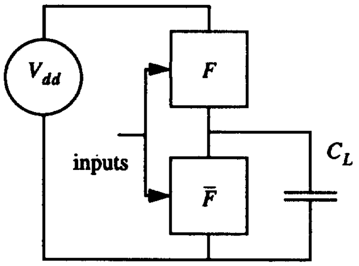


Fig. 1. Generic, conventional CMOS logic gate with pull-up and pull-down networks.

In the process of charging the output, a charge of size $Q = C_L V_{dd}$ is delivered to the load. The power supply must supply this charge at voltage V_{dd}, so the energy supplied is $Q \cdot V_{dd} = C_L \cdot V_{dd}^2$. The energy stored on a capacitance C_L charged to V_{dd} is only half of this, or (1/2) $C_L V_{dd}^2$. Because energy is conserved, the other half must be dissipated by the pFET's in the pull-up network. The same amount of energy is dissipated, regardless of the make-up of the network, the resistance of the pFET's, and the time taken to complete the charging. Similarly, when the inputs change again causing the output to discharge, all of the signal energy stored on the capacitance is inevitably dissipated in the pull-down network, because no energy can enter the ground rail ($Q \cdot V_{\text{gnd}} = Q \cdot 0 = 0$).

Thus, from an energy conservation perspective, the conventional case represents a maximum of wastefulness. All charge is input to the circuit at voltage V_{dd} and exists at voltage 0. The energy of the charge upon entry is $C_L V_{dd} \cdot V_{dd}$. The energy of the charge upon exit is 0. Energy dissipation from delivery to removal of the charge is $C_L V_{dd}^2$. All of this energy is dissipated as heat.

Since the energy dissipated when a signal capacitance is cycled is fixed at twice the signal energy, the only way to reduce the energy dissipation in a conventional CMOS gate is to reduce the signal energy. This increases the sensitivity to background noise and thus the probability of malfunction.

Switching energy can be made to be considerably less than signal energy. From the theory of charge control [13], charges can be distinguished as either controlling or controlled charge.

Manuscript received June 15, 1994; August 23, 1994. This work was supported by ARPA under Contract DABT63-92-C0052.

The authors are with the Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292 USA.

IEEE Log Number 9406367.

1063-8210/94$04.00 © 1994 IEEE

For MOSFET's, the controlling charge is on the gate, while the controlled charge flows through the channel. Dissipation is caused by the resistance encountered by the controlled charge. If charge transport is slowed down, energy that would otherwise be dissipated in the channel can be conserved for later reuse. The energy advantage can be readily understood by assuming a constant current source that delivers the charge $C_L V_{dd}$ over a time period T. The dissipation through the channel resistance R is then:

$$E_{\text{diss}} = P \cdot T = I^2 \cdot R \cdot T = \left(\frac{C_L V_{dd}}{T}\right)^2 \cdot R \cdot T$$
$$= \left(\frac{RC_L}{T}\right) \cdot C_L V_{dd}^2. \quad (1)$$

Equation (1) shows that it is possible to charge and discharge a capacitance through a resistance while dissipating less than $C_L V_{dd}^2$ of energy. It also suggests that it is possible to reduce the dissipation to an arbitrary degree by increasing the switching time to ever-larger values. We refer to this as the principle of *adiabatic charging*. We use the term "adiabatic" to indicate that all charge transfer is to occur without generating heat. As is the typical usage of the term in thermodynamics, fully adiabatic operation is an ideal condition that is asymptotically approached as the process is slowed down. To the best of our knowledge, Seitz and co-workers [1] were the first to formulate the relationship of (1) and to use the effect in practical digital circuits.

Switching circuits that charge and discharge their load capacitance adiabatically are said to use *adiabatic switching*. This article describes and analyzes the power dissipation and performance of adiabatically-switching logic circuits built in CMOS. The circuits rely on special power supplies that provide accurate pulsed-power delivery. It is important to note that adiabatic switching techniques can be an attractive alternative to other low-power design approaches only if the supplies can deliver power efficiently and recycle the power fed back to them. The circuits described here are compatible with energy-efficient, resonant power supplies that we have developed and prototyped [2].

Many factors must be considered when determining the conditions for which adiabatic switching offers superior low-power operation to other approaches. When signal-voltage swing is significantly greater than the threshold voltages of the CMOS devices, the advantages can be readily determined from the analysis of the *adiabatic amplifier*, which is discussed in detail in Section II. On the other hand, when signal-voltage swing can be scaled down to reduce power dissipation, the advantages of adiabatic amplification rapidly diminish for all but the slowest applications, as shown in Section III.

The adiabatic amplifier and the simple gate structure on which it is based can be straightforwardly generalized so that in addition to amplification or buffering, it can implement arbitrary logic functions. Section IV describes the transition from combinational logic to sequential logic. The transition constitutes a turning point, since conventional storage elements cannot be made fully adiabatic. Reversible-logic techniques may be used to avoid storage elements and thus make it possible to build fully adiabatic sequential systems. To this end, we have developed techniques for constructing reversible logic pipelines similar in organization to those developed and demonstrated by Younis and Knight [9]. Section IV also includes a description of a design exercise, where the reversible pipeline structure is used to construct a highly pipelined adder. The results of this exercise indicate that for reversible logic to be a competitive approach to low-power CMOS, either the premium on reducing the energy dissipated per operation must be extremely high, or logic styles and synthesis procedures need to be invented which result in circuit designs with much less logical overhead.

II. Adiabatic Amplification

The adiabatic amplifier, which is the fundamental circuit to our approach, is a simple buffer circuit that uses adiabatic charging to drive a capacitive load. It is useful in a stand-alone configuration, and also as a part of more sophisticated circuits. In this section, we analyse its efficiency and describe its stand-alone use as a line driver for an address bus on a memory board.

Equation (1) assumes that C_L is charged or discharged through a constant resistance. Conventional gates, such as that in Fig. 1, use pull-up and pull-down networks composed of nonlinear FET devices. Fortunately, a switch network can be linearized to a first approximation by replacing each FET with a fully restoring CMOS transmission gate (T-gate). The T-gate is built from an nFET and a pFET connected in parallel. To tie the T-gate with minimal on-resistance, the gate of the pFET is grounded and the gate of the nFET is tied to V_{dd}. For a small voltage drop across the T-gate, which is the intended region for adiabatic circuits, we may model the conductances, G_p and G_n, of the FET's as:

$$G_p = \frac{C_p}{K_p}(V_{\text{ch}} - V_{\text{th}}) \quad (2)$$

$$G_n = \frac{C_n}{K_n}(V_{dd} - V_{\text{ch}} - V_{\text{th}}). \quad (3)$$

C_p and C_n are the gate capacitances of the FET's, V_{th} is the threshold voltage of the nFET, and K_p and K_n represent process constants independent of gate capacitance and voltage. V_{ch} is the average channel voltage. We assume that the channel length is held constant, so that the gate capacitance is directly proportional to the channel width. These equations do not take into account body effects or the difference in threshold voltage between nFET's and pFET's. Accuracy is therefore limited to a factor of two.

By selecting the widths of the two FET's such that $K_n/C_n = K_p/C_p$, we may simplify the sum of the two conductances to:

$$G_p + G_n = \frac{C_n}{K_n}(V_{dd} - V_{\text{ch}} - V_{\text{th}} + V_{\text{ch}} - V_{\text{th}})$$
$$= \frac{C_n}{K_n}(V_{dd} - 2V_{\text{th}}) \quad (4)$$

The on-resistance of the T-gate is the reciprocal of this sum:

$$R_{\text{tg}} = \frac{K_n}{C_n(V_{dd} - 2V_{\text{th}})}. \quad (5)$$

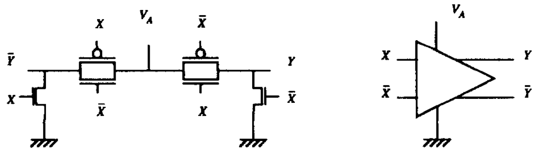


Fig. 2. Adiabatic amplifier: schematic and logic symbol. V_A should be connected to a pulsed-power supply.

This formulation assumes that both devices are conducting, which is not the case when V_{ch} is within one threshold voltage of either supply rail. However, at these extremes the on-resistance will be less than the value given by (5), so the formulation can be used as a worst-case estimate.

The adiabatic amplifier of Fig. 2 is a simple but important circuit element, which may be used to drive a capacitive load by adiabatic charging. The amplifier consists of two T-gates and two nFET clamps. The input is dual-rail encoded, since both signal polarities are needed to control the T-gates. The output is also dual-rail encoded; this choice helps keep the capacitive load seen by the power supply data-independent, which simplifies its design, and is also required when other T-gates (such as other amplifiers) are to be controlled by the output signal. The operation of the amplifier is straightforward. First, the input is set to a valid value. Next, the amplifier is "energized" by applying to V_A a slow voltage ramp from 0 to V_{dd}. If the ramp is slow compared to $R_{tg}C_L$, one of the load capacitances will be adiabatically charged through one of the T-gates. The output signal is now valid and can be used as an input to other circuits. Next, the amplifier is de-energized by ramping the voltage on V_A back to 0. The signal energy that was stored on the load capacitance flows back into the power supply connected to V_A. It falls on the power supply to recycle this returned energy.

We now analyze the energy efficiency of the amplifier, using (1) and (5). The dissipation caused directly by charging and discharging the load capacitance, E_{load}, can be expressed by combining (1) and (5). To account for a nonconstant charge current, the dissipation of (1) must be multiplied by a constant shape factor ξ (which takes the value $\pi^2/8 \approx 1.23$ for a sine-shaped current). The simple formula of (1) becomes:

$$E_{\text{load}} = 2\xi\frac{R_{\text{tg}}C_L}{T}C_LV_{dd}^2 = \frac{2\xi}{T}\frac{K_n}{C_n(V_{dd}-2V_{\text{th}})}C_L^2V_{dd}^2. \quad (6)$$

Additionally, parasitic effects such as diffusion capacitance of the T-gates and the clamp nFET's will contribute to the total load capacitance. To account for these contributions would unduly complicate the equations and would not contribute to an understanding of the general result. For the analysis that follows, these capacitances will be treated as negligible by assuming that total input capacitance is small relative to the effective load.

Energy is also dissipated to drive the input capacitance. For the moment, we will analyse the case when the inputs are driven conventionally to V_{dd}, so that the input energy, E_{in}, is dissipated. We furthermore assume that the total input capacitance is proportional to the gate capacitance of the nFET of the T-gate, with a constant of proportionality α; if no parasitics are considered, $\alpha = (C_n + C_p)/C_n$. Thus, $E_{\text{in}} = \alpha C_nV_{dd}^2$. The total energy dissipated per cycle is then:

$$E_{\text{total}} = E_{\text{in}} + E_{\text{load}} = \alpha C_nV_{dd}^2 + \frac{2\xi}{T}\frac{K_n}{C_n(V_{dd}-2V_{\text{th}})}C_L^2V_{dd}^2. \quad (7)$$

Equation (7) defines an important trade-off in adiabatic CMOS circuits. When the channel width, and thereby the input capacitance, is increased, the energy dissipated in charging the load decreases, but the energy dissipated in charging the inputs increases proportionately. If C_n is a free parameter in the design of the amplifier, the minimal energy dissipation is achieved when the two terms of (7) are equal:

$$C_{n_{\text{opt}}} = \sqrt{\frac{2\xi}{\alpha T}\frac{K_n}{(V_{dd}-2V_{\text{th}})}}C_L. \quad (8)$$

Inserting (8) into (7) yields the following expression for the minimum dissipation:

$$E_{\text{total}_{\text{min}}} = 2\cdot\alpha C_{n_{\text{opt}}}V_{dd}^2 = \sqrt{\frac{8\alpha\xi}{T}\frac{K_n}{(V_{dd}-2V_{\text{th}})}}C_LV_{dd}^2. \quad (9)$$

Equation (9) illustrates a fundamental limitation[1] for adiabatic charging with conventionally driven control signals: the switching energy will only scale as $T^{-1/2}$, as opposed to the T^{-1} of (1). An obvious improvement is to use two amplifier stages, where a smaller adiabatic amplifier is used to drive the input signals of the final stage. Let C_{n1} and C_{n1} be the gate capacitances of the nFET's of the T-gates of the first and second amplifier, respectively. The input capacitance of the second amplifier, αC_{n2}, then constitutes the load of the first amplifier. Allotting half of the total charging time to each stage, we get:

$$E_{\text{total}} = \alpha C_{n1}V_{dd}^2 + \frac{K_n}{(V_{dd}-2V_{\text{th}})}\frac{4\xi}{T}\left(\frac{\alpha^2C_{n2}^2}{C_{n1}} + \frac{C_L^2}{C_{n2}}\right)V_{dd}^2. \quad (10)$$

With both C_{n1} and C_{n2} as free variables, the total minimum energy dissipation is:

$$E_{\text{total}_{\text{min}}} = 8\left(\frac{\alpha\xi}{T}\frac{K_n}{(V_{dd}-2V_{\text{th}})}\right)^{3/4}C_LV_{dd}^2. \quad (11)$$

From (9) to (11), the energy scaling improves from $T^{-1/2}$ to $T^{-3/4}$. It can be shown that for a cascading of n such amplifiers, the energy dissipation scales as:

$$E_{\text{total}_{\text{min}}} \sim T^{2^{-n}-1}. \quad (12)$$

This result is not very useful in practice, since the parasitic capacitances can no longer be neglected when the capacitances of the final gate-drive FET's become comparable to that of the load. Also, if the switching is to be performed within a constant time interval, the ramp time for each of the steps must decrease at least linearly in the number of cascaded amplifier stages.

These results present a dilemma. The cost of using conventionally driven inputs with adiabatic logic is a scaling in switching energy which decreases sublinearly with increases

[1] A similar relationship has been derived previously for the case of resonant power supplies based on power MOSFET circuits [16].

in switching time. T^{-1} scaling may be asymptotically approached, even when the inputs of the amplifier cascade are driven conventionally, but to immediately achieve linear scaling requires the use of all-adiabatic circuits, where the input signals of all amplifiers are also switched adiabatically. The implications for all-adiabatic operation are serious and are discussed in detail in Section IV.

A. The Adiabatic Line Driver

As shown above, adiabatic amplifiers make it possible to reduce switching energy while signal energy is held constant, something which cannot be done with conventional CMOS switching. To investigate how well the theory reflects reality, we designed, implemented, and tested an adiabatic line driver chip (ALDC) for driving an address bus of a memory board [2]. This application involves relatively large driven capacitances and moderate speed requirements, which translates to considerable theoretical power savings. The signal energies are fixed by voltage swing requirements and input capacitances of the memory chips, so no improvements are possible with conventional drivers. The ALDC is also very simple, allowing us to characterize it fully and isolate the non-ideal phenomena.

The ALDC comprises eight adiabatic amplifiers in a 40-pin DIP, sharing the same supply voltage. It was designed to use 5% of the power of a conventional driver (according to (1)) when driving 100 pF per output line at 1 MHz, and is implemented in a MOSIS-supplied 2 μm process. A simple pulsed-power supply, outlined in Fig. 3, was used to test the circuit. The power supply and some parasitics not present in a conventional driver limited the actual dissipation gain from a factor of 20 to a factor of 6.3. It is instructive to consider the distribution of the dissipation over the different parts of the circuit: 51% of the power dissipation is in the ALDC, 29% is in the channel of the power supply FET switch used to time the energy transfer, 17% is in the circuit driving the gate of that switch, and 3% is in the parasitic resistance in the inductor. This breakdown emphasizes the importance of taking the dissipation of the entire system (including the power supply) into the account, not only that of the logic circuit.

III. Supply Voltage Selection

For many applications, signal energy levels are determined by the input/output requirements of the digital system. Industry standards for interfacing (such as for the ALDC) or the peculiar requirements of the physical properties of the interface device (such as for liquid crystal displays) mandate certain voltage swings. For these applications, the benefits of adiabatic switching improve with increasing voltage swing, as can be seen by comparing the energy required to cycle a signal conventionally to the adiabatic dissipation of (6):

$$\frac{E_{\text{conv}}}{E_{\text{load}}} = \frac{C_L V_{dd}^2}{E_{\text{load}}} = \frac{C_n T}{2\xi K_n C_L}(V_{dd} - 2V_{\text{th}}). \tag{13}$$

The energy savings ratio increases linearly with V_{dd} for an all-adiabatic system and increases as $V_{dd}^{1/2}$ or $V_{dd}^{3/4}$ for the mixed approaches described by (9) or (11).

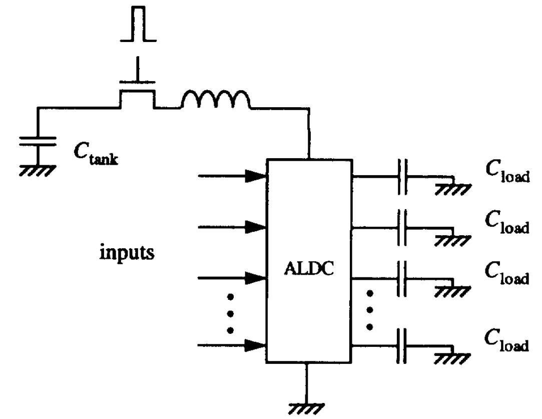


Fig. 3. Test setup of ALDC. The capacitor C_{tank} is held at V_{dd}. The ALDC inputs are set with all output capacitances discharged; the nFET switch is pulsed; the current pulse through the inductor is steered to the intended output capacitances. The nFET is pulsed again for the charge to return to the tank capacitor. Only thereafter may the inputs change.

On the other hand, if we are free to choose the voltage swing for minimum dissipation, we might try to scale down signal energies (by increasing V_{dd}) as far as the application permits and then apply adiabatic charging techniques to further reduce switching energies. This is, however, an unattractive strategy that does not lead to minimal dissipation for T-gate based circuits. In fact, there is an optimal supply voltage swing for T-gate-based adiabatic circuits. This voltage will vary with the approach taken; for simplicity, we will analyse the all-adiabatic approach. Equation (6) can be rewritten:

$$\begin{aligned} E_{\text{load}} &= \frac{2\xi}{T}\frac{K_n}{C_n(V_{dd} - 2V_{\text{th}})} C_L^2 V_{dd}^2 \\ &= \frac{2\xi}{T}\frac{K_n C_L^2}{C_n}\left[\frac{V_{dd}^2}{(V_{dd} - 2V_{\text{th}})}\right]. \end{aligned} \tag{14}$$

The energy has a minimum for $V_{dd} = 4V_{\text{th}}$. Again, neither the body effect nor the regions where only one FET device is turned on are accounted for. SPICE simulations and more detailed analysis which takes into account body effects, subthreshold regions, and waveform shapes all show that the minimum is reasonably shallow and that $4V_{\text{th}}$ yields close to the optimal dissipation. When $4V_{\text{th}}$ is substituted for V_{dd} in (6), the energy dissipation of the adiabatic amplifier reduces to:

$$E_{V_{\text{opt}}} = \frac{C_L^2}{C_n} \cdot \frac{16\xi K_n V_{\text{th}}}{T}. \tag{15}$$

We now wish to compare this dissipation to that of a conventional inverter driving the same capacitance, but where the voltage swing can be chosen freely. The conventional circuit will use devices of the same size as the adiabatic one, and its supply voltage will be selected to equalize the speed of the two drivers. To model switching delay as a function of supply voltage for conventional circuits, we use the following well-known approximation [3]:

$$T = \frac{C_L V_{dd}}{I_{\text{sat}}} = \frac{2K_n V_{dd}}{(V_{dd} - V_{\text{th}})^2}\frac{C_L}{C_n}. \tag{16}$$

This expression is based on the drain current of a device in saturation. Velocity saturation is unlikely to influence the delay for the low supply voltages we are interested in and is consequently not taken into account. Mathematical analysis confirmed by SPICE simulations of a ring oscillator indicates that this formula is accurate down to $V_{dd} \approx 1.3V_{\rm th}$. From (16), we can now express the required supply voltage and the resulting switching energy as functions of the allowed switching time:

$$V_{dd} = V_{\rm th} + \frac{rK_n + \sqrt{(rK_n)(rK_n + 2TV_{\rm th})}}{T} \qquad (17)$$

$$\begin{aligned} E_{\rm conv} &= C_L V_{dd}^2 \\ &= C_L \left(V_{\rm th} + \frac{rK_n + \sqrt{(rK_n)(rK_n + 2TV_{\rm th})}}{T} \right)^2 . \end{aligned} \qquad (18)$$

The parameter r relates the output and input capacitances ($r = C_L/C_n$). We see from (18) that when switching time is increased in a conventional CMOS circuit by reducing supply voltage, dissipation will initially fall off as T^{-2}. It will, however, tend asymptotically to $C_L V_{\rm th}^2$ when T grows very large and V_{dd} approaches $V_{\rm th}$. The adiabatic case has no such asymptote, as is seen in (15). Fig. 4 is a graph of the dissipation of the two approaches for a MOSIS 2 μm CMOS process and $r = 10$. Below 5 ns, (15) is inapplicable because the charging time is no longer sufficiently long compared to $R_{tg}C_L$ at 4 V_{th}. At 25 ns, V_{dd} for (18) equals 1.3 V_{th}, which is the lower limit for using (16).

We assumed in this analysis that the device sizes for both approaches are equal. Thus, the input capacitances for both circuits are approximately equal, if the conventional driver is ratioed according to carrier mobility. However, the adiabatic driver requires that dual-rail signalling be used, while conventional one does not. This requirement for dual-rail signalling tends to proliferate throughout the system, doubling the hardware amount. Assuming that both signal polarities of the input are available, the choice between the two approaches depends on the allowable switching time and on the input/output requirements. Initially, supply voltage reduction is very attractive, since there is a quadratic decrease in energy dissipation as switching time is increased. However, its scalability is limited and depends on having the flexibility to set supply voltages based on circuit latency requirements.

In practical circuits, several additional sources of overhead must be considered for an adiabatic circuit to offer lower energy dissipation than a conventional one with a freely chosen supply voltage. These include the FET clamps on undriven outputs and additional sequencing requirements so that energy can be efficiently delivered and recovered. The modest gains indicated by Fig. 4 may easily be overwhelmed. Furthermore, the situation becomes more severe when sequential functions with adiabatic charging are attempted, as will be seen in the next section.

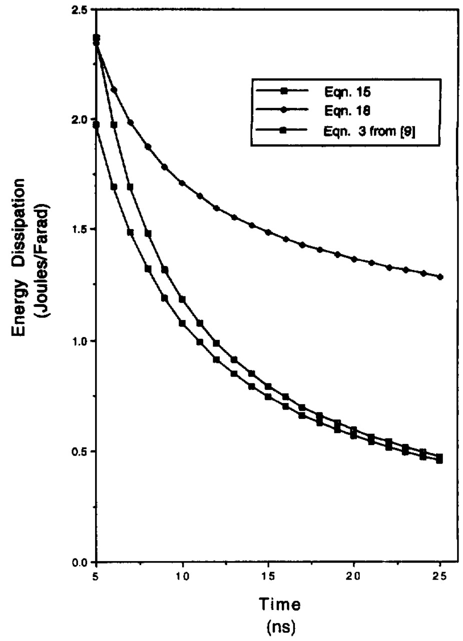


Fig. 4. Trading energy for delay for a conventional (18) and an adiabatic (15) driver. The transition time of the conventional driver is set by adjusting the supply voltage (V_{dd}). The voltage swing of the adiabatic driver is set at $4V_{th}$. Equation 3 from Younis [3] is the exact formula for energy dissipation when $T >> RC$ does not hold.

IV. Adiabatic Logic Circuits

A straightforward extension of the adiabatic amplifier lets us implement arbitrary combinational logic functions that use adiabatic charging. The circuit of Fig. 1 may be transformed into an adiabatically-switched circuit by replacing each of the pFET's and nFET's in the pull-up and pull-down networks with T-gates, and by using the expanded pull-up network to charge the true output and the expanded pull-down network for the complement output. Fig. 5 depicts this circuit transformation. Both networks in the transformed circuit are used to both charge and discharge the output capacitances. The dc V_{dd} source of the original circuit must be replaced by a pulsed-power source to allow adiabatic operation. The optimal sizes of the T-gates of the function networks can be determined from the equations of Section II.

As with conventional CMOS logic, it is desirable for reasons of performance and complexity management to partition a large block of logic into smaller ones and then compose them to implement the original larger function. However, if values are allowed to ripple through a chain of cascaded adiabatic logic gates, non-adiabatic flow of energy will occur. Adiabatic operation is possible only if the inputs of a gate are held stable while the gate is energized. As observed by Hall [5], a cascade of initially de-energized circuits may be energized in succession, but must then be de-energized in reverse order before the input values to the cascade may change. These "retractile cascades" are impractical for several reasons: they require a large and possibly indeterminate number of supply voltage waveforms; these waveforms all have different pulse widths;

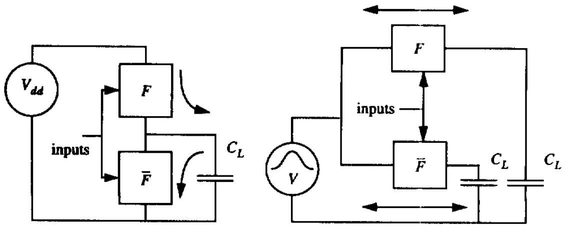


Fig. 5. (a) A conventional CMOS logic gate. (b) A corresponding adiabatic gate may be constructed by reorganizing the switch networks as shown, replacing the pFET's and nFET's of the conventional gate with T-gates. The arrows indicate the charge flow when the load capacitances are charged and discharged.

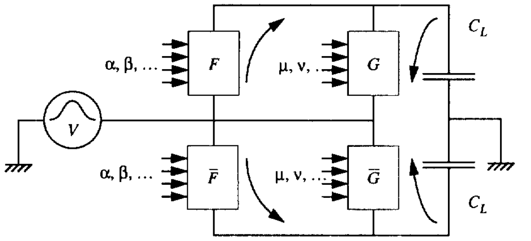


Fig. 6. A pipelinable adiabatic gate. The inputs $\alpha, \beta, \ldots$ control the F paths through which one of the load capacitances is charged. The G paths, used for discharge, are controlled by a different set of signals: $\mu, \nu, \ldots$ As in Fig. 5, arrows indicate charge flow.

and since an N stage cascade requires time proportional to N to produce each result, the latency is proportional to N and the throughput is proportional to $1/N$.

Pipelining can be advantageously used to improve the throughput of the system. By using latches to hold the inputs of a retractile stage, we could circumvent the requirement that the preceding stage stays energized until the current stage has been de-energized. Conventional latches, would, however, result in the unfavorable $T^{-1/2}$ or $T^{-3/4}$ scaling of the mixed adiabatic and conventional approaches. If the retractile stage is expanded to include an explicit discharge path for de-energizing, as shown in Fig. 6, the controlling inputs of the energizing paths need not be kept stable to handle de-energizing. A means for pipelining then exists that does not use latches, and T^{-1} scaling can be retained. The control signals for the de-energizing path must be stable throughout the discharge to ensure adiabatic operation. Therefore, they cannot be derived directly from the outputs of the stage being de-energized [6]. Any attempt at schemes based on the same idea is bound to fail for thermodynamic reasons [8]. The de-energizing path must instead be controlled by signals derived from the output of the *following* stage in the pipeline. A pure copy of the signal is sufficient, but this method only defers the problem and is not a general solution when resources are limited. The ability to derive control signals for the de-energizing path is guaranteed if all logic blocks implement functions that are invertible, so that their inputs may be recomputed from their output.

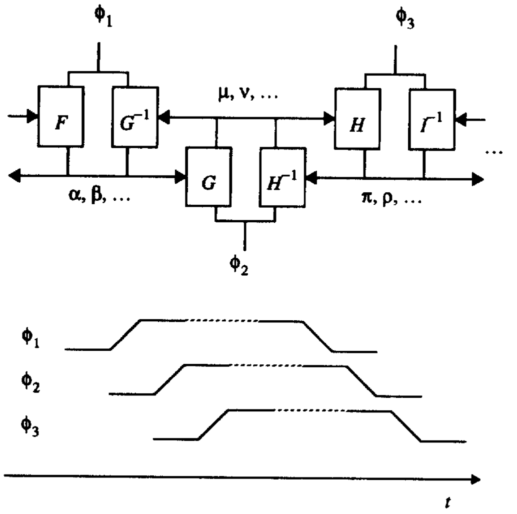


Fig. 7. Conceptual adiabatic pipeline using invertible functions. For simplicity, the multiple switch networks needed for multirail signaling are not shown. α, β, μ, etc. are logic variables computed by the pipeline. The corresponding pulse-power/clock signals, denoted by ϕ_i, are also shown. One stage must be completely energized before the charging of the next stage commences.

By constructing all of the logic stages according to Fig. 6 and restricting the function blocks to invertible functions only, it is possible to assemble a fully adiabatic pipeline as shown in Fig. 7. (A pipeline structure similar in organization but differing in the details has been presented by Younis and Knight [9].) Once energized, the output values of a stage are held for a certain time before de-energizing commences; an idle period separates subsequent energy cycles. During the hold period, the output values are used to set the next pipeline stage and to reset the previous stage. The pulsed-power supply voltages for the pipeline are shown in the figure. The requirement that the outputs of one stage are held while the surrounding stages are energized and de-energized naturally leads to set of supply voltage waveforms partially overlapping in time. Since they not only power the computations but also pace them, the supply waveforms may equally well be thought of as clock signals. The use of exclusively invertible functions makes the pipeline reversible: if the clocks are reversed in time, the pipeline runs backwards. Reversible logic has been studied in theoretical physics since Bennett showed that computations need not destroy information [7]. Since then, a large body of theory has been developed [10]–[12].

The complexities of two mutually inverse function blocks are usually similar, so the typical circuit overhead for the separate discharge path is a factor of two. This overhead is avoided if the de-energizing paths are replaced by diodes. The diodes would be connected with their anodes to the respective output node and their cathodes to the supply. They will not influence circuit behavior while the stage is energized, being back-biased for the uncharged output and shorted by the switch network for the charged output. During de-energizing, the high output will dump its charge back to the supply through the diode. Attractively simple and fast circuits using a similar approach have recently been published by Kramer *et al.* [14]. The dissipation caused by passing the output charge through a diode drop is a drawback, since it is essentially independent of the ramp time.

To validate the reversible pipeline structure shown in Figs. 7 and 8, we designed a small shift register as a CMOS chip. Shifting is an inherently reversible operation, and the simple structure allowed us to stay in the realm of small, highly controllable experiments. The small size also meant that all signal nodes inside the chip could be brought out to the off-chip pads for observation. Following the adiabatic amplification and switching principles outlined above, the shift register was implemented with dual-rail logic and T-gates for the separate charge and discharge paths. Four symmetric clock/supply signals, ninety degrees out of phase, were used, together with off-chip storage capacitances on all circuit nodes to preserve the dynamic data values during the clock transitions. Each "phase stage" of the register required eight transistors, and each "bit stage" was composed for four phase stages, for a total of thirty-two transistors for each stage. A complete cycle of a clock signal would move a bit from one bit stage to the next.

The principle of reversible pipelines assumes that every phase stage is followed by another stage that controls its de-energizing. In practical circuits, the last stage of the pipeline produces a final result of the computation, which is then output and discarded. This discharge of the output nodes of the last pipeline stage will therefore always be dissipative, but the dissipation will be proportional to the number of output bits rather than to the total number of bits in the pipeline. In order to characterize the efficiency of the shift register without having to consider these secondary effects, the input to the shift register was connected to the output, making it a circular and completely reversible machine with no net flow of information into or out of the system.

The shift register, although very simple, is to the best of our knowledge the first working sequential adiabatic CMOS circuit. Its performance was determined simply by measuring the dc current fed to the pulsed-power supply. For a four bit-stage chip, power dissipation (including that of the pulsed-power unit) was 130 μW at 63 kHz driving 82 pF loads to 4.5 V. The slow clocking speed was due in part to the timing generator used in the test setup. The ramp times were set at approximately 1 μs, which suggests that the design could have run at 250 kHz with little modification. At 63 kHz, the energy dissipation per clock cycle was 2.3 nJ. The electrical work done per cycle was to completely charge and discharge eight 82 pF loads to 4.5 V. The equivalent conventional work for driving these loads would be 13.2 nJ.

When logic functions more general than shifting are implemented in a reversible pipeline, signal encoding brings additional overhead compared to the conventional case. The outputs of a gate carry invalid values when the corresponding supply voltage is in its idle phase and therefore de-energized. To prohibit these values from activating devices in the switch networks of the neighboring stages, quad-rail encoding of the signal values can be used [9]. Active-low signals then control the P-channel devices, while active-high signals control the N-channel devices. One pair of signals change from their idle values to represent a zero, the other pair to represent a one. To avoid the extra complexity factor of two brought by quad-rail signalling, enable switches may be placed in series with the switch networks to prevent charge flow during the idle phase. This results in the gate style and clocking scheme depicted in Fig. 8 (two devices have been added to clamp the undriven output to ground).

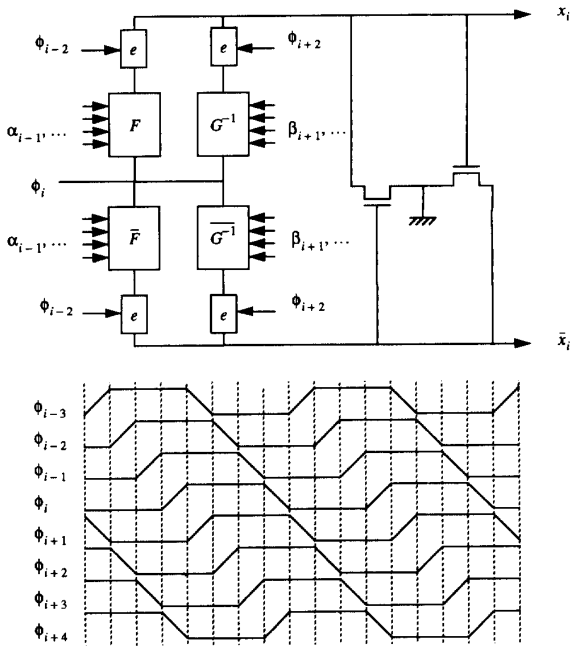


Fig. 8. Detailed dual-rail logic gate for use in reversible pipeline, and its eight-phase adiabatic clock signals. The "enable switches," marked *e*, are transmission gates, tied when the indicated clock signal is high. For data signals (α and β), the sub-index indicates phase of validity. The clamp devices at the output of the gate ensures that the undriven output stays at ground.

As a design exercise, we used the gate style of Fig. 8 for a highly pipelined FIR filter. We believe a filter, specified by transfer function and throughput rate, to be a realistic early application for adiabatic logic circuits: the switching time can easily be increased to allow operation at lower energy levels without changing the specification. This property was also noted by Duncan *et al.*, who demonstrated trading switching time for energy dissipation by varying the supply voltage of a filter implemented with conventional logic [15].

The adder is the central combinational building block of a FIR filter. Following the example of Duncan, we designed a bit-level-pipelined adder circuit to allow the largest possible switching time for a given filter throughput. Fig. 9 shows a gate-level diagram of the adder, complete with the delay cells needed to preserve a bit of information until it may be reversed. The three-bit version shown in the figure was fabricated and its functionality verified; the extension to a wider version is straightforward. To indicate the total circuitry overhead caused by the use of reversibility, those elements that would be needed in a conventional adder have been shaded in Fig. 9. Because

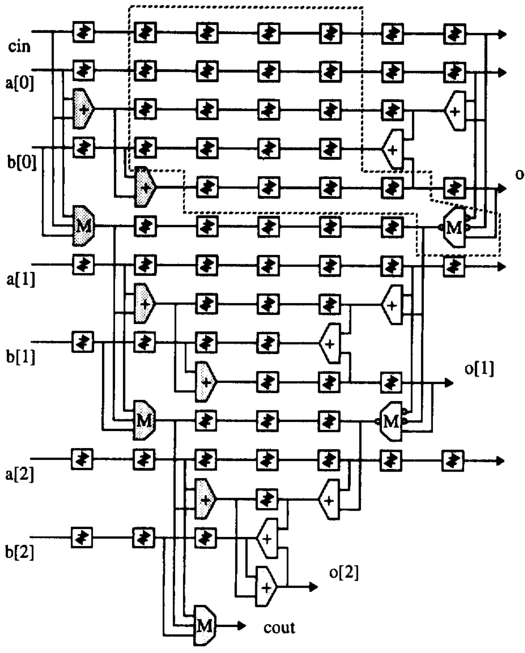


Fig. 9. Pipelined ripple-carry adder, implemented with reversible logic. Each circuit node is driven by two gates: one for charging and one for discharging. Blocks marked "+" are XOR gates; "M" denote majority gates; and bidirectional delay cells are shown with reciprocating arrows. The carry chain mandates a large number of delay cells. Only the shaded elements would be present in a conventional adder.

of dual-rail signalling, separate charge-up and charge-down paths, and the use of T-gates for switches, each gate is also more complex than its conventional counterpart. As a result, the three-bit reversible adder requires 20 times the number of devices and 32 times the area of a conventional adder using the same technology and laid out by the same designer. The overhead clearly gets worse for a wider adder, since the number of delay cells grows as the square of the adder width.

We may decrease the hardware overhead of the adder significantly by using reversibility more judiciously. For example, by allowing at least part of the signal energy of the carry signal from the least significant bit in Fig. 9 to dissipate (such as by using a passive diode charge-down path), we may remove all the elements within the dashed frame. Furthermore, the dissipation caused by operating these elements may actually outweigh the signal energy of the carry signal, unless the adder is operated *very* slowly. Layout extraction and circuit simulation of the reversible adder (to account for parasitics) indicate that a four-bit adder exhibits lower dissipation than two cascaded two-bit adders only if the charging times are thousands of times slower than those possible in a conventional adder. This comparison assumes that the energy of the non-reversed carry signal interconnecting the two two-bit adders is completely dissipated. Charging has to be much slower yet for a 32-bit adder to outperform two cascaded 16-bit adders.

It may be argued that the bit-level-pipelined adder is a particularly bad example for reversible computing because of the long carry-chain dependency. However, a version using carry-lookahead over two bits fares no better because of increased capacitance in the lookahead network itself. Similar assessments apply to other modified versions that retain the basic structure of Fig. 9. We conjecture that new breakthroughs in the composition of reversible logic gates are necessary for a fully-reversible 16- or 32-bit pipelined adder to be competitive with circuits that use reversibility on a lesser scale.

V. Conclusion

With the adiabatic switching approach, circuit energies are conserved rather than dissipated as heat. Depending on the application and the system requirements, this approach can sometimes be used to reduce the power dissipation of digital systems. For situations where the voltage swing is given by external constraints, adiabatic switching is the only known scalable approach to trade off energy dissipation for switching speed. Since the reduction in switching energy is linear in supply voltage (13), adiabatic switching is a promising approach for circuits that drive large capacitive loads to voltages above four times the threshold voltage.

The situation changes when supply voltages can be freely chosen to minimize dissipation for a given performance level. The optimum voltage swing for T-gate-based adiabatic circuits is close to $4V_{\text{th}}$ regardless of switching speed, whereas conventional circuits can reasonably operate with V_{dd} approaching V_{th}. This corresponds to an energy dissipation factor of 16 in favor of the conventional approach. However, the dissipation of adiabatic switching is not limited by the $C_L V_{\text{th}}^2$ asymptote of (18), and for sufficiently slow switching, the adiabatic circuit will operate at lower power levels than its conventional counterpart.

Other sources of overhead must also be considered when analyzing the energy efficiency of adiabatic circuits. The shape factor of the input waveform (ξ) accounts for an additional 23% of dissipation when sine-shaped currents (such as those generated by resonant power supplies) are used. The requirement for dual-rail signals and T-gates where single-rail signals and individual FET's could otherwise be used roughly doubles the area required for logic. With dual-rail signals, half of the circuit nodes will switch each cycle, while for single-rail signals, this number would be the worst case. Also, a significant part of the additional area will contribute to load capacitance that would not be present in a conventional circuit. A final source of overhead are the timing constraints placed on the inputs and outputs.

In total, the circuit overhead for combinational logic is at least 250%, not including the difference in duty factor between the dual-rail and single-rail signals or the added load capacitance due to the additional dual-rail circuitry. To compensate for this, switching time must be increased proportionately. If adiabatic logic is driven by conventional logic, the minimal increase in switching time will be greater because of the $T^{-1/2}$ and $T^{-3/4}$ scaling properties.

The desirable T^{-1} scaling, which can only be achieved with all-adiabatic operation, mandates that the logic be fully reversible. Requiring that inputs be held constant across multiple stages of cascaded logic would preclude using adiabatic switching for complex logic functions. Pipelining is possible and has been demonstrated in practice; if each combinational logic block is replaced by two, one for charging the output and one for discharging, the inputs may change while the output is valid. The inputs for the discharge block must then be driven by the outputs of the following pipeline stage. The stages can be sequenced by overlapping clock signals, which also supply energy to the circuits, to implement reversible pipelines. The overhead in logic is approximately a factor of two because of the need for separate charge and discharge blocks. However, since these blocks are activated at different times, each path is used only once for each full compute cycle. Hence, the energy dissipation is comparable to that of the single block which is used for charging in both directions.

Finally, the use of only reversible logic for fully-adiabatic pipelines puts additional constraints on logic implementation. This is especially noticeable when adiabatic switching and reversible logic are applied to logic functions with sequential dependencies across many stages, as in the example of the carry chain for a bit-level pipelined adder. The overhead of storing the intermediate bits so that they may be later reversed grows as the square of the adder width. Maintaining these bits does not contribute to the computation proper but only to the overhead of performing the computation reversibly. The energy needed for temporarily storing the bits in reversible latches may be greater than the energy the arrangement was supposed to recover.

In summary, if adiabatic switching is to be useful for realizing low-power digital systems in practice, it will most likely be in hybrid configurations which include conventional or partially adiabatic [6] latches and logic. The non-adiabatic latches and logic will have to operate with the smallest energy levels possible, since these energies will be dissipated in all or part. As was speculated about the applications for hot-clock nMOS [1], those parts of a chip which involve driving large capacitive loads, i.e., pads, data buses, and globally decoded signals, are the most promising candidates for adiabatic switching techniques. The analyses and experiments presented in this article are a first step towards understanding where these techniques can be practically used.

ACKNOWLEDGMENT

The authors wish to thank Prof. S. Mattisson of Lund University for many helpful discussions.

REFERENCES

[1] C. L. Seitz, A. H. Frey, S. Mattisson, S. D. Rabin, D. A. Speck, and J. L. A. van de Snepscheut, "Hot-clock NMOS," in *Proc. 1985 Chapel Hill Conf. VLSI*, 1985, pp. 1–17.

[2] W. C. Athas, J. G. Koller, and L. "J." Svensson, "An energy-efficient CMOS line driver using adiabatic switching," *Proc. Fourth Great Lakes Symp. VLSI Design*, pp. 196–199, Mar. 1994.

[3] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circ.*, vol. 27, no. 4, pp. 473–484, April 1992.

[4] C. A. Mead and L. Conway, *Introduction to VLSI systems.* Reading, MA: Addison-Wesley, 1980.

[5] J. S. Hall, "An electroid switching model for reversible computer architectures," in *Proc. ICCI'92, 4th Int. Conf. on Computing and Information*, 1992.

[6] J. G. Koller and W. C. Athas, "Adiabatic switching, low energy computing, and the physics of storing and erasing information," in *Proc. Workshop on Physics and Computation*, PhysCmp '92, Oct. 1992; IEEE Press, 1993.

[7] C. H. Bennett, "Logical reversibility of computation," *IBM J. Res. Dev.*, vol. 17, pp. 525–532, 1973.

[8] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Dev.*, vol. 5, pp. 183–191, 1961.

[9] S. G. Younis and T. F. Knight, "Practical implementation of charge recovery asymptotically zero power CMOS," in *Proc. 1993 Symp. on Integrated Syst.*, MIT Press, 1993, pp. 234–250.

[10] C. H. Bennett and R. Landauer, "The fundamental physical limits of computation," *Scientific American*, pp. 48–56, July 1985.

[11] R. Merkle, "Reversible electronic logic using switches," *Nanotechnology*, vol. 4, pp. 21–40, 1993.

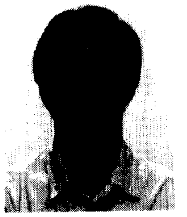
[12] C. H. Bennett, "Time/space trade-offs for reversible computation," *SIAM J. Computing*, vol. 18, pp. 766–776, 1989.

[13] E. M. Cherry and D. E. Hooper, *Amplifying Devices and Low-Pass Amplifier Design.* New York: Wiley, 1968.

[14] A. Kramer, J. S. Denker, S. C. Avery, A. G. Dickinson, and T. R. Wik, "Adiabatic computing with the 2N-2N2D logic family," in *1994 Symp. VLSI Circ.: Digest of Tech. Papers*, IEEE Press, June 1994.

[15] P. J. Duncan, S. Swamy, and R. Jain, "Low-power DSP circuit design using bit-level pipelined maximally parallel architectures," in *Proc. 1993 Symp. on Integrated Syst.*, MIT Press, 1993, pp. 266–275.

[16] D. Maksimovic, "A MOS gate drive with resonant transitions," in *Proc. IEEE Power Electron. Specialists Conf.*, IEEE Press, 1991, pp. 527–532.

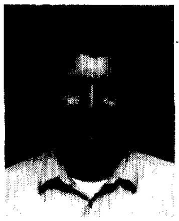


William C. Athas received the B.S. degree in computer science in 1978 from the University of Utah. He received the M.S. and Ph.D. degrees, also in computer science from the California Institute of Tecnology in 1984 and 1987, respectively.

From 1988 to 1989, he was with the University of Texas at Austin where he was an Assistant Professor with the Department of Computer Science. From 1989 to 1991, he was a Research Scientist at the Northrop Research and Technology Center in Los Angeles, CA, where he was engaged in the study of low-power computing techniques for applications in high-density embedded processing environments. During this time he also taught classes at the University of California, Irvine, in computer architecture and VLSI design. Currently he is with the Enterprise Integration Systems Division of the University of Southern California's Information Sciences Institute. In addition to his research into the theory and practice of adiabatic computing, he is involved in the hardware and software design of physically-compact, high-performance parallel computers and teaches VLSI design with the Department of Electrical Engineering at USC.

Lars "J." Svensson (S'86-M'89) received the B.Sc. degree in electrical engineering in 1983 and the Ph.D. degree in applied electronics in 1990, both from Lund University, Sweden. During his Ph.D. work, he was a regular visitor at the University of California at Berkeley.

From 1990 to 1992, he was with the VSDM division of IMEC in Leuven, Belgium. He joined USC/ISI in the beginning of 1993. He is currently with the Enterprise Integration Systems Division at ISI. His main research interest is the monolithic integration of electronic circuits, especially the power dissipation aspects.

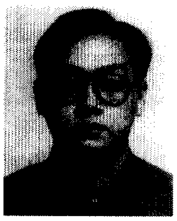


Jeffrey G. Koller received the B.Sc. degree in physics, mathematics and applied mathematics in 1978 and the B.Sc.(Hons) degree in physics in 1979, both from the University of Witwatersrand in Johannesburg. He received the Ph.D. degree in theoretical physics from the California Institute of Technology in 1984.

He spent three years at the Institute for Theoretical Physics at the State University of New York, Stony Brook, two years with the Caltech Concurrent Computation Program, and two years with the Northrop Research and Technology Center in Los Angeles, CA, before joining ISI in 1991. Currently, he is with the Advanced Systems Division at ISI, where his research interests are low power computing, device physics, and parallel hardware and software architectures.

Nestoras Tzartzanis was born in Volos, Greece, in 1966. He received the B.Sc. and M.Sc. degrees, both in computer science, from University of Crete, Greece, in 1988 and 1991.

From 1991 to 1992, he worked on the CPU design of a super-scalar multiprocessor in the Advanced Computer Research Institute, Lyon, France. Since 1993, he has been a Ph.D. student in computer engineering at the University of Southern California and a Research Assistant at USC/ISI, where he works on the application of adiabatic switching on digital systems. His research interests also include computer system architecture, VLSI design, and compilation techniques.



Eric Ying-Chin Chou was born in Hsin-Chu, Taiwan in 1968. He received the B.Sc. degree in computer science and information engineering from National Taiwan University in 1990 and the M.Sc. degree in electrical engineering in 1993 from the University of Southern California.

From 1990 to 1992, he worked as a weaponry officer in Taiwan Navy. He is currently a Ph.D. student in Computer Engineering at USC.