Adiabatic CMOS: Limits of Reversible Energy Recovery and First Steps for Design Automation

Ismo Hänninen(✉), Gregory L. Snider, and Craig S. Lent

Center for Nano Science and Technology, University of Notre Dame,
Notre Dame, IN 46556, USA
{ismo.hanninen,snider.7,lent}@nd.edu

Abstract. Standard CMOS technology discards all signal energy during every switching cycle, leading to heat generation that limits the operating speed and the achievable computing performance. Energy-recovery schemes avoid the heat generation, but are often burdened with the cost of significant increase in system complexity and the lack of automated design tools. In this paper, we propose to implement adiabatic CMOS circuits utilizing split-level rails and Bennett clocking, which enable energy-recovery in standard CMOS logic gates with only minor modifications. Using a pessimistic 32 nm bulk MOSFET technology model, a switching energy improvement factor of approximately 10X can be reached over standard CMOS, while we predict that emerging low-leakage transistor technologies potentially enable adiabatic energy improvements up to four orders-of–magnitude over the standard approach. The significant end-result of our method is that we can leverage the huge number of existing standard gate libraries and logic designs for energy-recovery circuits. We outline an approach to integrate the automatic generation of the adiabatic circuits into the standard circuit design flow, including standard gate logic synthesis and place-and-route.

Keywords: CMOS circuit design · Adiabatic charging · Reversible computing

1 Introduction

The computing performance offered by integrated circuits is tightly connected to energy-efficiency, and this invariant fact remains valid throughout the future predictions of the International Technology Roadmap for Semiconductors (ITRS) [1]. Energy use limits the computing performance through heat dissipation and practical considerations regarding cooling and the overall system electricity bill. Fundamentally, only developing devices and circuits with smaller energy loss can increase computing power. The roadmap predicts smaller transistors and switching energies, but the 40-years-old, and extremely successful, *static complementary transistor* approach, as currently embodied in the complementary metal-oxide-semiconductor (CMOS) technology, has a fundamental flaw: during a logic cycle like "0" → "1" → "0", *all signal energy* in the specific circuit node is irreversibly dissipated as heat. The signal energy must always be significantly higher than the noise floor, which sets a lower limit to the dissipation of the circuit.

© Springer-Verlag Berlin Heidelberg 2014
M.L. Gavrilova et al. (Eds.): Trans. on Comput. Sci. XXIV, LNCS 8911, pp. 1–20, 2014.
DOI: 10.1007/978-3-662-45711-5_1

Standard complementary transistor circuits can be improved by utilizing a reversible energy-recovery scheme, where the signal energy (charge) is adiabatically transferred between the supply rails and the internal circuit nodes. This involves a tradeoff between the switching speed and the energy, and typically results in a considerable circuit overhead [2]. While the general adiabatic logic style has been known for more than 20 years, the above-mentioned challenge has prevented wide utilization. However, standard CMOS has for the last 10 years been trading off speed for heat. For example, the 32 nm bulk MOSFET technology in this paper could easily switch at a frequency of 115 GHz, but heat generation limits the standard circuits to clocks in the range of 1–3 GHz, and the gates in combinatorial blocks are switched at least three orders of magnitude slower than the device speed. Given these new realities, the adiabatic structures previously considered "too slow" are not slow anymore.

In this paper, we construct adiabatic CMOS circuits, which are compatible with the standard CMOS design flow and existing electronic design automation tools. The circuits based on *split-level charge recovery* [3, 4] and *Bennett clocking* [3, 5] are benchmarked in SPICE simulation and compared to the standard CMOS: the switching energy is verified to decrease by a factor of approximately 10X in the pessimistic 32 nm technology, while low-leakage emerging transistors, including tunneling field effect transistors (TFETs), are predicted to offer up to 20,000X adiabatic improvement over a standard approach using the same "steep" transistor devices.

This article is organized as follows: Sect. 2 provides an overview of the power in the transistor circuits, while Sect. 3 describes the reversible clocking approach. In Sect. 4, the model parameters are defined and the circuit speed and leakage characterized, while in Sect. 5, the switching energy is analyzed. Section 6 provides an outline of the proposed automated design flow, while the conclusion follows in Sect. 7.

2 Power in Complementary Transistor Circuits

Standard CMOS logic style uses complementary N- and P-type transistors to construct static logic gates, which can be utilized also in the adiabatic CMOS. In the following, we give a short introduction into power dissipation in both CMOS variants.

2.1 Standard Static CMOS

Today's computers encode information using charge stored on the CMOS gate and interconnect capacitors. Power dissipation for standard CMOS is

$$P_{Total} = N(\gamma C V_{DD}^2 + P_{Passive}) \tag{1}$$

where γ is the activity factor, V_{DD} is the supply voltage, C is the load capacitance at the output of each logic gate, N is the number of gates and f is the operating frequency. The first term represents the active power, *i.e.,* the power dissipated in processing information. The second term, the passive power dissipation, is power that is simply wasted because a voltage is applied to the circuit. The dominant cause of passive power is the

subthreshold leakage current when the transistor is in the off state. Gate leakage has become less of an issue with the advent of high-k dielectrics.

Equation (1) highlights the twin problems faced by the CMOS electronics industry. The ITRS Roadmap [1] projects fully scaled CMOS to have a device density of 10^{10} cm^{-2}, a switching speed of 12 THz, and a switching energy of 3 aJ (750 k_BT). If all of the devices on such a chip were switched at full speed, the power dissipation of the chip would be approximately 150 kW/cm^2, and that is just the *active* power. Even lowering the switching energy to 100 k_BT, a practical limit resulting from the noise floor, will only reduce the active power to 20 kW/cm^2. Clearly the processing of information using the current methods does not provide a path to ultra-high-density high-speed computation where all devices are switched at their maximum frequency.

Historically, the primary approach the industry has taken to lower dissipation has been to lower V_{DD} so that the active power is lowered. This has worked well for many years because, as Eq. (1) shows, the active power is quadratically dependent on V_{DD}. This approach comes at a price, however, because the passive power (the second term in (1)) is exponentially dependent on the transistor threshold voltage V_{th}, which in turn is bounded by V_{DD} (V_{th} is typically 1/4 – 1/3 V_{DD}, where higher V_{th} leads to lower speed). Thus, lowering V_{DD} lowers the active power quadratically, but raises the passive power exponentially. The active and the passive power scale as functions of the gate length, and the passive power increases rapidly with the scaling. Because passive power increases exponentially when V_{DD} is decreased, V_{DD} cannot be scaled much below 1 V. Because of the linear increase in the active power with f, we have seen a virtual halt to the decades-long steady increase in clock frequencies. The practical limit for air-cooling is still about a constant 100 W/cm^2. The semiconductor industry is pursuing the following three approaches to alleviate the dissipation problem:

1. One can try to alter the connection between the supply voltage, threshold voltage, and leakage current by changing the physics of the on/off transition so that it is inherently more abrupt. By developing so-called "steep devices", transistors with a subthreshold slope steeper than the thermal 60 mV/decade, the hope is that a lower OFF current could be had for a given threshold. Then, V_{DD} could be further lowered to reduce the active power, but this is limited by the 100 k_BT practical limit for the switching energy. [6, 7]
2. One can try to improve performance through parallelism rather than single processor speed, introducing multi-core architectures. This amounts to a wager that *software innovations* will finally find a way to efficiently use many processors to accomplish a single task, in effect defeating Amdahl's law. David Patterson, winner of the 2008 IEEE/ACM Eckert-Mauchly Award, has called this wager the "Hail Mary pass" of the industry [8].
3. One can reduce the effective N in Eq. (1) by not using all of the gates, turning off areas of the chip that are unused. This approach is known as "dark silicon" and represents our inability to use existing circuit resources. [9]

It is important to understand the fundamental limits of dissipation in computation. In 1961 Landauer [10] postulated that energy must be dissipated as heat only when information is destroyed, an idea that has come to be known as the Landauer Principle (LP). The minimum amount of energy is related to a quantity known as the Ultimate

Shannon Limit [11], $k_BT \ln(2)$, the minimum energy to make a bit distinguishable from noise. If information is not destroyed, there is no fundamental lower limit to the dissipation in computation, just practical limits, which we analyze in this paper.

The electronics industry is locked into the dissipation limitation by the standard CMOS circuitry, which destroys information contained in the logic gate signal at every switching event. In a logic gate, the information is represented by an energy

$$E_{Bit} = \frac{1}{2} C V_{DD}^2 \tag{2}$$

stored on the capacitor C, and this entire amount is dissipated as heat twice in a switching cycle. The standard CMOS circuits unavoidably destroy information, so the only way to limit the active dissipation is to reduce the energy in a bit (reduce V_{DD}) or limit the rate at which bits are destroyed (limit f). Both methods have their downsides.

2.2 Adiabatic CMOS

There is another approach. Reversible computing with adiabatic clocking can reduce the active power and break the connection between active and passive dissipation. This approach offers energy efficiencies which are orders of magnitude better than current computational paradigms. Reversible or adiabatic computing is an idea that was proposed many years ago, but it has faced criticism that it is "slow," trading clock speed for dissipation, as well as assertions that it simply cannot reduce dissipation. However, since the industry has stopped increasing the clock speed, a trade-off of clock speed is already being made. Reversible designs once considered too slow can become attractive, and our experiments have shown that power savings are possible.

How much can adiabatic switching reduce the power dissipation? Power in an adiabatic CMOS system is given by the equation:

$$P_{Total} = N\left\{\gamma C V_{DD}^2 f\left[\alpha\frac{f}{f_o} + (1-\alpha)\right] + A\exp\left(-\frac{qV_{DD}}{4\eta kT}\right)\right\} \tag{3}$$

where N is the total number of logic gates in the system, α is the fraction of gates that are switched adiabatically, f_o is the characteristic frequency defined by the RC time constant of the gate, A is a constant, and η is the "ideality factor" for the subthreshold slope (η = 1 gives the ideal 60 mV/decade). As before, the active power depends on $CV_{\mathrm{DD}}^2 f$, but has now split into two terms. The factor α is the fraction of the system that is logically reversible, where in a fully reversible design α = 1 and no information is destroyed. These reversible transitions are described by the first term of the active power, where the adiabatic reduction factor f/f_o is a measure of how much of the bit energy is lost to heat. Our experiments have shown that this energy can be very small indeed. In a practical system, α will be less than 1, and the power dissipation due to the destroyed bits is given by the second term in the active power. Passive power, the last term in Eq. (2), has an exponential dependence on the transistor threshold voltage, and in the equation the threshold voltage is set to $V_{\mathrm{DD}}/4$.

Figure 1 shows a plot of calculated power dissipation vs. frequency for a standard CMOS system in which static, or passive, and dynamic power are equal at a critical frequency f_c = 10 GHz. Here we assume that f_0 = 100 GHz, V_{DD} = 1 V. The figure contrasts the standard circuit with a moderately reversible (α = 0.85) circuit with V_{DD} = 2 V. Because the active power is lower due to adiabaticity, V_{DD} can be raised enough to greatly suppress the static power. Static power for the reversible circuit is too low to be visible on the graph. The lower red line is the total power dissipated by the reversible circuit. The reversible circuit yields total power savings of more than an order of magnitude when f/f_c = 0.1, and more than two orders of magnitude when f/f_c = 0.01. Since the active and passive powers are now decoupled, V_{DD} and V_{th} can be raised to suppress the passive power without a dramatic increase in active power. The only fundamental constraints on V_{DD} will be gate breakdown and source-drain punchthrough, but it should be possible to keep V_{DD} above 1 V through the end of the roadmap. The key point is that the *signal energy can be very high*, and all is not lost.

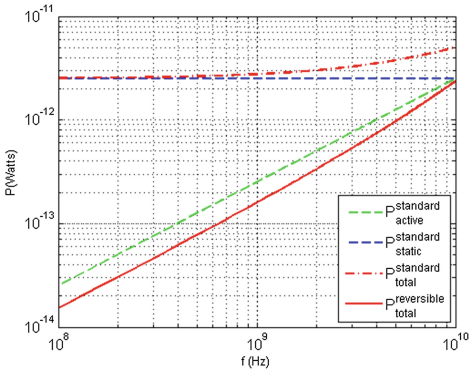

Fig. 1. Power dissipation for standard and reversible CMOS logic.

Interconnects are responsible for a significant fraction of the power dissipation of CMOS systems, making them a prime target for energy recovery. The high dissipation in interconnects is due to the high capacitance of the interconnect lines that leads to large bit energies, which are dissipated at each logic transition. For short interconnects, as within a logic block, the dissipation occurs in the CMOS driver transistors, not in the interconnect wires themselves since the transistors represent the dominant resistance. In this case the simple application of the adiabatic approach can dramatically reduce the power dissipation. For longer interconnect runs, the key factor is again the adiabatic reduction factor f/f_o. If the switching frequency is kept below f_o then power savings will be realized. Therefore, the resistance of the interconnect should be minimized and long wires should be avoided, but the interconnects are the best target for energy savings. As a related example, resonant clocks are being explored as a means to reduce dissipation in clock distribution networks [12].

Adiabatic CMOS provides a way to leverage existing technologies in ways that can greatly reduce the power dissipation. Because it attacks the active dissipation, any

transistor type can be used in the adiabatic circuits, such as tunnel FETs, III-V and graphene channel FETs. In fact, a "steep" device used in adiabatic CMOS could be used to address the issues of both active and passive dissipation. When reversible circuits have been investigated in the past, they were dismissed as slow. The time has come to re-examine reversible CMOS in light of the new end-of-scaling realities. Using these techniques one can take whatever transistors the industry can produce and greatly reduce the power dissipation compared to the standard CMOS.

3 Bennett Clocked Reversible Logic Circuits

Energy-recovery is possible only if specific information exists to control the process. In adiabatic transistor circuits, this information tells where the charge of a logic signal should be returned when the signal is relaxed, in an *erase-with-a-copy* operation. There are two approaches to provide this information: 1. Reversible logic gates using bi-directional double pipelines with the cost paid in gate complexity [13–15], and 2. Bennett clocking, which is used in retractile cascade circuits. For a classification of adiabatic logic circuits, the reader is referred to [2].

We chose to implement the split-rail Bennett clocked reversible approach to retain compatibility with the existing and extrapolated CMOS technology and to gain a high level of design automation with only relatively minor additions to a standard design flow. Basically, the top-level decision is to choose whether to use split-rails with a "null" voltage in the middle and standard CMOS type single-output gates, or a single-rail-to-ground approach with complementary output gates with more complex structure and typically twice the transistor count. Asymptotically adiabatic logic (the best class of energy-recovery circuits) requires three-state signal levels, which can be achieved without any internal circuit overhead using the split-rail approach. While this is excellent from the spatial design point-of-view, some penalty will be paid in more complex timing, which requires the Bennett clocking approach.

Split-Level Charge Recovery. Adiabatic CMOS requires that the output nodes of a logic gate are energized to the full signal level "1" or "0" and also de-energized back to the relaxed voltage level, which can overlap with one of the logic levels or be situated between of them. We chose to utilize three distinct levels {"1", relaxed "null", "0"} and two opposite ramped power-clock rails supplying the pull-up and the pull-down network of a standard static complementary transistor logic gate, as illustrated by the 1n1p-logic inverter in Fig. 2. While a standard CMOS two-level voltage convention would enable the energizing of the output node to the full logic level, the de-energizing step through the same specific transistor requires the three-level voltage convention to recover all charge. For example, an output energized to the logic level "1" through a PMOS transistor must be completely de-energized through the same device. If a single power clock is used the PMOS transistor would turn off prematurely below the threshold voltage. This is avoided if the output is de-energized only to the intermediate "null" level, while the gate is at "0" [3, 4].

Bennett Clocking. The energy recovery scheme requires that the inputs of a logic gate are driven to the correct value also during the de-energization, to select which of the two opposite rails is used to recover the charge. Bennett clocking is used to implement a retractile cascade circuit, in which all previous stages are retained in the energized state until the current stage has de-energized. Basically, the inputs of a logic gate hold on to their values and this information is used to control the de-energization, operating any standard CMOS gate reversibly. The cost is paid in timing, trading of speed for energy, with the consequence that circuits will have to be divided into several blocks, with every boundary presenting irreversible energy loss [3, 5].

A combinatorial block consisting of n logic levels requires n pairs of the power-clocks as illustrated in Fig. 2. One computation consists of a *compute phase* and an *un-compute phase*, during which the primary inputs must stay stable at the logic value "0" or "1", driven for example by a standard CMOS latch. [3, 5, 13, 14, 16].

During the compute phase, all power-clocks and nodes inside the combinatorial block begin in the relaxed "null" state. Starting from the input side, the first stage is energized by ramping active the corresponding pair of power-clocks. After that, the second stage has valid input and can begin to energize similarly. One by one, each stage gets energized until the valid output of the last stage can be stored in a latch.

During the de-compute phase, all power-clocks and nodes inside the combinatorial block begin in the energized state, which is identical to the situation in the standard CMOS. First, the last pair of power-clocks is ramped inactive and the last stage de-energized, while all previous stages are held energized and retain valid logic values. One by one, from the block output towards the input side, each logic stage gets similarly de-energized, always keeping the previous stages energized and providing the correct gate input to select which power-clock rail is connected to recover the charge. Finally, the first logic stage is de-energized and a new cycle can begin.

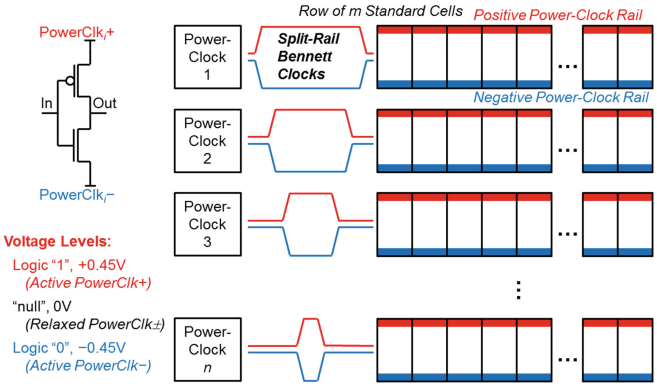

Fig. 2. Split-rail Bennett-clocked adiabatic CMOS: an inverter implemented with standard complementary transistors, the voltage levels, and a conceptualized combinatorial block of n logic levels, each energized by the corresponding pair of complementary power-clocks.

4 Model Characteristics (32 nm Bulk MOSFETs)

Asymptotically adiabatic recovery of the switching energy is possible by slowing down the ramp time and using ever larger reversible block sizes. In practice, leakage currents prevent achieving the theoretical efficiencies. We constructed the test circuits using standard 32 nm bulk MOSFETs, based on the Nano-CMOS SPICE models obtained from the Predictive Technology Model (PTM) library of the Arizona State University [17]. The models are pessimistic, with high leakage dominating the power.

The 32 nm bulk MOSFET model has a nominal operating voltage V_{DD} = 0.9 V, NMOS threshold voltage $V_{\mathrm{th,n}}$ = 0.23 V, and PMOS threshold voltage $V_{\mathrm{th,p}}$ = -0.23 V. The NMOS body is biased to 0 V and PMOS body to 0.9 V. For standard static CMOS, the signal levels are 0 V (logic "0") and 0.9 V (logic "1"). For adiabatic CMOS, we shift the relative voltages to better represent the concept of positive and negative rails: the swing of the positive power-clock is 0–0.45 V, the swing of the negative power-clock is 0–(−0.45) V, and the corresponding signal levels are −0.45 V (logic "0"), 0 V ("null"), 0.45 V (logic "1"). The NMOS body is biased to −0.45 V and PMOS body to 0.45 V. This convention does not affect the model operation. The simulator suite used in this work is LTspice IV, version 4.20i [18].

4.1 Load Circuit and Energization

Leakage power, switching speed, and switching energy were determined in parameterized load circuits, customized for standard CMOS and adiabatic CMOS separately. This approach enabled realistic loading scenarios, taking into account the different timing and energization states in the complementary transistor circuit variants. The differences between the load circuits mirror exactly the differences between real circuits, and the comparison between standard and adiabatic CMOS is valid.

The circuit operation was characterized in a chain of 12 inverters and in an 11-stage ring oscillator, with all the inverters sized identically in the fan-out-of-one (FO1) configuration of each stage. For both the standard and the adiabatic CMOS, the transistor channel length was $l = 32$nm, the smallest NMOS width $w_{\mathrm{NMOS,min}} = 32$nm, and the smallest PMOS width $w_{\mathrm{PMOS,min}} = 2.8 \times 32$nm. The pull-up and pull-down transistors were sized to sink/source approximately the same maximum absolute saturation current, with a constant PMOS/NMOS width ratio $w_{P/N} = 2.8$. The transistor widths were uniformly scaled to larger multiples of the minimum, with the NMOS width $w_{\mathrm{NMOS}} = k \times 32$nm and the PMOS width $w_{\mathrm{PMOS}} = 2.8 \times w_{\mathrm{NMOS}}$, where the parameter k was stepped across the range $k = 1 \ldots 1000$. Circuits consisting of the minimum devices to circuits having 100X the minimum size were fully characterized, while larger sizes were only partially tested.

Standard CMOS. The standard CMOS logic gates are always driving a fully energized load: the following logic gates have their supply rails at V_{DD} and ground. As a consequence, all off-transistors have subthreshold leakage all the time. During a switching event, the gate output transitions through the complete voltage range 0 V–0.9 V. The chain of inverters in Fig. 3 was used to determine the in-circuit input

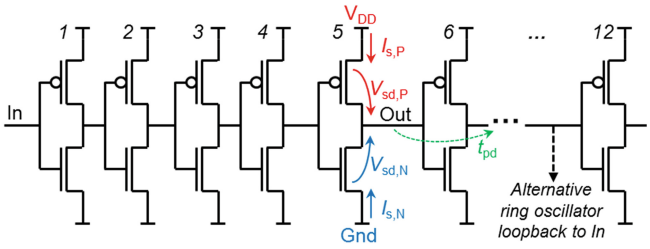

Fig. 3. Standard CMOS test circuit: FO1 inverter chain/ring oscillator.

and output edge rates, the propagation delays between the stages, and the switching energy, with the chain representing a realistic loading-of-load. These were verified in the ring oscillator with a characteristic frequency and uniform edge rates.

Adiabatic CMOS. The adiabatic CMOS logic gates have their load in an un-energized (relaxed to null) state before and during the energization ramp. During the energization ramp, the gate output transitions through half of the full voltage swing, for logic "1" the range 0 V–0.45 V and for logic "0" the range 0 V–(−0.45 V). After energization of the current stage and the following stages, the static voltages across all terminals correspond to the standard CMOS and result in identical subthreshold leakage. During the de-energization ramp (relaxation), the load gate is already in the relaxed state, and the output transitions to 0 V through half of the full voltage swing.

Switching energy was characterized using identical transistors and the same sizing as in standard CMOS, in a test setup illustrated in Fig. 4, consisting of the driver inverter and relaxed load and load-of-load inverters, which is the realistic situation during the energization and de-energization ramps. During the examined process, all power-clocks start in the relaxed state (0 V) and the input value to the driver gate is set to a fixed stable logic "1" (0.45 V) or "0" (−0.45 V). Following that, the power-clocks of the driver are ramped active, while the power-clocks of the load and the load-of-load stay relaxed. After energization, the driver output has a valid logic state and the load can begin switching, which we omit from this measurement, since the static state is exactly the same as in the standard CMOS. Instead, we keep the loads relaxed and proceed to ramp inactive the power-clocks of the driver, which normally would happen after the following stages have cycled through energization and de-energization.

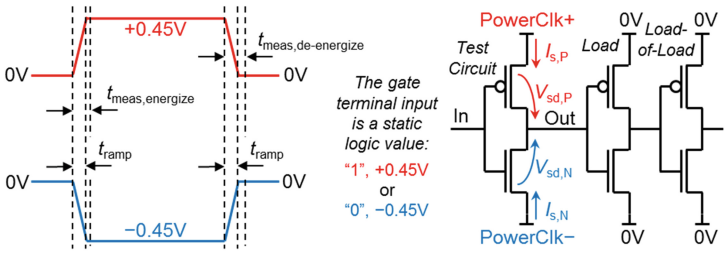

Fig. 4. Adiabatic CMOS test circuit: the power-clocks voltage levels and timing definitions, and the measured circuit followed by a relaxed load and a relaxed load-of-load.

4.2 Switching Speed and Leakage

Standard CMOS. The 32 nm bulk MOSFET transistors are fast, having a standard CMOS ring-oscillator in-circuit propagation delay under 9 ps. This, in theory, is the limit on how fast the circuits can transfer data, while the signal edge rate is 7.2 ps (20 %/80 % signal levels), extrapolated to a full swing transition time of approximately 12 ps. A standard circuit could have up to a 115 GHz clock frequency, or 38 logic levels at a 3 GHz clock. However, excessive heat generation prevents reaching these values.

Adiabatic CMOS. The above limits for the device speeds apply, but the gate input and power-clock ramps are purposefully kept slow enough to limit the resistive losses. The fastest possible circuit with a relaxed load, a prepared stable gate input, and ideal triangle power-clocks swinging without any hold or relax plateaus would fully switch the output with extremely short power-clock ramp times: a load of $k = 1$ could run at the corresponding frequency 20.4 GHz, $k = 10$ at 20.0 GHz, $k = 100$ at 17.5 GHz, and $k = 1000$ at 0.78 GHz. Of course, a circuit this fast could not be called adiabatic.

Leakage. The ON/OFF ratio determined as $I_{\mathrm{d,sat}}/I_{\mathrm{d,cutoff}}$ with nominal voltages between the terminals for the NMOS transistors was about 16500 and for the PMOS transistors about 11500. This indicates that the 32 nm bulk MOSFET model has relatively high leakage, which is shown later to dominate the circuit power.

5 Switching Energy

The baseline for the switching energy is formed by the standard static CMOS with a fixed characteristic switching speed, while the adiabatic circuits were energized and de-energized using a variable ramp time for the power-clocks. The transistor size parameter k was stepped across the range $k = 1\ldots1000$, from minimal to 1000X.

5.1 Standard CMOS

The standard CMOS switching energy was measured in the middle inverter of the FO1 inverter chain with the characteristic edge rate 7.2 ps (20 %/80 % levels), corresponding to a full swing time of 12 ps (transistor sizing $k = 1\ldots100$). However, the circuit does not reach the static state this fast, so the measurements used a window of 100 ps, starting at the beginning of the change in the input signal, to capture all the transients related to a single switching event towards a logic level. The full cycle contains two transitions, cycling the output through the logic values "0" → "1" → "0".

The energy dissipated in a specific transistor during a switching event was computed as a numerical integral of the momentarily time-dependent power $P(t)$:

$$E_{\mathrm{part}} = \int |P(t)|dt = \int |V_{\mathrm{sd}}(t) \times I_{\mathrm{s}}(t)|dt, \tag{4}$$

where $V_{sd}(t)$ is the momentarily potential difference between the source and the drain, and $I_s(t)$ the momentarily source current. The energy dissipated during a logic transition was computed as a sum of the above defined individual transistor energies:

$$E_{rise} = E_{rise,P} + E_{rise,N}, \qquad E_{fall} = E_{fall,P} + E_{fall,N}. \tag{5}$$

The total energy of a full logic cycle "0" → "1" → "0" was summed up as

$$E_{cycle} = E_{rise} + E_{fall}. \tag{6}$$

The resulting switching energies are presented in Table 1. The energy of the full cycle of the minimal FO1 inverter ($k = 1$) is approximately 300 aJ. As expected, the energy scales nearly linearly in the load range $k = 1...100$.

Energy Breakdown. For the typical transistor sizes, approximately 55 % of the full cycle energy is dissipated during the rising transition and 45 % during the falling transition. Consistently, 55 % of the total energy is dissipated in the PMOS transistor and 45 % in the NMOS. During a rising transition, 97 % is dissipated in the PMOS and 3 % in the NMOS, and during a falling transition, 96 % is dissipated in the NMOS and 4 % in the PMOS. The dissipation in the transistor moving into the cut-off region originates mostly from the crowbar current, unavoidable in the standard CMOS.

Table 1. Switching energy in the standard CMOS FO1 inverter, summarized for size factors k.

Sizing k	**Partial Energies (aJ)**				**Transitions (aJ)**		**Total (aJ)**
	$E_{rise,P}$	$E_{rise,N}$	$E_{fall,P}$	$E_{fall,N}$	E_{rise}	E_{fall}	E_{cycle}
	(P→on)	*(N→off)*	*(P→off)*	*(N→on)*			
1	149	5.10	4.50	119	154	124	278
10	1680	49.8	55.3	1360	1730	1420	3150
100	16500	502	647	13300	17000	14000	30900
1000	107000	5290	392000	373000	112000	765000	877000
Linear Model for the Full Cycle Energy (J)			$E_{cycle} = \left(3.09 \times 10^{-16}\right)k + 1.01 \times 10^{-17}$				

* *Transistor sizing range* $k = 1...100$ *has a characteristic edge rate 7.2* ps *(between signal levels 20 %/80 %), and a measurement window of 100* ps *for each transition. Linear model for* $k = 1...100$

5.2 Adiabatic CMOS

The adiabatic CMOS switching energy was measured in an FO1 inverter, driving a similar but relaxed load inverter followed by a relaxed load-of-load, which is the realistic situation during the energization and de-energization ramps. The gate input of the measured inverter was first set to a stable logic value "0" or "1", and following that, the power-clocks of that inverter were ramped from relaxed to active. After reaching a stable output state with the logic value "0" or "1", the power-clocks were ramped back to the relaxed voltage and consequently the output ramped to the "null" state.

The logic "1" cycle contains two voltage ramps, where the output signal transitions "null" → "1" → "null", while the logic "0" cycle contains two ramps and the output transitions "null" → "0" → "null". The full adiabatic cycle comparable to the standard CMOS contains all four transitions: "null" → "1" → "null" → "0" → "null". (The voltage levels are: –0.45 V for logic "0", 0 V for "null", 0.45 V for logic "1".)

The measurements used a varying length time window defined in Fig. 4, beginning at the start of the power-clock ramp and lasting a fixed 200 ps after the end of the energizing ramp and 2 ns after the end of the de-energizing ramp, to capture all transients. The longer after-ramp time was necessary in the de-energizing step, since a residual charge was left unrecovered during the actual ramp, due to a reduced conductance in the transistors compared to the energization step. With the additional time, the voltage relaxed fully to "null". As a consequence, the measured energy of short ramps contains relatively more contribution from the after-ramp time, in comparison to the long ramp times, and this produces a relatively small systematic measurement error. The fully characterized transistor sizing range was $k = 1 \ldots 100$, while the power-clock ramp time t_{ramp} was stepped through six decades from 1 ps — 1 μs, ten data points per decade.

The energy dissipated in a specific transistor during a transition event was computed as a numerical integral of the momentarily time-dependent power as for standard CMOS (Eq. 4). The energy of the logic "1" cycle was computed in two parts:

$$E_1 = E_{\mathrm{en},1} + E_{\mathrm{de},1}, \tag{7}$$

where $E_{\mathrm{en},1}$ is the energy related to the energization transition "null" → "1" and $E_{\mathrm{de},1}$ the energy related to the de-energization "1" → "null", computed by summing up the contributions of the individual complementary transistors:

$$E_{\mathrm{en},1} = E_{\mathrm{en},1,\mathrm{P}} + E_{\mathrm{en},1,\mathrm{N}}, \qquad E_{\mathrm{de},1} = E_{\mathrm{de},1,\mathrm{P}} + E_{\mathrm{de},1,\mathrm{N}}. \tag{8}$$

Similarly, the energy dissipated during the logic "0" cycle was computed as:

$$E_0 = E_{\mathrm{en},0} + E_{\mathrm{de},0}, \tag{9}$$

where $E_{\mathrm{en},0}$ is the energy related to the energization transition "null" → "0" and $E_{\mathrm{de},0}$ the energy related to the de-energization "0" → "null", computed by summing up the contributions of the individual complementary transistors:

$$E_{\mathrm{en},0} = E_{\mathrm{en},0,\mathrm{P}} + E_{\mathrm{en},0,\mathrm{N}}, \qquad E_{\mathrm{de},0} = E_{\mathrm{de},0,\mathrm{P}} + E_{\mathrm{de},0,\mathrm{N}}. \tag{10}$$

The total energy dissipated during the full adiabatic cycle, excluding the static time periods in the energized or relaxed state, contained all four transitions "null" → "1" → "null" → "0" → "null" and was computed by summing the half cycle energies:

$$E_{\mathrm{cycle}} = E_1 + E_0. \tag{11}$$

Energy vs. Ramp Rate. The total dissipated energy and the energy related to the "1" cycle and the "0" cycle separately are shown in Fig. 5. The example is for the transistor

sizing factor $k = 10$, but the scaled curves are similarly shaped for the other sizes, as shown in Fig. 6 on a log-log scale. The shortest (fastest) ramp times in the range 1 ps—10 ps correspond to the abrupt switching of the power-clocks and result in the maximum energy dissipation, while the longer (slower) ramp times up to about 10 ns bring energy savings. With even longer ramp times, the leakage currents cause increased dissipation and the benefits of the slowing down are lost. The lowest full cycle energy 400 aJ of the adiabatic CMOS is reached near the ramp time $t_{\text{ramp}} = 9\text{ns}$. This ramp rate results in minimum energy also for the other explored transistor sizes.

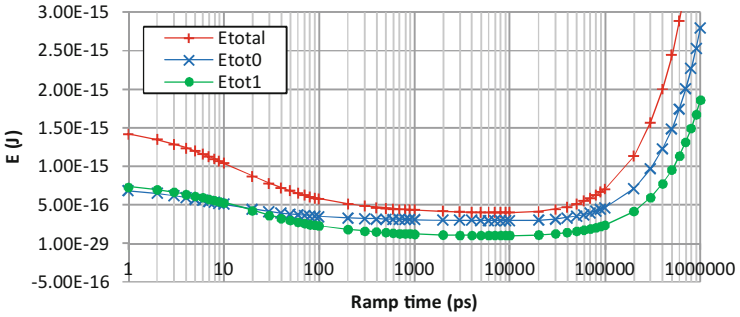

Fig. 5. Switching energy in the adiabatic CMOS FO1 inverter with transistor sizing $k = 10$: total, "1" cycle, and "0" cycle. The corresponding standard CMOS full cycle energy is 3.15 fJ.

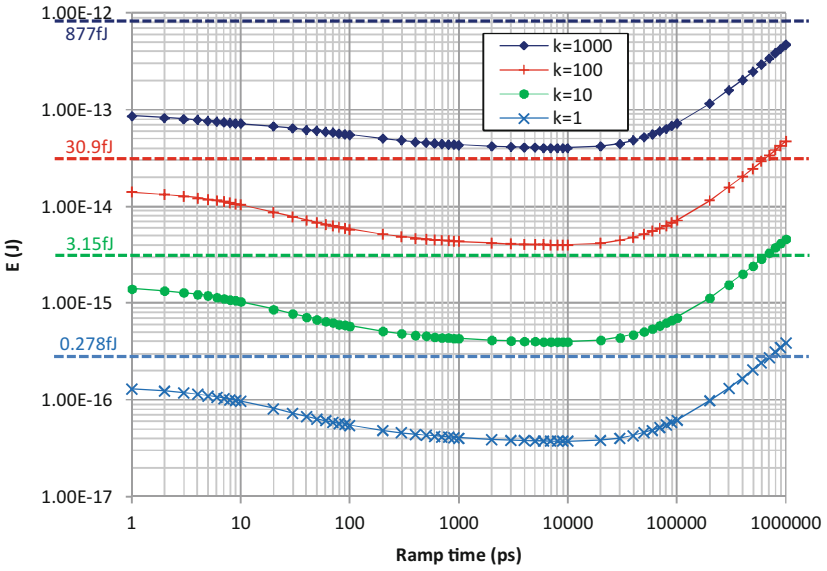

Fig. 6. Total switching energy in the adiabatic CMOS FO1 inverter with transistor sizing $k = 1, 10, 100, 1000$, plotted on log-log scale. The corresponding standard CMOS full cycle energy is marked as a dashed horizontal line above each adiabatic CMOS curve.

Energy vs. Standard CMOS. For the transistor sizing factor $k = 10$, the lowest full cycle energy 400 aJ of the adiabatic CMOS results in approximately 87 % energy savings in comparison with the 3150 aJ full cycle energy of the static CMOS, bringing an improvement factor of 7.9X. Similar improvements are achievable throughout the explored sizing space, with all the minima occurring near the ramp rate $t_{\text{ramp}} = 9\text{ns}$, as summarized in Fig. 7. It should be noted that even with the fastest power-clock ramp rates, the adiabatic CMOS energy is less than half of the standard CMOS energy, and an 82 % improvement can be reached with a ramp rates as fast as 100 ps. Therefore, significant savings are possible even with a 200 ps clock cycle time, containing a ramp up and a ramp down, corresponding to a frequency of 5 GHz.

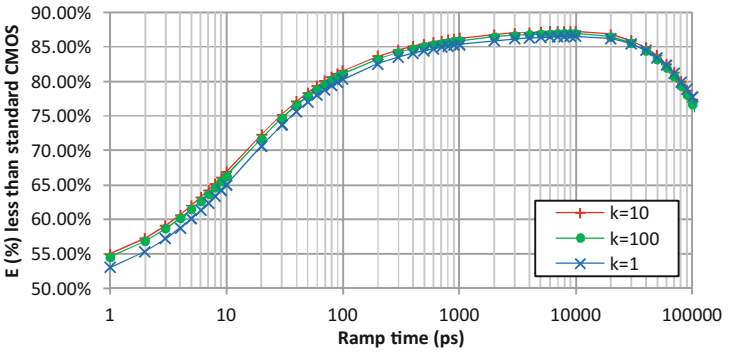

Fig. 7. Relative improvement in the switching energy, adiabatic CMOS vs. standard CMOS, defined as $\left(E_{\text{cycle,standard}} - E_{\text{cycle,adiabatic}}\right)/E_{\text{cycle,standard}}$

Relative Energy of the "1" / "0" Cycles. The absolute and relative energy contributions of the half cycles are shown in Fig. 8 for the transistor sizing factor $k = 10$. With realistic ramp rates above 10 ps, the "0" cycle dissipates more energy than the "1" cycle, which is consistent with the fact that the PMOS in cutoff leaks more than the NMOS in cutoff. At its worst, the "1" cycle contribution forms approximately 50 % of the full cycle energy, and at its best, only 25 % of the energy, exactly at the ramp rate with the minimum overall dissipation. The "0" cycle benefits less from the longer ramp rates. Same applies for all sizing factors k.

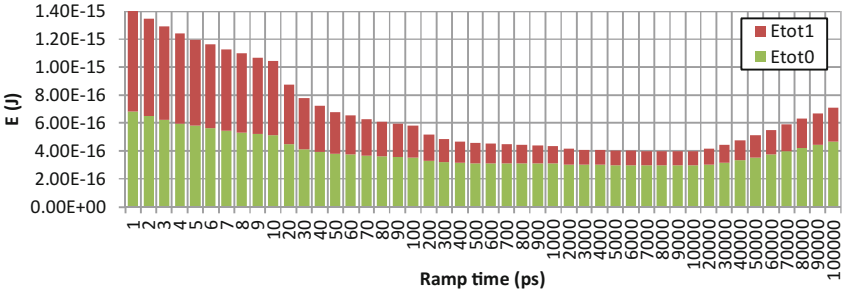

Fig. 8. Absolute switching energy of the "1" cycle and the "0" cycle in adiabatic CMOS with transistor sizing $k = 10$. Total bar height corresponds to the total energy.

Relative Energy in the P/N Transistors. During the full adiabatic switching cycle, the PMOS transistor contributes over 70 % of the total dissipation, with all sizing factors. However, in each separate half cycle, the specific OFF transistor will dominate the energy in slow ramp rates. During the logic "1" cycle, the fastest ramp rates under 10 ps dissipate approximately 85 % of the energy in the active PMOS, while the rates slower than 200 ps dissipate less than 50 % in the PMOS. During the logic "0" cycle the ramp rates under 10 ps dissipate approximately 50 % in the active NMOS, while the rates slower than 200 ps dissipate less than 10 % in the NMOS. ($k = 10$.)

In standard CMOS, both complementary transistors switch at the same time, resulting in a crowbar current which effectively shorts the rails. This does not happen in the adiabatic CMOS, since only one of the transistors switches at a time. However, we still measure a significant energy loss in the specific transistor which should be OFF, as illustrated in Figs. 9 and 10. This passive switching energy originates from the leakage currents, which are inherent to the modeled 32 nm bulk MOSFET technology, and it turns out to dominate the total switching energy, since the voltage drop is large over the OFF transistor. The passive energy prevents the adiabatic CMOS circuits on this technology from reaching more than 10X improvement over the standard CMOS. The passive energy only increases when the ramp rate is slowed down.

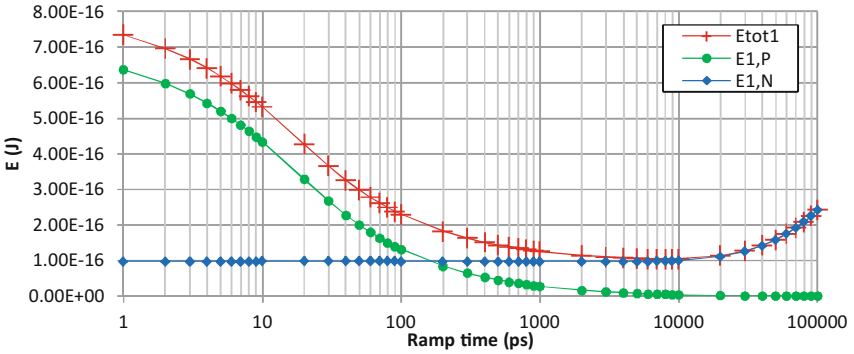

Fig. 9. Logic "1" cycle, switching energy: total, active PMOS, and passive NMOS. The corresponding standard CMOS *half* cycle energy 1.6 fJ is about 16X higher than the minimum.

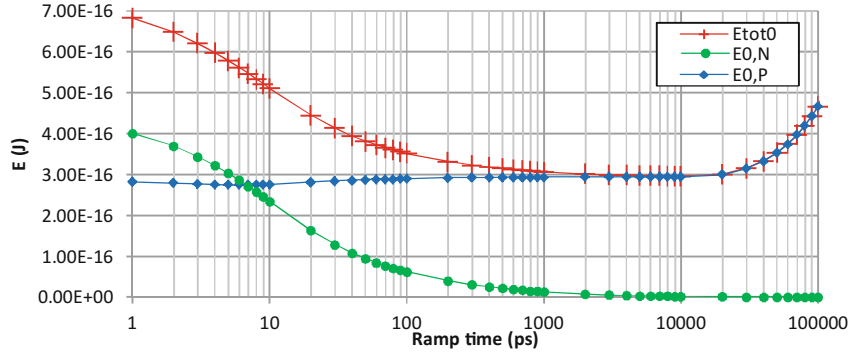

Fig. 10. Logic "0" cycle, switching energy: total, active NMOS, and passive PMOS. The corresponding standard CMOS *half* cycle energy 1.6 fJ is about 5X higher than the minimum.

Energy Using Low-Leakage Transistors. The active switching energy scales well with the slow-down and would enable extremely large energy benefits in the adiabatic CMOS. The energy improvement factor over standard CMOS is defined as:

$$X = E_{\text{cycle,standard}} / E_{\text{cycle,adiabatic}}, \tag{12}$$

where $E_{\text{cycle,standard}}$ is the full cycle energy of the standard CMOS and $E_{\text{cycle,adiabatic}}$ the full cycle energy of the adiabatic CMOS. The improvement is computed separately for the total energy in the active and passive transistors, and for the active transistors alone, in the adiabatic CMOS. The improvement factors are reported in Fig. 11.

The passive transistors limit the energy benefit over the standard CMOS to about 10X, while the active transistors would enable up to 20,000X improvement. The conclusion is that low leakage transistors are necessary for adiabatic CMOS, but we *can* expect exponential improvements of the switching energy vs. the ramp rate. Based on the full cycle active energy of the 32 nm bulk MOSFET model, a reasonable prediction for the approximate adiabatic improvement factor is:

$$X_{\text{estimate}} = 0.81 \times t_{\text{ramp}}^{0.72}. \tag{13}$$

This prediction is pessimistic regarding the active transistor, since reducing the leakage in the passive transistor reduces also the current in the active transistor, which was not taken into account here. Interpreted as total energy, this estimate is optimistic regarding the OFF device, which in practice would always have some leakage.

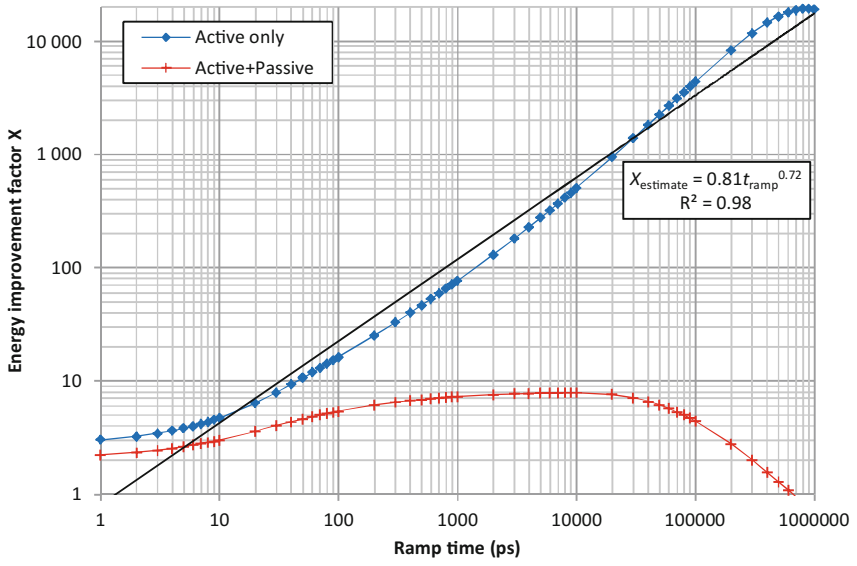

Fig. 11. Switching energy improvement over the standard CMOS, during the full cycle with transistor sizing $k = 10$. The passive transistors limit the benefit to at most 10X, while the active transistors alone would enable improvements up to 20000X over the standard CMOS.

6 First Steps for Automated Design Flow

The split-rail Bennett circuits can be synthesized using a standard CMOS tool flow with only minor modifications. With gate and logic level compatibility, the approach brings energy benefits over standard circuits even in the near-term future. Figure 12 illustrates the design process, which has been preliminarily tested in a standard CMOS design environment from the Cadence Design Systems.

Design Entry and Logic Synthesis. The design entry proceeds using for example structural or behavioral specification in a hardware description language (HDL). A standard logic synthesis tool produces a structural gate-level netlist, based on the gate library characteristics. Since the transistors in our approach are sized and the logic gates constructed exactly as in the standard CMOS, all automatic sizing optimizations and balancing of the delays in the combinatorial networks are directly valid.

Bennett Placement Constraints. The Bennett placement constraints ensure that the standard gates will be placed to the physical part of the floorplan where they can be efficiently wired to the correct power-clocks. For each instantiated gate, this is determined by the relative logic level. The floorplanning could be very flexible, but in our alpha-level process, we chose to implement the most straight-forward approach to place each specific level in a logic block to one physical row, as illustrated in Fig. 12. Inside one Bennett block, each pair of power-clocks is driving only one row, which simplifies the wiring complexity to approximately the same level as in the V_{DD} and ground rails of the standard CMOS. Each power-clock can drive also several separate Bennett blocks, depending on the circuit architecture choices.

The logic gate dependency information exists inside the standard synthesis tool, but this information is not generally accessible from the outside, due to commercial reasons. Therefore, we decided to implement our own software, which reads in the structural netlist produced by the logic synthesis and constructs a graph representation of the dependencies between the gates of a design. The tool basically tags each gate with a placement constraint defining which logic row the gate is to be placed in. This step should involve the balancing of the number of gates in each row to obtain a block with as fully utilized rows as possible, but our alpha-level tool does not yet implement this. Another important feature to be implemented is the placing of several logic blocks together, which would enhance the physical row utilization significantly.

Place-and-Route. The structural netlist containing the Bennett placement constraints can be fed into a standard place-and-route tool, which constructs the physical layout of the logic part of the design and connects the standard cells with wires. The standard optimizations for combinatorial logic are valid for the adiabatic circuits. The wires for the power-clocks can then be added to drive each appropriate row of logic, for example by using the automatic functions for the clock tree synthesis. However, the power-clock routing complexity is significantly smaller than the complexity of a standard clock in a block of random logic without the clock-per-row placement constraint.

Interfacing Sequential Logic. The combinatorial synthesis and placement are relatively straight-forward, but accommodating the sequential elements like flip-flops and latches in the synthesized standard netlist requires considerations of the circuit timing

and architecture. Basically, the standard CMOS flip-flops and latches all are compatible with the proposed approach, but their timing has to be controlled synchronously with the power-clocks. However, the location of the sequential elements in the output netlist of the standard logic synthesis has not been optimized for the retractile cascade circuits, and the best performance can be obtained only by giving additional constraints for the standard synthesis. We have not implemented automation for this yet.

Clock Tree and Physical Synthesis. In standard CMOS, the clock trees are usually balanced by an H-shaped branching structure, which can be used also for the distribution of a specific Bennett clock with uniform delay. However, the Bennett circuits typically utilize a significantly larger number of the clock phases, which the standard automatic tools are not currently able to synthesize without designer help. Our preliminary study of larger designs (a complete microprocessor) using the split-rail Bennett approach indicates that it is practical to arrange the standard cells in such a way that only one specific Bennett clock drives each logic row. This brings considerable uniformity to the standard cell design to help the physical synthesis: to summarize, the cells have the CMOS standard voltage and ground rails for the well-taps, while parallel to them, the positive and negative power-clock rails run through to drive the logic.

Circuit Architecture. Generally, the proposed augmented design flow has relatively small overhead vs. the standard CMOS flow, but to obtain the best performance and energy offered by the Bennett clocked circuits requires some additional considerations. One of the tradeoffs between the computing performance and the energy is related to the size of each block and the number of power-clocks: the larger the block, the more energy recovered, but the smaller the number of complete computations results per clock cycle.

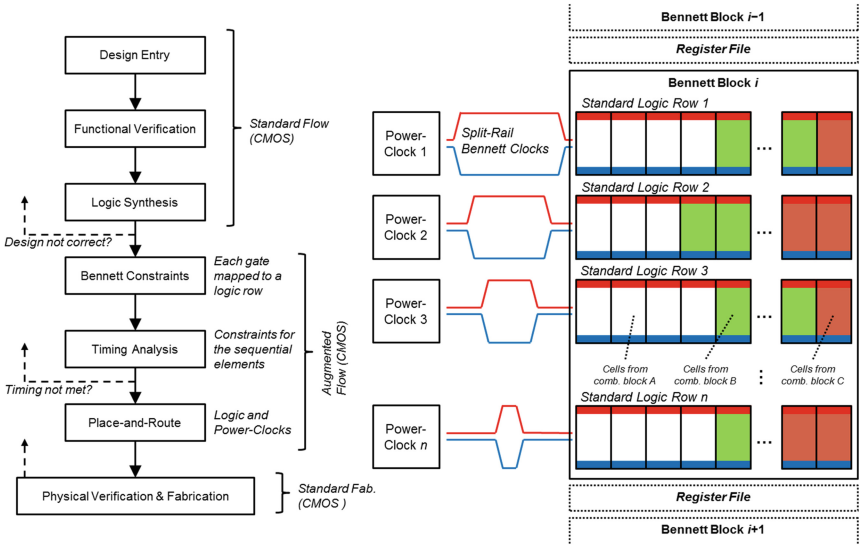

Fig. 12. Proposed design flow for the split-rail Bennett clocked adiabatic CMOS, based on the standard flow, and the physical floorplan. A Bennett block can contain standard cells from several logic blocks, while the power-clocks can be shared between several Bennett blocks.

This tradeoff is not simple because of the heat limit: a large block size might enable higher computation performance per wall-clock time unit, since a small block would not be able to run as fast as theoretically possible.

7 Conclusion

The static complementary transistor circuit structure is extremely reliable and scalable, as the standard CMOS technology has proven. However, the operating principle of throwing away *all* signal energy in a logic gate during *every* switching cycle is not a sustainable solution. We believe that already in the short-term future, energy-recovery becomes one of the most important methods for increasing the computing performance of the integrated circuits. As the main contribution, this paper has demonstrated how to extend the standard CMOS circuits to incorporate energy recycling and how to adjust existing design tools to attain compatible design automation.

The adiabatic circuits analyzed here utilize the standard CMOS logic gates, but the static supply rails have been replaced by the split-rail Bennett power-clocks, enabling logical reversibility in the circuit. Implemented in a 32 nm transistor technology which is *not* well-suited for the adiabatic circuits due to leakage, up to 10X switching energy improvements can be achieved compared to the standard CMOS. Unfortunately, the necessary overheads related to the generation of the complex power-clocks can waste even this amount of saved energy, reducing the system improvement significantly.

Low-leakage transistors are necessary to achieve the full potential offered by our approach. The transistors which should be OFF consume most of the energy and this passive power limits the analyzed technology. However, the active switching energy scales extremely well in the elementary energy-speed tradeoff, and the ON transistor efficiency would enable up to 20,000X improvements over the standard CMOS.

Since the transistor leakage current in the OFF state sets a lower bound on the energy efficiency of the adiabatic circuit, it is important to consider devices with low leakage. An area of intense research currently is that of tunnel FETs (TFETs), which use energy filtering to produce a subthreshold swing that is lower than the conventional 60 mV/dec [6, 7]. Our ongoing work characterizes the system level power and performance of the adiabatic complementary transistor circuits using these "steep" devices. Preliminary results indicate that the low-leakage devices could achieve the two-to-four orders of magnitude switching energy improvement over the standard CMOS, which would easily justify the overheads related to our adiabatic circuit style.

Acknowledgments. This work was supported in part by the National Science Foundation, grant number CHE-1124762.

References

1. ITRS, International Technology Roadmap for Semiconductors, ITRS report (2012). http://www.itrs.net/Links/2012ITRS/Home2012.html
2. Starosel'skii, V.I.: Adiabatic logic circuits: A review. Russ. Microelectron. **31**(1), 37–58 (2001) http://dx.doi.org/10.1023/A:1013857006906

3. Valiev, K.A.: Starosel'skii, V.I.: A model and properties of a thermodynamically reversible logic gate. Russ. Microelectron. **29**(2), 83–98 (2000)
4. Younis, S.G.: Asymptotically Zero Energy Computing Using Split-Level Charge Recovery Logic. Ph.D. Thesis. http://dspace.mit.edu/handle/1721.1/7058
5. Lent, C.S., Liu, M., Lu, Y.: Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling. Nanotechnology **17**(16), 4240–4251 (2006)
6. Seabaugh, A.C., Zhang, Q.: Low-voltage tunnel transistors for beyond CMOS logic. Proc. IEEE **98**(12), 2095–2110 (2010)
7. Ionescu, A.M., Riel, H.: Tunnel field-effect transistors as energy-efficient electronic switches. Nature **479**(7373), 329–337 (2011)
8. Patterson, D.: The trouble with multicore. IEEE Spectr. **47**, 28–32 (2010)
9. Esmaeilzadeh, H., Blem, E., Amant, R., Sankaralingam, K., Burger, D.: Dark silicon and the end of multicore scaling. In: 38th Annual International Symposium on Computer Architecture. pp. 365–376. ACM, San Jose, CA (2011)
10. Landauer, R.: Irreversibility and heat generation in the computing process. IBM J. Res. Dev. **5**, 183–191 (1961)
11. Costello, D.J., Forney, G.D.: Channel coding: The road to channel capacity. Proc. IEEE **95**, 1150–1177 (2007)
12. Sathe, S.A.V., Ouyang, C., Papaefthymiou, M., Ishii, A., Naffziger, S.: Resonant clock design for a power-efficient high-volume x86-64 microprocessor. In: 2012 IEEE International Solid-State Circuits Conference (ISSCC). p. 68–70. IEEE, San Francisco, CA (2012)
13. Athas, W.C., Svensson, L.J., Koller, J.G., Tzartzanis, N., Chou, E.Y.-C.: Low-power digital systems based on adiabatic-switching principles. IEEE Trans. VLSI Syst. **2**(4), 398–407 (1994)
14. Ferrary, A.: Adiabatic Switching, Adiabatic Logic, 20 March 1966
15. Younis, S.G., Knight, T.F.: Asymptotically zero energy split-level charge recovery logic. In: Proceedings of 1994 International Workshop on Low Power Design, pp. 177–182 (1994)
16. Bennett, C.: Logical reversibility of computation. IBM J. Res. Dev. **17**, 525–532 (1973)
17. Predictive Technology Model (PTM) library of the Arizona State University, Nano-CMOS. http://ptm.asu.edu/
18. Linear Technology LTspice IV, version 4.20i. http://www.linear.com/designtools/software/