Practical Considerations of Clock-Powered Logic

William Athas

House Ear Institute

2100 West Third Street

Los Angeles, California

athas@hei.org

ABSTRACT

Recovering and reusing circuit energies that would otherwise be dissipated as heat can reduce the power dissipated by a VLSI chip. To accomplish this requires a power source that can efficiently inject and extract energy, and an efficient power delivery system to connect the power source to the circuit nodes. The additional circuitry and timing required to support this process can readily exceed the power-savings benefit. Clock-powered logic is a circuit-level, energy-recovery approach that has been implemented in two generations of small-scale microprocessor experiments. The results have shown that it is possible and practical to extract useful amounts of power savings by leveraging the additional circuitry for other compatible purposes. The capabilities and limitations of clock-powered logic as a competitive low-power approach are presented and discussed in this paper.

Keywords

Energy-recovery CMOS, clock-powered logic, adiabatic charging, microprocessors, ER-CMOS, supply-voltage scaling.

1. INTRODUCTION

The energy dissipated to transport energy from one place to another is proportional to the speed of energy transport [1]. This observation led to research that started circa 1990 into the low-power implications for CMOS VLSI. Since then a number of approaches have been proposed and carried forward to different stages of developments. Many names have been coined to describe the various approaches, but they all have shared the common principle that circuit energies inside a CMOS chip, which would otherwise be dissipated as heat, could instead be recovered and reused. In this paper I will generally refer to all of the approaches that use this idea in some form or another as energy-recovery CMOS (ER-CMOS).

Few working prototypes have been demonstrated that operationally recover and reuse energy to reduce power. Fewer

still have shown indications of competitiveness with other low-power approaches when faced with solving the same design problem. In this paper I will argue that the practical benefit to ER-CMOS is as a complementary approach for managing signal energies inside a CMOS VLSI chip through the judicious use of special driver circuits for the large on-chip capacitive loads. Specifically, ER-CMOS is advantageous for signals that require high fan-out. The most compelling examples are driving the enable lines and the bit lines for a write operation in a high-density memory array.

The organization of this paper is to first concisely define the ER-CMOS advantage and then review the practical barriers that must be overcome to make an ER-CMOS scheme a competitive low-power approach. This discussion will point out the need for efficient and minimally intrusive ways to introduce ER-CMOS into an otherwise conventional design. The paper concludes with a constructive plan for introducing clock-powered logic into a conventional design, first, as a means for reducing the power needed to drive the usual clock loads, and secondly as a unique way to reduce power dissipation for high-capacitance circuit nodes inside the chip.

2. REVIEW

In the simplest terms possible, the real-world benefit to ER-CMOS is that it is possible to reduce the power dissipated to drive a signal without having to reduce the signal voltage swing. Since signal energy in CMOS varies as the square of the voltage amplitude, this perspective on the benefit supports the claim that ER-CMOS is an approach for managing signal energies. Applications which have the most to gain are those where the internal signal energies span a wide range.

The question is then when is it necessary or advantageous to *not* reduce the voltage? There are some niche applications, such as VLSI pin drivers for legacy (3.3V and 5V) systems [2], and row and column drivers for LCD panels [3]. In these kinds of applications, the voltage levels are, by definition or by physics, "non-negotiable." ER-CMOS has proven to be an effective approach for reducing power dissipation in these applications. An important observation though is that in the implementations for these applications only one stage of ER-CMOS circuitry, i.e., a signal driver, was used while the logic circuitry was implemented in conventional static CMOS.

There have been a large number of logic styles proposed that have used some form of energy recovery to implement low-power digital logic functions. From a practical standpoint, it is mandatory to compare their speed and power performance to

those of the other more conventional means for reducing power, most notably reducing the supply voltage of fully-restored (full-swing) static CMOS logic. This standard logic style is moderately high speed, has good noise immunity, and has good energy-versus-delay scalability across a wide and useful operating frequency range. To compare supply-voltage-scaled CMOS (SVS-CMOS) to ER-CMOS there are three key aspects to consider:

1. Energy-versus-delay performance
2. Power delivery
3. Power generation

At the lowest level of design in an ER-CMOS system, energy dissipation is proportional to $(R_{ON}C/T)$ where R_{ON} is the path on-resistance starting from the power generator and through the power deliver system to the circuit node, C is the capacitive load of the circuit node, and T is the transport or transition time. The charging or discharging of a circuit node in this manner is called *adiabatic charging*. In bulk CMOS technology the small-signal source-to-drain on-resistance scales as

$$\frac{L}{V_{GS} - V_{TH}},\qquad \text{(Eq. 1)}$$

where L is the channel length, V_{GS} is the gate-to-source voltage, and V_{TH} is the threshold voltage. From this equation and for constant transport time T, the benefit of ER-CMOS improves with increasing voltage swing (V_{GS}) and decreasing feature size. Given fixed requirements for voltage swing, channel width, and transport time, the potential benefit increases with each channel-length reduction in the CMOS fabrication technology. This relationship between fabrication technology and energy performance is the reason why ER-CMOS has become an effective solution for the niche applications of VLSI pin drivers and row and column drivers for LCD panels.

For the general case of digital logic, signal energies are a function of the voltage swing, which for fully-restored logic, equals the supply voltage. Reducing the supply voltage is a direct and simple method to decrease power with no added circuitry required of the logic. Furthermore, the energy-versus-delay tradeoffs are more attractive than ER-CMOS when the supply voltage and logic swing are well above V_{TH}. In this regime, delay varies approximately linearly with voltage while signal energy, as always, varies as the square of the voltage. This is in contrast to the linear-versus-linear relationship in ER-CMOS. Given the choice of turning down the voltage or increasing the transition time, the better choice for this regime is to reduce the signal energies through supply voltage reduction.

3. BEYOND VOLTAGE SCALING

Threshold voltage plays a decisive role since as the gate-to-source voltage of the transistor approaches V_{TH}, charge-carrier diffusion rather than drift dominates the charge (and energy) transport through the transistor channel. The practical implication is that delay then varies exponentially with voltage. The onset of this exponential delay can be staved off at the CMOS fabrication technology level by modifying the implant step that controls the threshold voltage. The cost of the threshold reduction is an exponential increase in sub-threshold conduction current. For supply-voltage-scaled CMOS (SVS-CMOS) it is possible to minimize power dissipation though by optimizing the choice of V_{TH}. The optimum has been shown to be a function of the anticipated circuit activity factor and the number of cascaded logic stages (logic depth) [4].

The exponential sensitivity to supply and threshold voltage in this regime suggests that the linear sensitivity of ER-CMOS could offer better energy-versus-delay performance. It is important to note though that on an absolute time scale, other sources of power loss, such as that due to leakage currents through reversed-biased p-n junctions may dominate the total dissipation at these extremely low-power and low-frequency operating levels. A careful device-physics-level analysis is warranted for this region of operation.

Aside from the power issues due to leakage, there is the theoretical and practical limitation that the circuitry must be operated in a purely reversible fashion to sustain the linear energy-versus-delay tradeoff in this operating regime. Pure reversibility can be done simply with retractile cascades [5] or with sophisticated reversible pipelines [6,7]. The limitations imposed by retractile cascades are that each logic stage requires its own power rail and nested power pulse and that the throughput is inversely proportional to the number of logic stages.

Reversible pipelines on the other hand require only a fixed number of power rails and their timing is also fixed. However, typically more than ten power rails are required and the timing relationships between the power signals are complex. There is also the overhead that each logic function and its inverse must be implemented.

The overhead for pure reversibility is high. There are two general strategies for reducing the overhead by eliminating the condition that all circuit nodes must be adiabatically charged. The first strategy is to impose a "one-way-ness" into the current flow through the use of diodes, "diode-connected FETs," or circuit structures that function as diode-connected FETs. In each case the net effect is to only partially recover charge from some nodes. The power dissipated in charging or discharging through a diode-type structure is $I \cdot V_D$, where V_D is the voltage drop of the diode. The magnitude of the unrecoverable, "marooned," signal energy is $(1/2)CV_D^2$, which limits the energy-versus-delay scalability.

The other general approach is to mix conventionally powered circuit nodes with adiabatically charged ones. The principle here is that "all circuit nodes are not created equal." Large capacitance nodes are adiabatically charged while small capacitance nodes are conventionally powered from dc sources. Asymptotically, the energy dissipation of the conventionally powered nodes limits the degree to which energy dissipation can be reduced through adiabatic charging.

There are also the efficiency issues of the power generation to be considered. The power generator must first deliver the energy efficiently and then efficiently recover the energy. During each charge-discharge cycle the power generator must have a means to temporarily store the energy. Two approaches have been successfully reduced to practice: resonant charging and stepwise charging [8]. Resonant charging uses inductors or transmission lines to temporarily store the energy and stepwise charging uses banks of switched capacitors. In both approaches power FETs control the energy delivery and recovery process. The efficiency of the overall system depends on minimizing the losses through these components. In the case of resonant charging, the losses are due to the I^2R_{ON} dissipation where R_{ON} is the on-resistance of the

power FETs that steer the current between the inductors and the capacitive loads. To a first approximation, the on-resistance is inversely proportional to the gate capacitance, C_G. If the gates of the FETs are conventionally powered, then there is a tradeoff between the I^2R_{ON} losses and the C_GV^2 gate losses. When the two source of dissipation are optimally balanced, the linear energy-versus-delay scaling relationship for ER-CMOS becomes a square-root tradeoff, which further diminishes the practical benefit [9].

There are exceptions. First, it is possible to design power generators that can produce arbitrarily complex waveforms entirely from passive components, i.e., inductors, capacitors, and transmission-line components [10, 11]. Second, it is possible to use a resonant gate drive. This has been done in the case of the *blip circuit*, which generates two-almost non-overlapping resonant clock pulses [12]. The output of one clock-phase driver directly controls the pulldown power FET of the other clock-phase driver in a complementary fashion. Although there are some small losses, which are tediously difficult to account for, empirical laboratory measurements of operating blip circuits for different loads, voltages, and frequencies have shown that the energy-versus-frequency scaling is close to linear.

The power generator connects to the circuit loads through a power delivery system. The same is true for SVS-CMOS, but in the SVS-CMOS case the supply and ground rails remain at constant, dc potentials. Parasitic capacitances associated with the supply rails are beneficial because these capacitances reduce the impedance of the power supply. In contrast, the power delivery system of an ER-CMOS chip uses one more supply rails which repeatedly injects and extracts energy. Parasitic capacitances on these rails contribute to the total power dissipation since they are synchronously cycled with the circuit nodes.

More importantly, in SVS-CMOS the energy on the supply rails is always present and there are no constraints on when the circuits operate. An ER-CMOS scheme places tight constraints on the timing since FETs may only be turned on or off when the potentials across the channels are nearly zero. There must be careful synchronization between the timing of the logic and the timing of the ac power.

These problems are mitigated with stepwise charging, which is an alternative to resonant charging. Stepwise charging uses multiple dc power rails that are equally spaced between ground and the maximum voltage swing. Through charge sharing between the loads and the large tank capacitors attached to the dc rails, charge and energy are injected and extracted in a stepwise fashion. The energy-versus-delay scaling of an optimized stepwise charger varies as one over the cube-root of the transport time, which is much worse than resonant charging. However, stepwise charging may be used in a modular, asynchronous fashion, and most importantly, it does not require inductors. For these reason stepwise chargers have been used in the applications of VLSI pin drivers and LCD panel drivers despite the theoretically inferior energy-versus-delay scalability.

4. CLOCK-POWERED LOGIC (CPL)

Energy-versus-delay scalability in SVS-CMOS is practically limited by V_{TH}. However, except for the case of purely reversible logic, ER-CMOS approaches are also constrained by the same V_{TH} limit with higher overhead factors for other parts of the
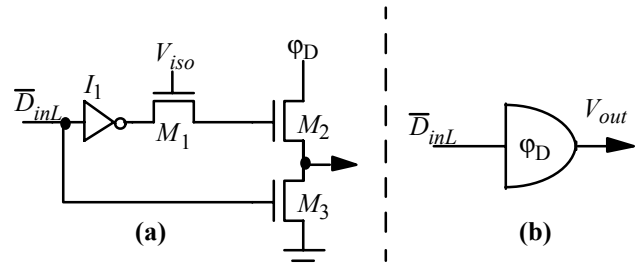
Figure 1. The clocked buffer as it is used for power-efficient energy injection and extraction (a) and its logic symbol (b). M_1 is the isolation transistor, M_2 is the bootstrapped clock-pass transistor, and M_3 is the clamp device.

system. Unless and until the problems of efficiently implementing reversible logic in CMOS are solved, ER-CMOS approaches are subject to the same limitation.

For a practical ER-CMOS solution to compete with SVS-CMOS, it must offer a sufficiently high power-savings benefit to offset the unavoidable overhead which is required to support the energy delivery and recovery process between the power generator and circuit nodes. Resonant charging is attractive in that it offers the highest efficiency, but it requires a system-level approach for integration into an otherwise conventional system. Stepwise charging can be introduced in a modular fashion, but the additional area and power required to operate the control circuitry precludes from using it at a level of fine-granularity inside a VLSI chip.

The instances for which ER-CMOS has been successfully deployed on a significant scale inside a chip have been in the case of signal drivers.

The basis for clock-powered logic is the combination of the highly-efficient blip circuit for resonant power generation, a two-phase clock grid for power delivery, and the clocked buffer circuit [13] of Figure 1. The energy-recovery aspects to the operation of this circuit have been described in many places elsewhere [14]. The important point for ER-CMOS feasibility is that the bootstrapped transistor M_2 provides the lowest on-resistance per unit gate area for the path from the power-delivery system to the output node. The tradeoff between the gate capacitance and the on-resistance in this situation is the same as that of the power generator, i.e., energy-versus-delay performance will vary as the square root of $1/T$ when the power dissipation is minimized by optimizing the size of transistor M_2.

The combination of the clocked buffer and the resonant two-phase clocking approach leads naturally to a style of ER-CMOS in which the clock signals serve as a source of ac power for the on-chip circuit nodes. Since the clock powers these nodes, they are called *clock-powered* nodes. Note that the overhead of the clock-pass transistor (M_2), clamp device (M_3), and inverter (I_1) limits the granularity at which this circuit can be reasonably applied. Note also that the receiving logic must interface to a pulsed signal rather than a dc level.

CPL has been implemented in two generations of small-scale prototype microprocessors. In the first generation, the resonant clocks ran "hot" in that the voltage swing for the clock rails was greater than that of the supply voltage [15]. The design rationale for this decision was to drive as many of the on-chip nodes as possible directly with the clock signals since the clocks were
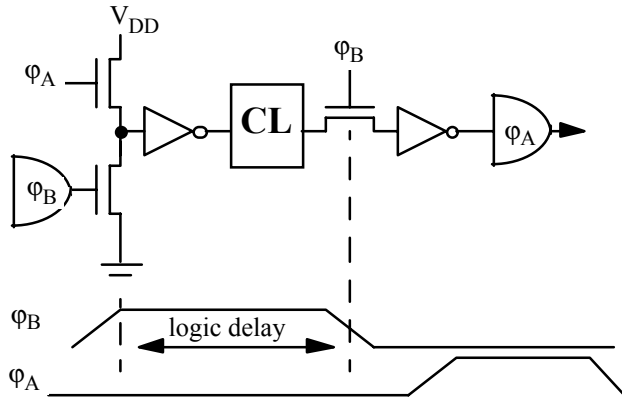
intended to be a source of "cheap" energy. The logic was dynamic with nFETs used as precharge pull-up devices and also for pass gates. The amplitude of the clock signals was the supply voltage plus one threshold voltage (plus body effect) to fully restore the outputs when nFETs were used to drive signals high. The direct generation of the negative versions of the clock signals was not possible with the blip-circuit power generator. This was an important motivator for using nFETs for precharge transistors and pass gates.

To put the overall results into context, assume a 90% efficiency in the power generator. This is equivalent to a ten-fold power in the clock circuit. However, realistically there is a 50% overhead in the added capacitance of the clock grid, and a 50% overhead in the clock swing. The latter may seem excessive, but is reasonable for a supply voltage of $4V_{TH}$ and after the body effect has been taken into consideration. The power savings are then only a factor of three from the original factor of ten.

An additional benefit of the hot clocks was that the circuits were smaller since less pFETs were needed.

Figure 2 depicts the logic arrangement and timing. The major drawback to this style is that there is less than a single clock phase for the propagation of signals through the logic block **CL**. The input pulldown network does not start to switch until the voltage swing of the clock-powered pulse rises above V_{TH}. The pass-gate

latch on the output stops loading when the clock pulse ramps down to V_{TH}. For the pass gate there is also the body effect to be considered which further shortens the usable logic propagation delay time per clock phase. To compensate, the supply voltage needs to be increased to make the logic run faster, which in turn requires the clock voltage to be increased. A further consideration is that the hot clocks can potentially cause long-term damage to the chip by exceeding the CMOS process specification for maximum allowable voltages across different junctions and oxides.

Second-generation CPL [16] was designed to address these problems. It offers more compute time per clock cycle for the logic and also has provisions for decoupling the clock swing from the logic swing. Figure 3 depicts a representative circuit that highlights the improvements. There were no changes made to the clocked buffers, but the precharged receiver circuit of Figure 2 was replaced with a static pulse-to-level converter circuit (called a *jam latch*). The dynamic latch of a pass-gate and inverter was replaced with an n-latch (also shown in Figure 4). The drawback is that the number of transistors required to convert and to latch a clock-powered signal increased from a minimum of seven to eighteen. The benefits are that the amount of signal-propagation time per clock cycle increased from less than one clock phase to nearly the entire clock period and the clock-signal voltage swings were completely decoupled from the supply-voltage amplitude. Consequently they could be optimized independently.

Extensive H-SPICE simulations of the driver, input, and latch circuitry of Figure 3 indicated better scalability than a simple SVS-CMOS inverter for equal propagation delay [14]. Figure 4 summarizes the result of the simulation experiments between SVS-CMOS and ER-CMOS driver/receiver pairs. In this simulation experiment, the supply voltage for the CMOS inverter was first reduced and the end-to-end delay was simulated. Since the input to the inverter had a 3.3V swing, this handicapped the pFET as the supply voltage was reduced. To compensate, an nFET was placed in parallel with the pFET and driven by the complement of the input. The simulation results for this modification are shown as the "Mod. Invertor" line and data points in the graph of Figure 4.
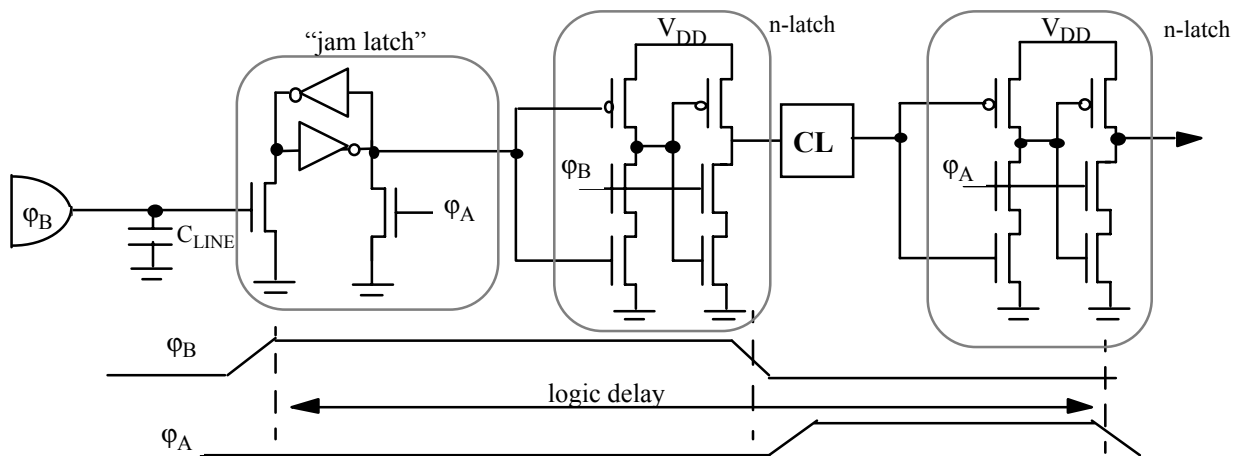
Figure 3. Second generation clock-powered logic. "Jam" latches convert clock-powered pulses to levels and also perform voltage level conversion. n-latches allow the entire system to be clocked with low-swing clock signals.
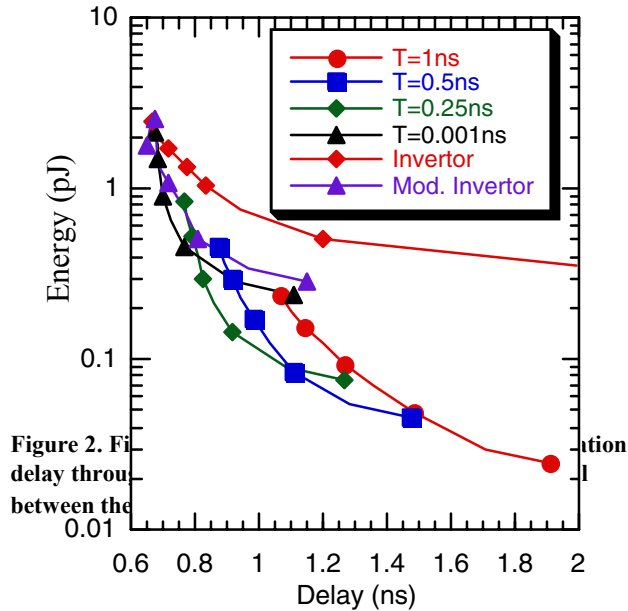
Figure 4. H-SPICE simulation results for end-to-end delay comparison between CPL driver/receiver and SVS-CMOS inverter.

Voltage scaling of the clock amplitude was also applied to the CPL driver/receiver pair of Figure 3, however, when energy-versus-delay became sub-linear, the voltage swing was increased and the transport time was also increased. By applying both techniques to this "driver problem," it was possible to extend the energy-versus-delay scalability to two orders of magnitude. One order of magnitude was due to the voltage scaling and the second was due to the energy recovery (time scaling).

5. ANALYSIS

The second-generation CPL style was implemented in a commercial application in which the threshold voltages for the CMOS fabrication technology were relatively high (800mV and -900mV) compared to the supply voltage requirement of 1.5V. Figure 5 shows a block diagram of the microprocessor and where clock-powered buses were used. Unlike first-generation CPL, we were not able to extensively use clocked buffers for recovering energy from the long lines driven by the controller because of the high circuitry overhead of the pulse-to-level converters. Many of those control lines were driven directly by conventional buffers.

The main purpose of the circuitry shown in Figure 3 is to recover the energy associated with the line capacitance (C_{LINE}). An alternative method that would have been appropriate for this one-to-one type of connection would have been to use a small-swing differential driver and receiver pair. Furthermore, a CMOS VLSI process that supported transistors with high thresholds (low leakage) and low threshold (high speed) would have further improved the scalability for the SVS-CMOS inverter.

These alternative techniques would have offered similar power savings for the operand busses between the Register File and the Function Units (see Figure 5), and between the Function Units and the Memory Access Unit. These techniques would not have required the overhead of the ac power delivery system or of the resonant clock generator. The timing constraints would have been

similar, e.g., clocking the voltage-level converters on the receiver side to reduce short-circuit power.

There was one important contributor to the power dissipation, however, for which the alternative techniques would not have been directly applicable which was in the write back into the Register File from the Memory Access Unit. This case is an example of a one-to-many (high fan-out) condition in which the receiver or input circuitry is to be minimized while the overhead of the driver becomes increasingly less important as the fan-out increases. The practical alternative would have been to write the register cells by using pulldown transistors as is done in the jam latch. This would have increased the area, delay and power of the register file for decoding and reading.

For completeness there is also the third case of many-to-one (high fan-in) configurations. ER-CMOS is not a good choice since these situations require many drivers and a single receiver. The driver overhead in ER-CMOS is high because it must include circuitry for the energy-recover process, as well as connections to the power delivery system. Here it is better to use a simple driver, such a pulldown transistor on a precharged line, and then limit the swing on the line. A sense-amp can then be used to convert the signal from small swing to normal swing.

6. CONCLUSIONS

High density in the register file is desirable but typically not a critical requirement in a microprocessor implementation. However, density is a first-order concern for memory arrays. Based on the results of the CPL microprocessor chips, low-power synchronous memory applications that have a high activity factor would be good candidates. It is highly desirable for density and speed performance reasons to make the memory cell as simple and small as possible. The major sources of dissipation at the array level are in driving the enable lines and the bit lines.

As was the case with the microprocessor register files, clock-powered signals can be used effectively to drive the enable lines for both read and writes operations because the enable lines must be driven full swing. Clock-powered signals can also be used to drive the bit lines for write operations.

There is a technique to provide for a low-voltage swing during the write operation [17], but this technique will only work for memory cells which can fully restore the signal levels inside the memory cell. The technique will not work for dynamic memory cells.
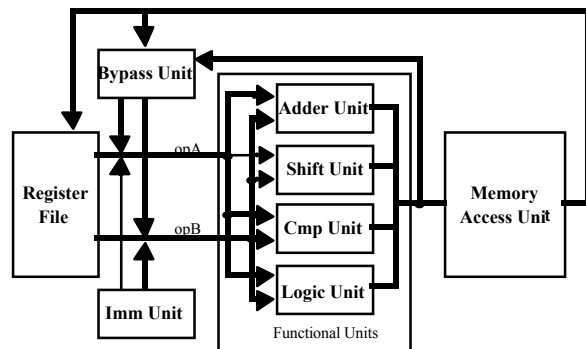

Figure 5. Block diagram of the datapath for the 2nd-generation microprocessor showing the clock-powered data busses.

In conclusion, the best application for clock-powered logic now appears to be memory arrays for embedded or portable applications that have continuous streaming throughput requirements at very low power operating levels. Ironically, one of the original motivations for this research direction was reversible computing and the avoidance of the erasure of information. In the final analysis, the unique low-power advantage to these types of circuits may be their superior efficiency for carrying out the process of erasing information.

8. REFERENCES

[1] Koller, J., and W. Athas, "Adiabatic Switching, Low Energy Computing, And The Physics of Storing and Erasing Information," *Proc. of the Workshop on Physics and Computation, PhysCmp 92*, IEEE Computer Society Press, Los Alamitos, CA (1993).

[2] Svensson, L., W. Athas, and R.S-C. Wen, "A Sub-CV2 Pad Driver With 10 ns Transition Time," *Proc. of the International Symposium on Low-Power Electronics and Design*, 105-108 (August 1996).

[3] Lal, R., W. Athas, and L. Svensson*, A Low-Power Adiabatic Driver System for AMLCDs*, ACMOS Technical Report, Information Sciences Institute, University of Southern California, Marina del Rey, CA (February 2000).

[4] Burr, J., and A. Petersen, "Energy Considerations in Multichip-Module Based Multiprocessors," *IEEE International Conference on Computer Design*, 593-600 (1991).

[5] Hall, J., "An Electroid Switching Model for Reversible Computer Architectures*," Proc. ICCI'92, 4th International Conf. on Computing and Information*, (1992).

[6] Younis, S. and T. Knight, "Practical Implementation Of Charge Recovery Asymptotically Zero Power CMOS," *Proc. of the 1993 Symposium on Integrated Systems*. MIT Press, 234-250(1993).

[7] Koller, J., W. Athas, and L. Svensson, "Thermal logic," *Proceedings of the 1994 Workshop on Physics and Computation*, IEEE Computer Society Press, Los Alamitos, CA 119-127 (November, 1994).

[8] Svensson, L. and J. Koller, "Driving A Capacitive Load Without Dissipating fCV2," *Proc. of the IEEE 1994 Symposium on Low-Power Electronics*, (October, 1994).

[9] Athas, W., J. Koller, and L. Svensson, "An Energy-Efficient CMOS Line Driver Using Adiabatic Switching," *Proc. of the Fourth Great Lakes Symposium on VLSI Design*, IEEE Computer Society Press, Los Alamitos, CA 159-164, (March, 1994).

[10] Younis, S. and T. Knight, "Non-Dissipative Rail Drivers for Adiabatic Circuits," *Proc. of the 16th Conference on Advanced Research in VLSI*, Chapel Hill, NC, 404-414, (March, 1995).

[11] Glasoe, G., and J. Lebacqz, eds., *Pulse Generators*, McGraw-Hill, 175-224 (1948).

[12] Athas, W., L. Svensson, and N. Tzartzanis, "A Resonant Signal Driver For Two-Phase, Almost-Non-Overlapping Clocks," *Proc. of the 1996 International Symposium on Circuits and Systems*, Atlanta, GA, 129-132 (May, 1996).

[13] Glasser L., and D. Dobberpuhl*, The Design and Analysis of VLSI Circuits*, Addison-Wesley, Reading, MA, (1985).

[14] Tzartzanis, N., "Energy-Recovery Techniques for CMOS Microprocessor Design," Ph.D. Thesis, EE-Systems Dept., University of Southern California, Los Angeles, CA (August, 1998).

[15] Athas, W., N. Tzartzanis, L. Svensson, and L. Peterson, "A Low-Power Microprocessor Based on Resonant Energy," *IEEE Jnl. of Solid-State Circuits*, 1693-1701, (November, 1997).

[16] Athas, W., N. Tzartzanis, W. Mao, R. Lal, K. Chong, L. Peterson, and M. Bolotski. "Clock-Powered CMOS VLSI Graphics Processor for Embedded Display Controller Application," *International Solid-State Circuits Conference*, San Francisco, CA 296-297 (February, 2000).

[17] Alowersson, J., and P. Andersson, "SRAM Cells for Low-Power Write in Buffer Memories," Symposium on Low-Power Electronics and Design, (1996).