Low Power Design Issues and Solutions for Future VLSI Systems

Anantha Chandrakasan, James Goodman, James Kao, Amit Sinha, Paul Peter Sotiriadis
Department of EECS,
Massachusetts Institute of Technology, Cambridge

1. Introduction

Over the past decade, there has been significant advances in the design of low power circuits and systems [1]. Power supply voltages have scaled due to improvements in process technology (e.g., reduced and multiple threshold devices) and the use of increased parallelism. The use of low swing signaling has further reduced the energy dissipation in many applications. Several techniques have also been proposed to reduce the switched capacitance through the use of conditional clocks, balanced topologies, data coding, etc.

This paper highlights some of the emerging trends in low-power design. The notion of energy scalable computing is introduced where the system (both software and hardware) provides flexibility to trade-off energy dissipation and the quality of results. This enables the user to control the battery lifetime. This paper also presents techniques to deal with increased leakage in low voltage systems and the use of data coding techniques.

2. Energy Scalable Architectures and Circuits

In many contexts it is important to structure algorithms and systems to allow for the possibility of trading off between the accuracy or optimality of the result and the energy dissipation. This allows the user to *dynamically* control the battery lifetime of a portable system.

For such systems, scalable architectures are required to minimize energy consumption of processors when the quality requirements on computational results change. Energy scalable computing exploits variations in the computation (e.g., precision, throughput, data statistics) to minimize the energy dissipation per input sample. Minimizing the energy dissipation requires the development of architectures that can adapt to match the current requirements of the application by varying either switched capacitance per sample, or the operating voltage.

As an example, consider an encryption processor where the level of security (i.e., quality) and energy consumed to encrypt a bit can be traded-off dynamically. The energy scalable encryption processor is based on a variable-width quadratic residue generator (QRG) which is a cryptographically-secure pseudo-random bit generator that is based upon the work in [2]. The QRG operates by performing repeated modular squarings. The modular squaring is performed using an algorithm based on Takagi's iterated radix-4 algorithm [3] which requires $(\log_2 Q)/2$ iterations to compute the result $P = X \cdot Y \bmod Q$. The least significant $\log_2\log_2 Q$ bits of each result can be extracted and used as a strong reproducible pseudo-random source for applications such as a stream cipher.

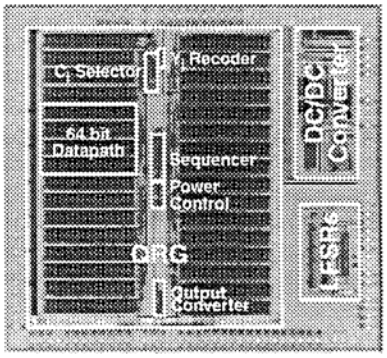

Figure 1. Layout of an energy scalable QRG.

For the QRG processor the quality scales sub-exponentially with the modulus length, while the energy consumption scales polynomially. A fully scalable QRG architecture was developed where the width ($w = \log_2 Q$) can be reconfigured on the fly to range from 64 to 512 bits in 64 bit increments (Figure 1) [4]. The design makes extensive use of clock gating to disable unused portions of the QRG. Hence the switched capacitance of the QRG is minimized and energy scalability is achieved.

Further energy/security scalability is achieved through the use of an adaptive supply [5]. Rather than designing a system with a static supply to meet a specific timing constraint under worst case conditions (i.e., establishing the feedback around the power converter to fix the output voltage), it is more energy efficient to allow the voltage to vary such that the timing constraints are just met at any given temperature and operating conditions; this is accomplished by establishing the feedback around a fixed processing rate or delay. In the aforementioned example, when operating at a reduced width, the number of cycles required per multiplication is reduced and therefore the supply voltage can be reduced for a given throughput. The supply is varied using an embedded custom DC/DC converter. Figure 2 shows the energy /security scalability trade-off. This plot is obtained by varying both the bitwidth and supply voltage dynamically.

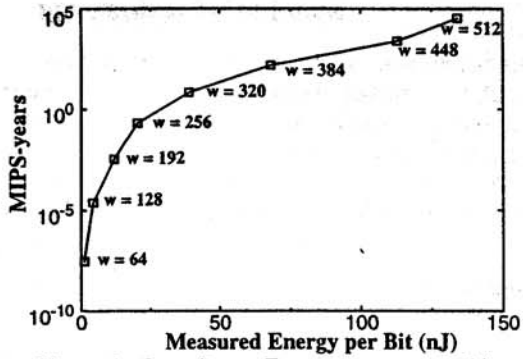

Figure 2. Security vs. Energy to encrypt a bit.

3. Energy Aware Software

While the design of application specific circuits clearly results in the minimum energy solution, the flexibility provided by software is often highly desirable. The notion of energy scalable computing using adaptive supply voltages can be exploited at the software level as well. The basic idea is to trade off the energy dissipation to run a program and the time required for completing the task and/or the accuracy required. Low power processors (e.g., StrongARM) allow the clock frequency to be set using software control. If a longer latency is acceptable, the energy dissipation can be lowered by reducing both the frequency and the voltage dynamically. It is important to note that decreasing frequency alone does not reduce the energy consumption since voltage is fixed. In fact, energy increases as frequency is reduced at a fixed supply since the leakage energy increases as the execution time increases.

The energy consumption for a 1024 point FFT as a function of the supply voltage and frequency is shown in Figure 3. When the operating frequency is fixed and the supply voltage is scaled, the energy scales almost quadratically. However, not all frequency, voltage combinations are possible. For example, the maximum frequency of the StrongARM is 206 MHz and it requires a minimum operating voltage of 1.4 V. The line across the surface plot demarcates the possible operating regions from the extrapolated ones (i.e. the minimum operating voltage for a given frequency). Future Operating Systems for portable systems must have the ability to schedule the supply voltage based on the current workload in the system and the performance demand / energy bound set by the user. If computation can be traded off efficiently with quality, energy savings can be obtained by executing the reduced computations at lower voltage set by the OS based on user input/battery state.

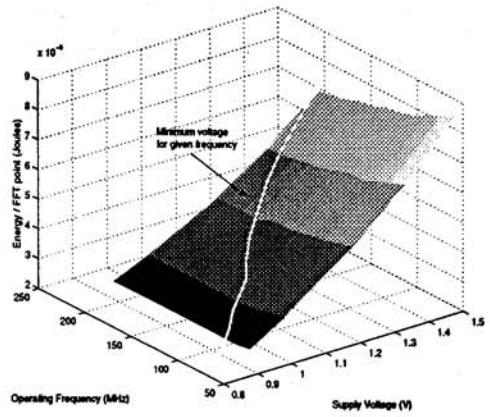

Figure 3. FFT energy consumption as a function of V_{DD} and frequency on the SA-1100.

As the supply voltages and thresholds are reduced, system designers have to pay increasing attention to leakage currents. For the StrongARM, at maximum duty cycle and minimum voltage (for a given frequency), the leakage energy is about 10%. However, the leakage energy rises exponentially with supply voltage and decreases linearly with frequency as shown in Figure 4. Therefore, operating at a voltage, above the minimum possible, for a given frequency, is not advisable. This might be an issue in fixed supply, variable frequency systems. For low duty-cycle systems, the overall energy consumption becomes increasingly dominated by leakage effects. The fixed task consumes a certain amount of switching energy per execution while the system leaks during the idle mode between tasks. Extensive clock gating techniques, such as those present in the StrongARM, reduce the unnecessary switching energy in the "idle" mode. The StrongARM does also have a "sleep" mode where the processor state is stored and the supply voltage is reduced to zero for most circuits. This significantly reduces the leakage problem. However, reverting to sleep mode between duty cycles may incur a lot of overhead (in terms of cycles and energy) or may not be supported by the target processor.

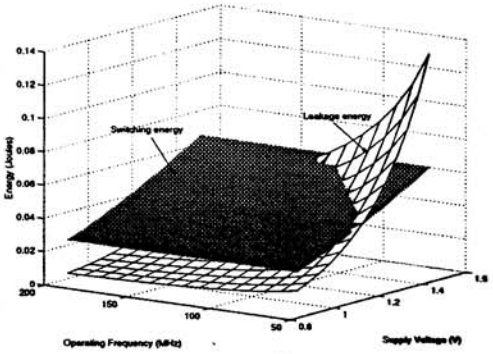

Figure 4. FFT energy components.

4. Leakage Control Techniques

Subthreshold leakage power will become an increasingly larger component of total power dissipation in future technologies. Large leakage currents are especially wasteful in burst mode type systems, where the majority of the time circuits are in a standby mode where no computation is taking place, yet power is continuously drained through subthreshold leakage currents.

One effective way to reduce subthreshold leakage currents in future technologies is to utilize multiple threshold devices. Low V_T devices can be used for high speed operation, while high V_T devices can be used to reduce leakage currents. One simple application of this technology is to partition a circuit into critical and non critical regions, and to only use low V_T devices where necessary [6]. Another approach at reducing leakage currents during standby mode is Multi-Threshold CMOS (MTCMOS), which involves using high V_T sleep transistors to gate the power supplies for a low V_T block [7]. Leakage currents can be reduced by several orders of magnitude during the standby mode by turning off the high V_T power switches. However, optimal sleep transistor sizing is difficult for larger circuits and sequential circuit must be modified to prevent data loss during standby modes [8].

4.1 Dual V_T Domino Logic

Dual V_T domino logic is yet another application of dual threshold voltages, which avoids the sizing problems associated with conventional MTCMOS [9]. Dual V_T domino requires no complex sizing requirements, incurs no performance penalty over an all low V_T design, and provides the standby leakage characteristic of a purely high V_T implementation. The conversion from an existing low V_T domino design to a dual V_T implementation is straightforward and simply requires a modification of some devices to be high V_T. No resizing is necessary, and timing relations remain unchanged. In this approach low V_T devices are used for all transistors that can switch during the evaluate mode and high V_T devices are used for all transistors that can switch during the precharge modes.

Figure 5 shows a typical gate in a clock delayed domino methodology where subsequent clock signals to downstream gates are delayed. By matching the clock delay with the data propagation to the next gate, NMOS footswitches can be eliminated for subsequent domino gates, thereby improving performance.

The critical path delay through the domino gate remains fast since all transitions occur through low V_T devices. However, the precharge times are slowed down, but this is acceptable since they are not in the critical path. This approach allows a trade-off between precharge time and standby leakage currents. By placing the gate in the evaluate mode, and forcing all gate inputs to a logic 1,

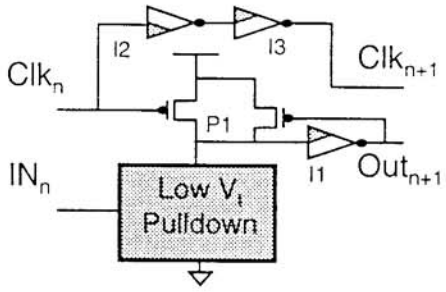

Figure 5. Dual V_T Domino Gate (LVT shaded).

then all high V_T devices in the gate will be strongly turned off and subthreshold leakage will be very small during the standby mode. Similarly for an entire pipeline block, simply forcing the first level inputs to the pipeline to be logic 1's will cause all subsequent gates to be evaluated, which will then cause all high V_T devices in the entire pipeline to be strongly turned off. As a result, dual V_T domino can provide the performance of an all low V_T design with the leakage characteristic of an all high V_T one.

4.2 Adaptive Body Biasing

By applying reverse bias to the body of a transistor, the threshold voltage of a device can be dynamically modified. Thus during the standby mode, reverse body bias can be applied to limit subthreshold leakage currents without the need for multiple threshold devices [10]. One attractive feature with this leakage suppression technique is that no extra devices are required, and that circuits during the standby mode remain on so that sequential circuits retain their data. However, body bias effectiveness is reduced as technology scales because the body factor drops with lower V_T and shrinking channel lengths.Because body biasing allows dynamic threshold control over a continuous range, it can also be used to adjust circuit performance during active mode. As technology scales, process variation will have a large impact on V_T, and overall circuit performance will vary greatly. A significant number of circuits will be faster than the target frequency, which can result in excess leakage currents.

Figure 6 shows an adaptive body biasing circuit block (similar to [11]) that can dynamically adjust threshold voltages to compensate for process variations and effectively tighten the distribution of delays in a collection of circuits. A delay feedback mechanism develops the appropriate reverse body bias such that the circuit critical path delay, represented by a matched delay line, meets the target operating speed. By applying reverse bias to fast circuits, excessive leakage currents can be reduced and the power dissipation of the active mode can be reduced as well.

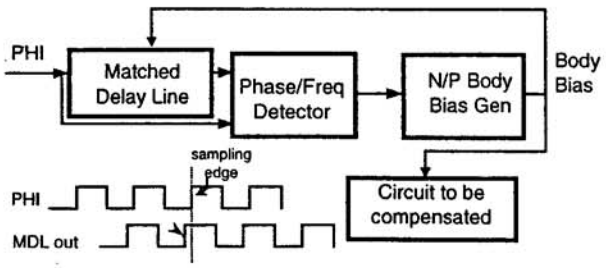

Figure 6. Adaptive Body Biasing Diagram.

5. Energy Reduction in Interconnect

An important emerging component of the power consumption in digital processors involves the transmission of data through high capacitance busses. Several techniques have been previously proposed for reducing interconnect power through the reduction of voltage swing and the reduction of transition activity via bus encoding (e.g., [12]). As the energy cost of performing logic reduces (e.g., with continued voltage scaling), more advanced schemes will be employed that perform more computation to reduce the energy associated with communication. Figure 7 shows an overall block diagram of the coding approach.

The original data bus of *m-lines* is extended by addition of *a-more* lines. Every initial *m-bit* word $D(k)$ corresponds now to a set of *r, (m+a)-bit* words where *r* is the number of *encoding functions*. The *control function* calculates the energy cost for transmitting each of these *r* codewords through the data bus and chooses the one with the lowest cost. The choice of the particular codeword $L^D(k)$ is encoded into the *a* additional lines, $L^C(k)$. This information enables the *receiver* part of the bus to recover the original data $D(k)$. Our initial analysis shows

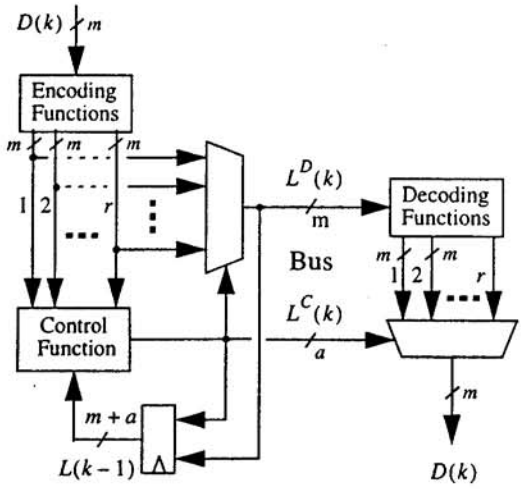

Figure 7. General Coding Scheme for Low Power Data Busses.

that we can reduce up to 40% the power dissipation using 4 bits for the control word.

6. Conclusion

Future systems will require the hardware and software to support energy scalable computing. The ability to trade-off the quality of results and the energy will enable the user to evaluate the computation demands and control battery lifetime. Future low-power systems must pay particular attention to the increasing contributions due to sub-threshold leakage and driving high capacitance interconnect. Feedback techniques to control leakage and control threshold process variations will become an integral part of future systems. Interconnects will require optimization between computation and communication. Low-power logic techniques will favor adding computation to reduce the communication/system power.

References

[1] A. Chandrakasan, R. Brodersen, Low Power CMOS Design, IEEE Press, 1998.

[2] L. Blum, M. Blum, and M. Shub, "A Simple Unpredictable Pseudo-Random Number Generator," SIAM Journal on Computing, v. 15, n. 2, pp. 364-383, 1986.

[3] N. Takagi, "A Radix-4 Modular Multiplication Hardware Algorithm for Modular Exponentiation", IEEE Transactions on Computers, pp. 949 - 956, August 1992.

[4] J. Goodman, A. Dancy, A. Chandrakasan, "An Energy/Security Scalable Encryption Processor Using an Embedded Variable Voltage DC/DC Converter," IEEE JSSC, pp. 1799-1809, November 1998.

[5] V. Gutnik, A.P. Chandrakasan, "Embedded Power Supply for Low-Power DSP," IEEE Transactions on VLSI Systems, pp. 425-435, December 1997.

[6] W. Lee, et al., "A 1V DSP for Wireless Communications," IEEE ISSCC, pp. 92-93, February 1997.

[7] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, J. Yamada, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," IEEE JSSC, vol. 30, no. 8, pp. 847-854, August 1995.

[8] J. Kao, A. Chandrakasan, D. Antoniadis, "Transistor Sizing Issues and Tool For Multi-Threshold CMOS Technology," ACM/IEEE DAC, pp. 409-414, June 1997.

[9] J. Kao, "Dual Threshold Voltage Domino Logic," IEEE European Solid State Circuits Conference, pp. 118-121, September 1999.

[10] T. Kuroda et al, "A 0.9V, 150MHz, 10mW, 4mm2, 2-DCT Core Processor with Variable VT Scheme," IEEE JSSC, vol. 31, no. 11, pp. 1770-1778, November 1996.

[11] M. Miyazaki, H. Mizuno, K. Ishibashi, "A Delay Distribution Squeezing Scheme with Speed-Adaptive Threshold-Voltage CMOS (SA-Vt CMOS) for Low Voltage LSIs," IEEE/ACM ISLPED, pp. 48-53, 1998.

[12] M. Stan, W. Burleson, "Bus-Invert Coding for Low Power I/O", IEEE Tran. on VLSI Systems, pp. 49-57, March 1995.