Robust and Efficient Dynamic Voltage Scaling Architecture

Mohamed Elgebaly, Amr Fahim*, and Inyup Kang* and Manoj Sachdev
Department of Electrical and Computer Engineering, University of Waterloo, Ontario, N2L 3G1 Canada
* Qualcomm Inc., San Diego, CA, 92121 U.S.A.

Abstract— **An efficient dynamic voltage scaling (DVS) architecture to track process and temperature based on a lookup table is presented. If environmental conditions change faster than what the DVS system can track, the supply voltage must be raised to satisfy the worst case process and temperature conditions. In the proposed system, an efficient method of identifying the process is used, independent of temperature conditions. Voltage can be adjusted based on split information extracted. The proposed system can achieve up to 29% energy savings compared to the conventional system.**

I. Introduction

Portable devices such as personal digital assistants (PDAs), cellular phones, and portable computers are becoming part of the daily life. The high demand from users for more applications and functions integrated into these portable devices has pushed the design trend towards more integration and building SoCs for such devices. Yet, the more sophisticated the portable device is, the more energy consumption and the less battery life. Long battery life is a very important design and marketing parameter. A great deal of design effort is devoted to extending battery life time while keeping the same level of performance.

Designing more versatile portable devices is becoming more feasible as the technology scales. With smaller feature size, more integration and more functions can be built within the same area. Energy reduction techniques are essential to the design of such systems. The most effective energy reduction method is supply voltage scaling due to the quadratic dependence of energy on voltage. The active dynamic energy dissipation for CMOS circuits is given by

$$E_{act} = C_{avg} V_{DD}^2 \quad (1)$$

where V_{DD} is the supply voltage. C_{avg} is the average switching capacitance and is given by $C_{avg} = C_{gate} + C_{diff} + C_{wire}$ where C_{gate}, C_{diff}, and C_{wire} are the average gate, diffusion, and wire capacitance for the chip respectively. C_{diff} is highly voltage dependent but its average value used here is assumed to be voltage independent.

Peak supply voltage is selected based on peak performance requirements. Occasionally, peak performance is not required by the processing unit. Therefore, supply voltage can be scaled when maximum performance is not required. The software interface can provide information about performance requirements. This information can be used to reduce supply voltage based on the required speed.

Dynamic voltage scaling (DVS) is a feedback system that is used to dynamically control supply voltage according to performance requirements [1–4]. By exploiting the variation in computational needs, the average energy of the system can be reduced while maintaining the same throughput. When supply voltage and operation frequency are dynamically controlled according to the computational load, the average energy can be reduced significantly. As a result, battery life time can be extended.

There are situations, however, when the microprocessor suddenly demands high performance. This requirement can exceed the time response of the DVS system. In such situations, often called panic mode, the supply voltage must be raised to the maximum supply, and then lowered to the optimal value. This paper describes a robust dynamic voltage scaling architecture which efficiently selects the maximum voltage during panic mode. Section II discusses DVS systems in general. Section III describes the architecture of the proposed system. The analysis of the proposed system and comparison to the conventional system are given in Section IV.

II. Dynamic Voltage Scaling System

DVS systems adjust supply voltage according to throughput requirements. Fig. 1 shows the overall architecture of a DVS system. The performance manager uses a software interface to predict performance requirements. Once performance requirement for the next task is determined, the performance manager sets the voltage and frequency just necessary to accomplish the task. The target frequency is sent to the phase-locked loop (PLL) to accomplish frequency scaling. Based on the target voltage, the voltage regulator scales supply voltage to meet performance target.

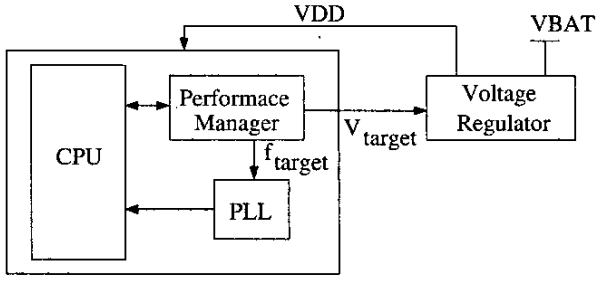

Fig. 1. Architecture of a Dynamic Voltage Scaling System

Setting the target voltage to the voltage regulator can be a one time setting based on a lookup table (LUT) or can be based on the actual system performance. In the LUT configuration, different target frequencies and the corresponding voltage

settings are used to set the voltage to the correct value. The LUT entries must be determined through characterization. In addition, enough margin to accommodate for worst case process and temperature variation is added. Such a margin can be large enough to significantly reduce energy savings. When the target voltage is set, an analog-to-digital converter is used to indicate when the voltage regulator has completed adjusting the voltage.

Instead of using a LUT, the second performance manager configuration is based on measuring the actual system performance using a ring oscillator or a replica of the critical path. When the system performance needs to be changed, the performance manager sets the target frequency. The error between the target and the measured frequencies is used by a feedback system to adjust the voltage to achieve the target performance.

Different parameters are involved when selecting between the two different configurations. Stability against temperature change is a main design parameter. The conventional LUT DVS stores the worst case performance numbers. Therefore, worst case process variation is covered and temperature stability is guaranteed. The large margin added to compensate for process and temperature variation can reduce energy savings significantly.

When monitoring the actual system performance, the system compensates for temperature variation. Closed loop parameters and system response determine the time required by the feedback system to adjust for a temperature change. If the rate of change in temperature is faster than the closed loop response time, the system has to ramp up the voltage to its worst case setting. That worst case setting corresponds to the worst case process and worst case temperature. The panic mode, where voltage has to guarantee the maximum performance under all circumstances, has not been addressed properly in closed loop DVS systems [1, 3].

In this paper, a hybrid between the one time voltage setting and continuous performance monitoring is presented. The proposed system saves more energy by automatically identifying the process. In this case, the system selects frequency and voltage data points which correspond to the split at hand and not the worst case. Once the voltage reaches the target dictated by the LUT, the system starts performance monitoring via a critical path replica to compensate for temperature variation. During panic mode, the system switches back to the LUT and ramps up supply voltage to the maximum specified according to the split. The proposed system is described in detail in the next section.

III. Proposed Dynamic Voltage Scaling Architecture

The proposed system works in two configurations, LUT mode and the performance monitoring mode. During calibration, an automated mechanism to identify the process corner to which the chip belongs is executed. The LUT uses the process identification information extracted to determine the closest performance and voltage data points based on worest case temperature. When performance needs to be changed, the system starts in the LUT configuration. Once voltage is adjusted according to the LUT setting, the system switches to performance monitoring to fine tune performance and compensate for temperature variations. Details of the proposed system shown in Fig. 2 are described below.

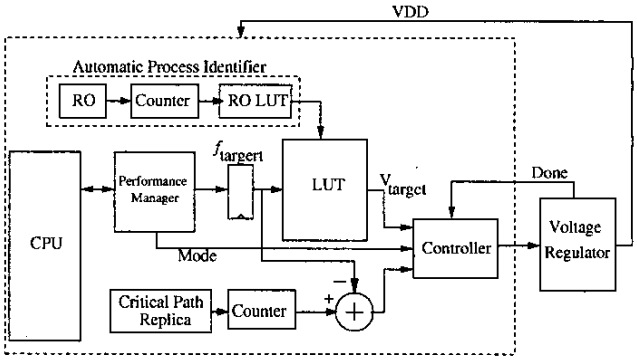

Fig. 2. Architecture of the proposed system

The automatic process identifier is used to identify the process corner during calibration. Process variations and temperature are the main factors directly affecting performance. For example, a slow split at cold temperature can be faster than a typical split at hot temperature. Fig. 3 shows the simulated frequency vs. voltage characteristics for a critical path at different splits in 0.13μm CMOS process. Process identification is difficult to accomplish at high voltages. The distinction between the different frequency characteristics becomes fuzzy due to the larger impact of temperature on performance at high voltages. During calibration, it is necessary to fix temperature at a certain level in order to identify the process corner. Temperature adjustment adds extra time and cost to the calibration process.

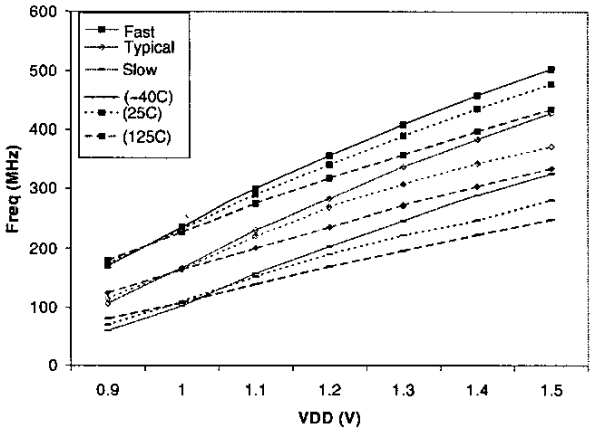

Fig. 3. Critical Path frequency scaling with voltage for different process splits and different temperature

The extra calibration time can be saved when the process corner is identified by measuring performance at a specific voltage for which performance is insensitive to temperature [5, 6]. When temperature changes, performance is affected by two main technology parameters, threshold voltage and channel

mobility. Frequency at which a logic path can operate is given by

$$f = \frac{I_{\text{avg}}}{L_D C_{\text{avg}} V} \tag{2}$$

where L_D is the logic depth. The average capacitance, C_{avg}, can be assumed independent of temperature. The average current, I_{avg}, is proportional to

$$I_{\text{avg}} \propto \mu(T)(V_{\text{DD}} - V_{\text{TH}}(T))^{\alpha} \tag{3}$$

where V_{DD} is the supply voltage, $\mu(T)$ is the channel mobility at temperature T, and $V_{\text{TH}}(T)$ is the threshold voltage at zero bias and at temperature T. Channel mobility and threshold voltage dependence on temperature are given by [5]

$$\mu(T) = \mu(T_0)\left(\frac{T}{T_0}\right)^{-M} \tag{4}$$

and

$$V_{\text{TH}}(T) = V_{\text{TH}}(T_0) - \kappa(T - T_0) \tag{5}$$

where $T_0 = 300K$, M is the mobility temperature exponent, κ is the threshold voltage temperature coefficient. Typical values for M and κ are 1.5 and 1.8 mV/K respectively.

By lowering the supply voltage, the temperature effect on threshold voltage starts to cancel out the temperature effect on mobility. At a specific voltage, logic performance becomes insensitive to temperature. This voltage is independent of the type of logic implemented. Therefore, it can be used to identify the split since temperature effect has been canceled out and the only influence on performance is through process variations.

Fig. 3 indicates that at approximately 1.0 V, the critical path frequency is insensitive to temperature across all process corners. A small ring oscillator (RO) is used to identify the process. A RO LUT is built using the characterized RO frequencies for different splits at the temperature insensitive voltage as shown in Table I. The LUT entries are indexed by the RO frequency. During calibration, the voltage is set to the temperature insensitive value and the RO frequency is read. The split can be identified using the RO frequency. If the frequency lies between two expected values then the split happens to be between two corners. The system must select the slower corner. For example, if the split lies between fast and typical corners, the typical corner is selected.

TABLE I
RO LUT FOR PROCESS IDENTIFICATION

Index	Process
f_{ro_1}	Process_1
f_{ro_2}	Process_2
...	

Split compensation can be accomplished by characterization of different splits at worst case temperature. The example of three splits (slow, typical, and fast) is considered here. Characterization data is then stored in a lookup table (LUT). The form of the LUT is indicated in Table II. Total number of rows in the table is equal to the required number of the different target frequencies set by the software interface. The voltage settings corresponding to the identified split are selected and all other voltage entries in the LUT are ignored.

TABLE II
LUT FOR SPLIT COMPENSATION

f	Slow	Typical	Fast
f_1	V_{s1}	V_{t1}	V_{f1}
f_2	V_{s2}	V_{t2}	V_{f2}
...			

The proposed system works as follows. The automatic process identifier identifies the process during system calibration phase. The target voltage is set to the temperature insensitive value. The RO frequency indexes the different values stored in the RO LUT. Accordingly, the RO identifies the process to the main LUT. The target voltage is set according to the target frequency for the process identified . Target voltage is used by the voltage regulator to adjust supply voltage to reach the target.

Once the voltage settles at the LUT target voltage, the system switches to the performance monitoring mode. The target frequency is compared to the frequency of a critical path replica for voltage fine tuning. A small voltage margin is added to compensate for any mismatch between the real critical path and its replica. The system switches back to the LUT when performance is to be increased or during panic mode to handle a drastic temperature change. The voltage is set to the peak voltage required by the split rather than the worst case split. The energy saving of the proposed system compared to the conventional system is analyzed in detail in the next section.

IV. ANALYSIS AND COMPARISON

When the proposed DVS system works in the LUT mode, energy can be saved compared to conventional systems by setting supply voltage to the worst case temperature for the closest split. On the other hand, conventional DVS systems set supply voltage for worst case split (slow). Fig. 4 shows the voltage distribution as a result of process variations (assuming a Gaussian distribution) at a fixed frequency. For the slow split, the voltage is maximum while for a fast split, voltage can be reduced to V_{min} and still maintains the same performance.

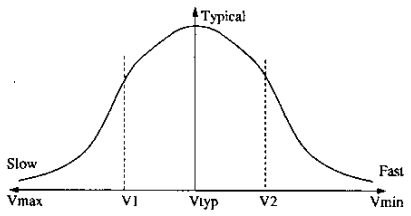

Fig. 4. Voltage Distribution due to Process Variation at a fixed frequency

From (1), energy has a quadratic dependence on voltage. As the number of entries in the LUT is increased, the expected savings are increasing due to the finer granularity the system can offer. The effect of increasing the number of entries on

energy savings is analyzed by considering the 3-σ process distribution.

The LUT contains information about three different splits, slow, typical, and fast. Assuming a 3-σ distribution, then the cumulative probability density function (CDF) of having the voltage between fast and typical voltage is 50% and having the voltage at fast conditions is only 1%. Therefore, 50% of the parts can save energy by reducing supply voltage from V_{max} to V_{typ} while only 1% of the parts would benefit from reducing voltage from V_{max} to V_{min}. Then, the energy saving resulting from using voltage-frequency data of three different splits in the LUT is given by

$$E_{savings} = 0.5\left[1-\left(\frac{V_{typ}}{V_{max}}\right)^2\right] + 0.01\left[1-\left(\frac{V_{min}}{V_{max}}\right)^2\right] \quad (6)$$

Taking $V_{min} = 1.0V$ and $V_{max} = 1.5V$ for a frequency of 200 MHz as shown in Fig. 3, energy saving is around 15%. If the number of splits stored in the LUT is increased, energy savings are increased. For example, assuming that 4 different splits, slow, -σ, +σ, and fast, are used in the LUT. CDF for -σ and +σ are 62.5% and 18.75% respectively. Energy saving becomes 20%.

Fig. 5 shows the trend of energy savings when the number of entries (splits) of the LUT is increased. Energy saving is limited to approximately 29% using voltage and frequency data points for 40 different splits. However, using more than 10 different splits adds only 1% of savings. Furthermore, since the temperature insensitive voltage is not exactly equal for all splits, increasing the number of entries in the LUT might result in a less accurate process identification.

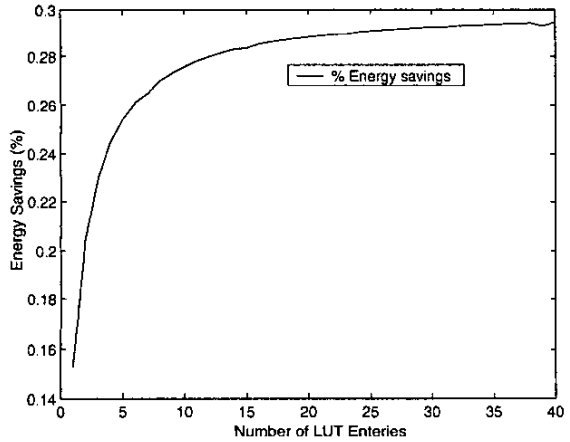

Fig. 5. Energy Savings vs. number of entries in the LUT

When the proposed system enters the panic mode and switches from the LUT mode to the performance monitoring mode, energy saving becomes highly application depndent. However, energy savings can still be achieved compared to the conventional system. Fig. 6 shows the voltage waveform of both systems when entering the panic mode. Assuming that the split which include the proposed system is faster than the conventional. The voltage ramps up to a lower level than in the conventional case. Eventually, both systems settle at the same voltage level but the conventional takes a longer time, $t_2 > t_1$, since it goes to V_2 while the proposed system goes to V_1, where $V_2 > V_1$. The discharge rate of both conventional and LUT case is assumed to be the same for the same load.

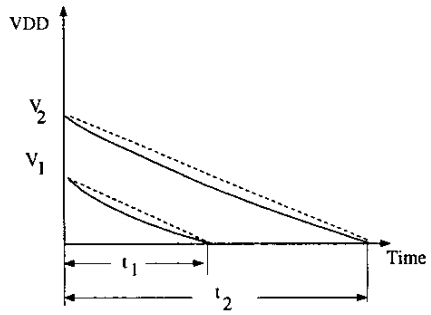

Fig. 6. Voltage Waveform when going to panic mode

Energy saving during performance monitoring mode depend on the number of times performance needs to be increased and how often the system enters the panic mode. The more the unnecessary performance glitches, the more energy savings the proposed system can achieve. This is highly application dependent and can vary from one system to the other.

V. Conclusion

In order to meet the challenges of increased energy dissipation in portable devices, a dynamic voltage scaling architecture was presented. The architecture regains the extra energy lost due to worst case characterization used in conventional systems. A lookup table based approach was presented. The LUT holds characterization data for performance vs. voltage scaling of the critical path for three different process corners. Number of entries can be increased to gain more savings. Process is identified using an automated process identifier. The frequency-voltage entries corresponding to the identified split are used to set the voltage when performance is to be tuned. Energy savings can be up to 29% compared to conventional systems.

References

[1] T. Burd, T. Pering, A. Stratakos, and R. Brodersen, "A Dynamic Voltage Scaled Microprocessor System," *IEEE Journal of Solid State Circuits*, vol. 35, no. 11, pp. 1571–1580, November 2000.

[2] A. Dancy, R. Amirtharajah, and A. Chandrakasan, "High-Efficiency Multiple-Output DC-DC Conversion for Low-Voltage Systems," *IEEE Journal of Solid State Circuits*, vol. 8, no. 3, pp. 252–263, June 2000.

[3] G. Wei and M. Horowitz, "A Fully Digital, Energy-Efficient Adaptive Power-Supply Regulator," *IEEE Journal of Solid State Circuits*, vol. 34, no. 4, pp. 520–528, April 1999.

[4] J. Kim and M. Horowitz, "An Efficient Digital Sliding Controller for Adaptive Power-Supply Regulation," *IEEE Journal of Solid State Circuits*, vol. 37, no. 5, pp. 639–647, May 2002.

[5] A. Bellaouar, A. Fridi, M. Elmasry, and K. Itoh, "Supply Voltage Scaling for Temperature Insensitive CMOS Circuit Operation," *IEEE Transactions on Circuits and Systems II*, vol. 45, no. 3, pp. 415–417, March 1998.

[6] K. Kanda, K. Nose, H. Kawaguchi, and T. Sakurai, "Design Impact of Positive Temperature Depedence on Drain Current in Sub-1-V CMOS VLSIs," *IEEE Journal of Solid State Circuits*, vol. 36, no. 10, pp. 1559–1564, October 2001.