16.5 Ultra-Dynamic Voltage Scaling Using Sub-threshold Operation and Local Voltage Dithering in 90nm CMOS

Benton H. Calhoun, Anantha Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA

Dynamic voltage scaling (DVS) has become a standard approach for reducing power when performance requirements vary. Voltage dithering was proposed to provide near-optimum DVS power savings with much less overhead [1]. Voltage dithering refers to operating for different fractions of time at two discrete voltage and frequency pairs to achieve an intermediate average frequency. Previous implementations apply voltage dithering to entire chips and require many microseconds to change operating voltage [1,2]. In this paper, a 90nm adder test chip that demonstrates the proposed concept of local voltage dithering *(*LVD*)* and couples LVD with sub-threshold operation to achieve ultra-dynamic voltage scaling (UDVS), is described.

The LVD technique uses embedded power switches to toggle between a small number of voltage levels at the local block level. Distributing these switches locally allows each block to minimize energy consumption using voltage dithering based on its own workload rather than a chip-wide workload. Since workload commonly differs among blocks, the locally optimum dithering rates reduce power relative to the chip-wide approach. Also, the embedded switches can respond more quickly to workload variations than a global switch. Finally, locally embedded dithering switches can be turned off to provide fine-grained power gating essentially for free.

Figure 16.5.1 shows the 32b Kogge-Stone adder block that can be configured as an accumulator for testing LVD on the chip. Two PMOS header switches select between two supply voltages for the adder and for a critical path replica ring oscillator that generates the correct clock frequency for each voltage. The die micrograph in Fig. 16.5.7 shows the accumulators with different numbers of header switches used for testing, and the approximate area of a single header switch is highlighted for reference.

Figure 16.5.2 illustrates the savings that LVD provides for the adder block on the test chip when the normalized clock frequency, or *rate*, varies. The dotted line shows operation at the highest rate followed by ideal shutdown. The solid line shows the measured energy for DVS assuming ideal voltage and frequency scaling. Selecting two rates from the ideal curve (1 and 0.5 in the figure) and operating for the correct fraction of time at each rate achieves the desired average performance with energy that nears the ideal. This *dithering* results in the dashed curve. A local block with three headers can achieve closer to optimum savings by selecting three rates and then dithering to connect those points on the plot. Figure 16.5.2 also shows measured ring oscillator frequency versus V_{DD} across temperature. Higher temperature has less impact on the frequency above threshold, but the increase in V_T with temperature in the sub-threshold region increases the frequency.

Figure 16.5.3 shows the test circuit used to measure the delay overhead of LVD. While the adder runs a long accumulation, V_{DD} dithers between V_{DDL} and V_{DDH}. When the headers toggle V_{DD}, a counter gates the clock for a specified number of cycles to ensure settling at the new voltage. Measurements shows that the correct value is accumulated with only 1/2 cycle (minimum possible using the test circuit) of clock gating for V_{DDL} above 0.6V. Thus, even conservative settling times for this LVD implementation are on the order of nanoseconds. The scope plot shows the dithering signal and the output clock divided by 1024. Measurements show that the effective switched capacitance of the entire switching network to toggle the header switches is 3.7× the average switched capacitance for an addition, so the switching energy overhead is compensated by spending only a small number of cycles at the lower voltage (12 cycles for V_{DDL}=0.9V, rate=0.5).

The discussion to this point has assumed that the varying rate remains at or above roughly a few percent of one. However, some applications require brief periods of high performance and also spend significant amounts of time operating at effective rates that are orders of magnitude below one. Minimum energy operation is demonstrated to occur in the sub-threshold region in [3]. In the absence of a performance constraint, energy is minimized by operating at the minimum energy point, which occurs because of increased leakage energy at low frequency, and then shutting down if there is more timing slack.

The minimum energy per operation point measured for the adder appears in Fig. 16.5.4 at V_{DD}=330mV (f=50kHz) and 0.1pJ per addition for 25°C. Figure 16.5.4 also shows the measured effect of temperature on the total energy per cycle and leakage energy per cycle. The leakage energy increases quickly with temperature for $V_{DD}>V_T$ because of the exponential rise of sub-V_T current. In sub-V_T operation, the increased current also decreases the cycle delay exponentially, so the measured effect of temperature on the minimum energy point is small, validating the model in [4]. Figure 16.5.5 shows a scope plot of sub-threshold operation just below the minimum energy point.

LVD is less effective when it spans into the subthreshold region. The exponential increase in delay in sub-threshold means less time is spent at the lower voltage for a given rate, so the dithering curve is closer to the shutdown case than to the ideal DVS curve. Operating at the minimum energy point is still worthwhile for systems that infrequently require high performance, only it makes more sense as a near-standby mode rather than for dithering to average the processing rate.

Since LVD works well for high speed operation and operating at the minimum energy point is optimal for low performance situations, ultra-dynamic voltage scaling (UDVS) using local power switches is proposed. Figure 16.5.6 provides one example of measured UDVS characteristics for the adder. In this example, dithered voltages are chosen at 1.1V, 0.8V, and 0.33V, which is the optimum voltage for minimum energy. When the adder block is performing operations with no timing deadline, it functions at the minimum energy point at 50kHz and saves 9× the energy versus the ideal shutdown scenario. When performance becomes important, the adder dithers between 1.1V and 0.8V within 30% of the optimal energy consumption while adjusting for variations in the rate above 0.1. When the average rates for a given application are known, the dithered voltages can be set to match those rates. This brings the dithered curve closer to the optimum DVS curve for the common cases. UDVS is achievable using three header switches. When transitions to and from high-performance mode are infrequent, two header switches may be paired with one adjustable DC/DC converter for the same functionality. For example, during low-speed operation, the headers dither between 0.33V and 0.6V. When the rare transition to high speed occurs, the DC/DC converter switches one voltage so that dithering is between 1.1V and 0.6V. Application of UDVS combined with LVD provides flexible energy-awareness capabilities for a variety of operating scenarios.

Acknowledgements:
This work was funded by DARPA through a subcontract with MIT Lincoln Laboratory and by Texas Instruments. We thank Texas Instruments for chip fabrication.

References:
[1] V. Gutnik and A. Chandrakasan, "Embedded Power Supply for Low-Power DSP," *IEEE Trans. on VLSI Systems*, vol. 5, no. 4, pp. 425-435, Dec., 1997.
[2] H. Kawaguchi et al., "A Controller LSI for Realizing VDD-Hopping Scheme with Off-the-Shelf Processors and Its Application to MPEG4 System," *IEICE Trans. Electron*., vol. E85- C, no.2, pp. 263-271, Feb., 2002.
[3] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuit Techniques," *ISSCC Dig. Tech. Papers*, pp. 292-293, Feb., 2004.
[4] B. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," *ISLPED*, pp. 90-95, 2004.

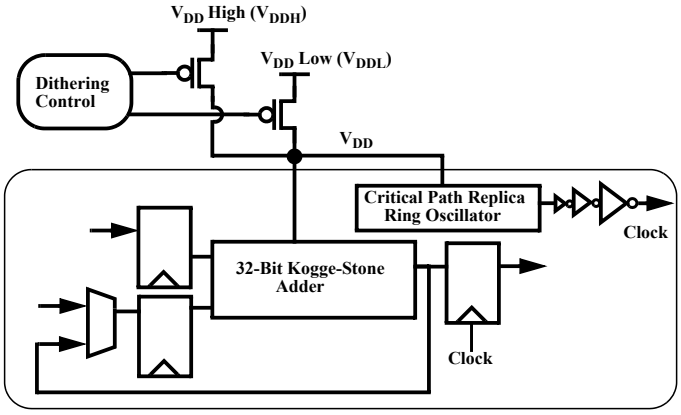

Figure 16.5.1: Block diagram of voltage dithered adder and critical-path replica using two local header switches for local voltage dithering (LVD).

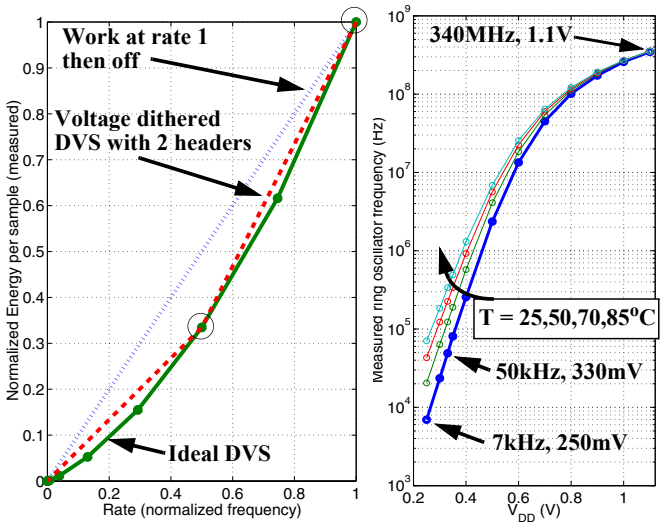

Figure 16.5.2: Characterizing voltage dithering using measured results for the adder. For rates near 1, frequency does not vary much with temperature.

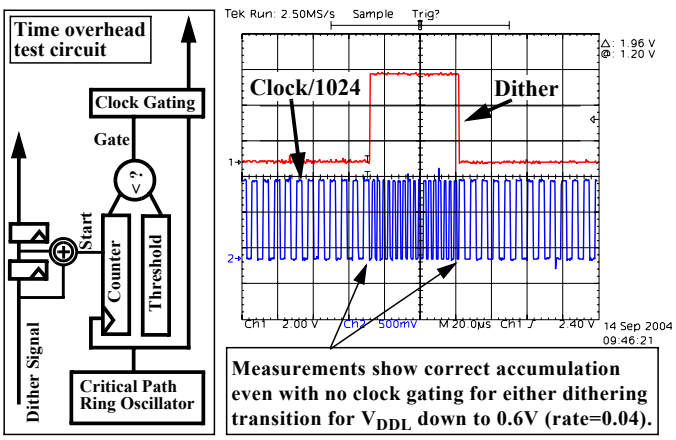

Figure 16.5.3: Measuring timing overhead of LVD between rate 0.5 (170MHz) and rate 1 (340MHz).

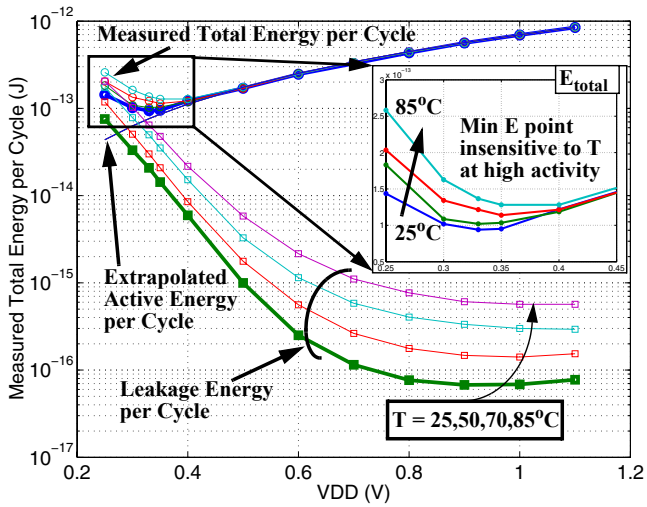

Figure 16.5.4: Measured energy per cycle in the sub-threshold region for input activity of one (accumulator mode). Minimum energy point occurs at 330mV (50kHz) and 0.1pJ per operation at 25°C.

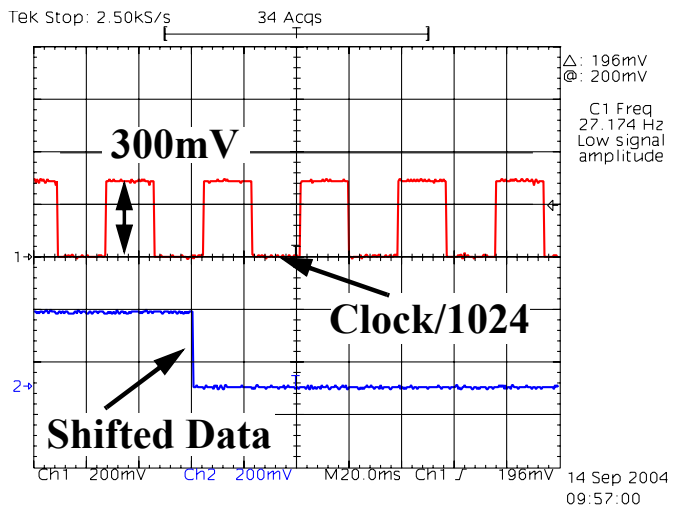

Figure 16.5.5: Sub-threshold operation just below the minimum energy point.

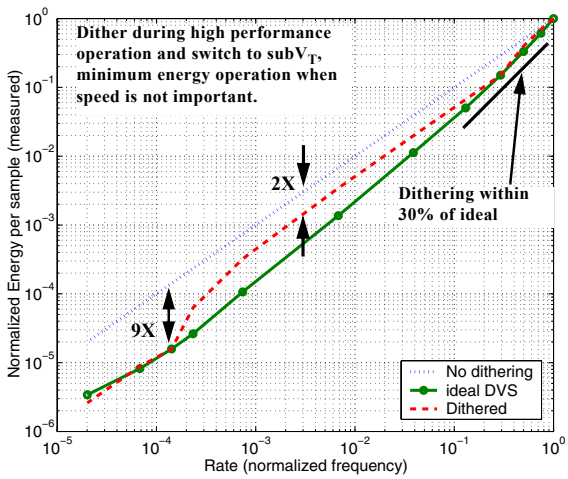

Figure 16.5.6: Ultra-dynamic voltage scaling (UDVS) using two headers with one variable DC-DC converter or using 3 headers. Different choices of dithered voltages make dithering closer to the ideal case over different ranges of rates.

Continued on Page 599

16

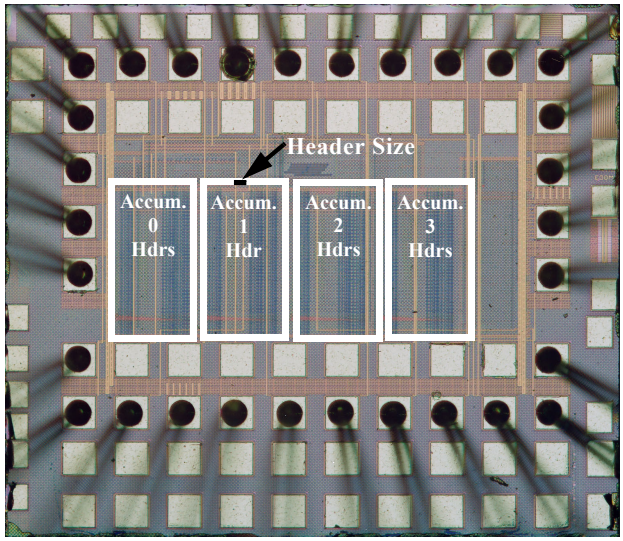

Figure 16.5.7: Annotated die micrograph showing accumulators with 0, 1, 2 and 3 headers. The size of one header is highlighted for reference.

