9

System Examples

This chapter gives two examples of systems that operate in sub-threshold. First, a sub-threshold FFT is described. The FFT system employs the concept of energy-aware architectures. Energy-aware architectures contain hooks that allow the user to gracefully scale energy and quality depending on the system conditions. Exercising these power hooks causes changes in the workload and activity factor of the system, which in turn impacts its energy characteristics. Measurements of the FFT test chip show that, as activity factor varies, the minimum energy point also changes.

Second, a Local Voltage Dithering (LVD)-UDVS system extends the voltage range of traditional DVS systems. LVD improves on existing voltage dithering systems by taking advantage of faster changes in workload and by allowing each block to optimize based on its own workload. Additionally, measurements show that the time and energy overhead of LVD are small. UDVS also provides a practical method for extending DVS into the sub-threshold region. For many emerging energy-constrained applications, lowering energy consumption is the primary concern under most conditions. Thus, operating at the minimum energy point conserves energy at the cost of lower performance (frequency). This type of application works at the minimum energy point primarily and only jumps to higher performance voltages in rare cases. Chip measurements have shown the effectiveness of UDVS for this scenario.

9.1 A Sub-threshold FFT Processor

Sub-threshold operation is, as previous discussed, well-suited for wireless sensor nodes. The lifetime of a sensor node depends on the battery capacity and the ability of the node to compute and communicate in an energy-efficient fashion. Communication through an RF link expends a great deal more energy than computation, therefore an efficient sensor system performs sensor signal processing on the data and only transmits the resulting necessary infor-

mation. This reduces the sensor bandwidth considerably and leads to longer sensor lifetimes.

To achieve minimal energy dissipation, a sub-threshold DSP is needed for sensors. A highly flexible node architecture contains both hardware accelerators and a programmable DSP [14]. Hardware accelerators can do signal processing algorithms extremely efficiently both in speed and energy dissipated. Sensor signal processing algorithms that are commonly used can be implemented in accelerators. The programmable DSP provides system control and implements functions that are not covered by the accelerators.

Examples of algorithms used in sensor nodes are beamforming, classification, direction sensing, etc. The Fast Fourier Transform (FFT) is an algorithm that is commonly used in sensor signal processing. This section shows details of a sub-threshold FFT implementation. The FFT uses various concepts described in Chapter 6 and Chapter 7. First, an energy-aware FFT architecture is described [189]. Energy-awareness provides power hooks into the architecture to allow the user to trade-off between energy and quality. This becomes important for sensors that must operate over a wide range of operating conditions. Next, the energy-performance contours for the FFT are analyzed. The energy-performance analysis describes the optimal operating point which occurs in the sub-threshold region at 400mV. Finally, chip measurements show functionality of the FFT down to 180mV and the optimal operating voltage as a function of varying the FFT bit precision and computation length.

9.1.1 The Fast Fourier Transform

The Fast Fourier Transform (FFT) is a widely used algorithm that appears in applications such as speech processing, signal detection, communications, and tracking. The FFT extracts the frequency and phase information from the sensor signals. Dedicated low-power FFT processors are able to sustain low-power requirements of various embedded applications [190].

Sensor data is considered "real-valued," which means that the imaginary part is zero. Traditionally, the FFT assumes complex input data. When an FFT is performed on real-valued data, the output is conjugate-symmetric. This means that, for an N-point FFT, the first N/2 points are unique, and the last N/2 points are symmetrically redundant. This symmetry is exploited by the real-valued FFT algorithm (RVFFT). The Real-Valued FFT (RVFFT) uses the symmetries inherent to computing the complex-valued FFT (CVFFT) on real-valued inputs to reduce overall computation. For example, a 1024-pt. RVFFT is efficiently performed by computing a 512-pt. CVFFT and then transforming the outputs back for 1024-pt. The computation effort of the RVFFT is approximately half that of the CVFFT.

A simple architecture for a RVFFT processor consists of a traditional CVFFT followed by backend processing. Figure 9.1 shows the conventional radix-2 butterfly architecture for the CVFFT using in-place computation. In-place computation occurs when a value is read out from the memory and is

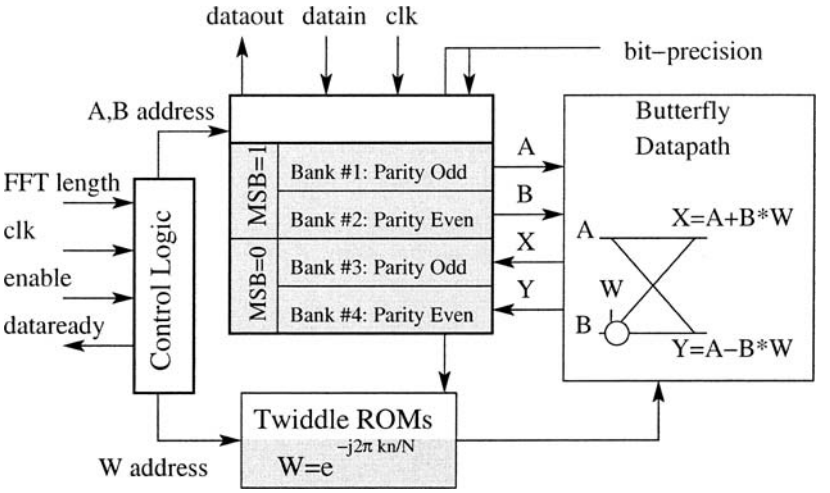

Fig. 9.1. Radix-2 butterfly FFT architecture. (© 2005 IEEE)

re-written to the same location in the next clock cycle. The advantage of in-place computation is that it requires the minimum sized buffer. For a vehicle-tracking sensor application, the maximum memory size was 1024-Words x 16-bit, which is the size of one frame of data. The FFT processor memory uses the register file design in Section 7.1. The memory was simulated and designed to operate to 100mV with a typical transistor model.

Additionally, a read-only memory (ROM) is needed for the storing twiddle factors (W). Twiddle factors are complex values used in the FFT to shift the phase of the input values.

9.1.2 Energy-Aware Architectures

In a sensor network, the environment is constantly changing. Energy-aware architectures are used to efficiently trade-off between energy and quality given current operating conditions. For example, a sensor with a full battery can provide very high quality sensor results and performance. When the battery energy is low, then the sensor can output lower quality results at a lower energy consumption and stretch out its battery lifetime.

The FFT architecture is designed with various power hooks that allow the architecture to gracefully scale FFT length and bit precision. On the same architecture, the sensor can perform a short 128-pt. FFT with 8-bit precision for a low-quality/low-energy result or perform a 1024-pt. FFT with 16-bit precision for a high-quality/high-energy result.

One of the hooks designed into the energy-aware FFT processor is variable bit-precision. The concept of variable bit-precision is showcased by the Baugh-Wooley (BW) multiplier in the butterfly datapath. A traditional fixed

bit-precision BW design optimizes for the worst case scenario by building a single multiplier for the largest bitwidth. This design is non-optimal because, when lower precision multiplications are performed, the sign extension bits cause significant switching energy overhead. The proposed scalable BW multiplier design recognizes that the MSB quadrant contains a lower bit-precision multiplier. (The MSB quadrant includes those gates associated with the MSB inputs).

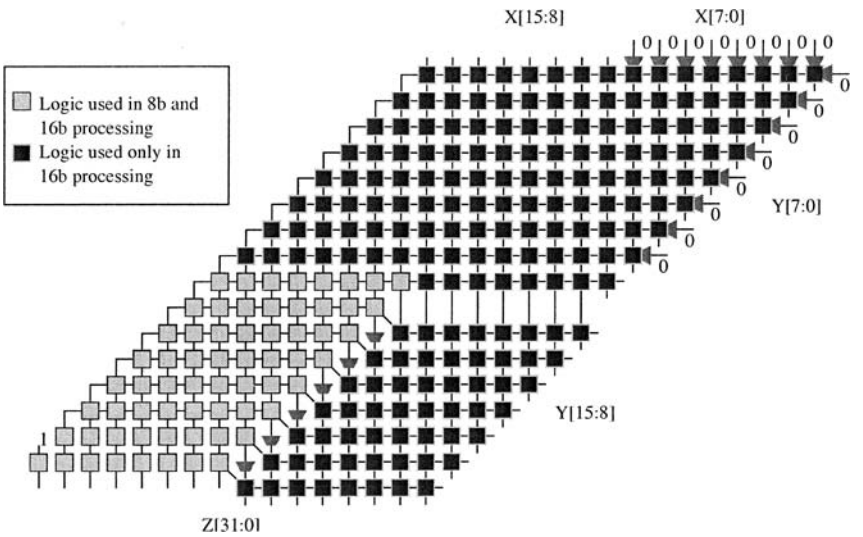

Fig. 9.2. 8b and 16b scalable Baugh-Wooley multiplier. (© 2005 IEEE)

To minimize switching in the LSB adders, the LSB inputs are gated, and only the MSB inputs are used to process data. Figure 9.2 demonstrates this technique for a BW multiplier that is scalable for 8-bit and 16-bit precisions. Similar bit-precision scalability was applied to the entire butterfly datapath, data memories and Twiddle ROMs. There is a 44% savings for 8-bit processing, but at the expense of 3% cost at 16-bit processing. 16-bit energy-aware processing has overhead due to the additional gates added to enable energy-awareness.

9.1.3 Minimum Energy Point Analysis

The FFT processor is designed to operate at the optimal operating point that minimizes energy dissipation. Analysis of the energy and performance of the FFT shows that the minimum energy point occurs at supply voltage levels below the threshold voltage. The energy and performance of the FFT were analyzed based on the theory discussed in Section 4.1. Figure 9.3 shows

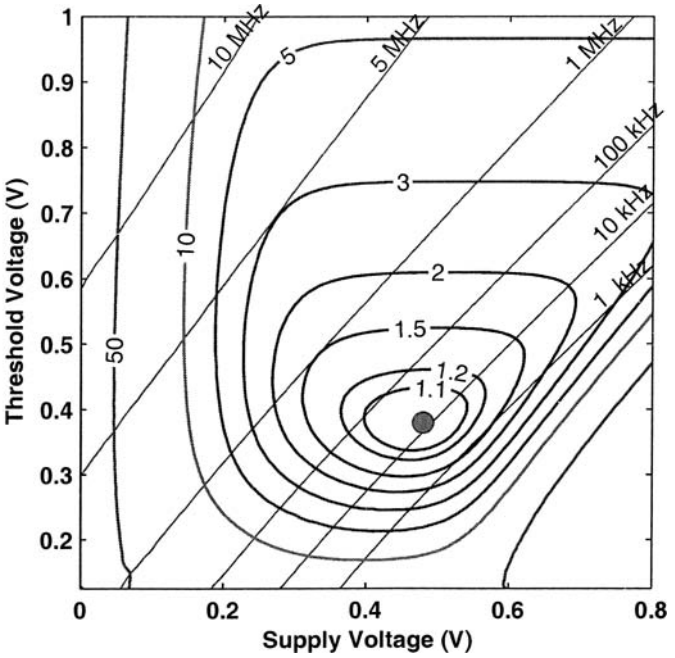

Fig. 9.3. Minimum energy point and constant energy and performance contours of the 16-b and 1024-pt. FFT. (© 2005 IEEE)

simulated energy contours of the 16-bit 1024-pt. FFT for a supply voltage range of 100mV-1V and threshold voltage range of 0V-800mV. The FFT was designed in a 0.18μm process. The figure shows the simulated average energy per FFT and performance across the entire supply and threshold voltage range using the switching and leakage models from Section 4.2.1.

The energy contours (circular) show that the minimum energy dissipation point occurs at V_{DD}=380mV and V_T=480mV. The performance contours show the frequency at the minimum energy point to be 13kHz.

The FFT processor is designed and fabricated in a standard 0.18μm bulk CMOS process with a fixed nominal threshold voltage of 450mV. Figure 9.4 shows an energy simulation at a fixed threshold voltage of 450mV. The predicted minimum energy point of the FFT processor occurs at 400mV. The FFT was designed and simulated at voltages much below 400mV to allow a thorough exploration of the space around the minimum energy point.

9.1.4 Measurements

The FFT processor was designed using a sub-threshold standard cell library, custom multiplier generators, and custom register file and ROM generators.

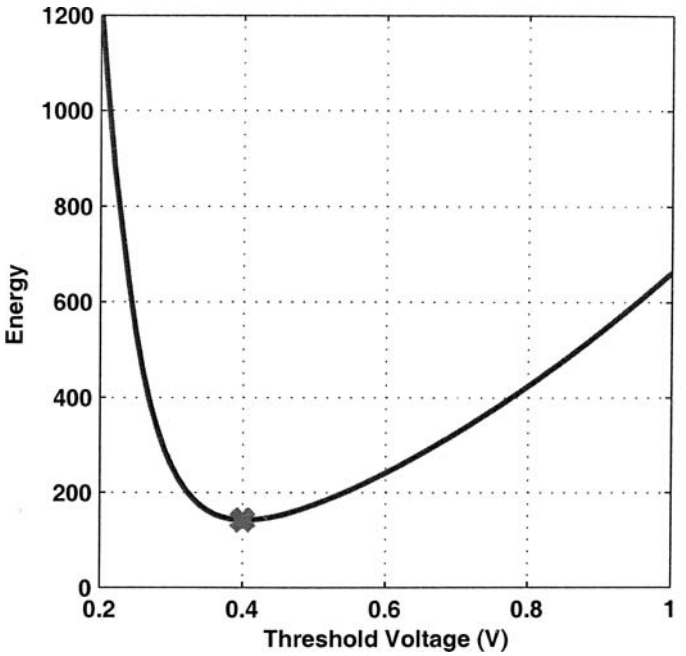

Fig. 9.4. Estimated minimum energy point for fixed V_T=450mV. (© 2005 IEEE)

The generators use specialized logic cells and ensure compact layout. The functional logic blocks in the datapath, control logic and memories were synthesized using the cell library.

The 0.18-μm CMOS FFT processor occupies 2.6 X 2.1 mm^2 and contains 627K transistors. It is fully functional at 128, 256, 512, and 1024 FFT lengths and for 8-b and 16-b precision, at voltage supply levels from 180 to 900mV with clock frequencies of 164Hz to 6MHz at these respective voltages. The power dissipated at 180mV is 90nW for 16-b 1024-pt. operation.

Figure 9.6 shows the measured energy consumption for 8-bit and 16-bit processing as a function of voltage. 8-bit processing has a lower activity factor and thus has lower switching energy. However, because the leakage energy is the same for both 8-b and 16-b processing, the minimum energy point increases. The minimum energy point for 16-bit occurs at 350mV and for 8-bit occurs at 400mV. The power dissipated at the 16-b optimum is 600nW at a clock frequency of 10 kHz, and the energy dissipated is 155nJ/FFT.

For an FFT sensor application benchmark, it proved to be 350X more energy efficient than a typical low-power microprocessor and 8X more energy efficient than a standard ASIC implementation.

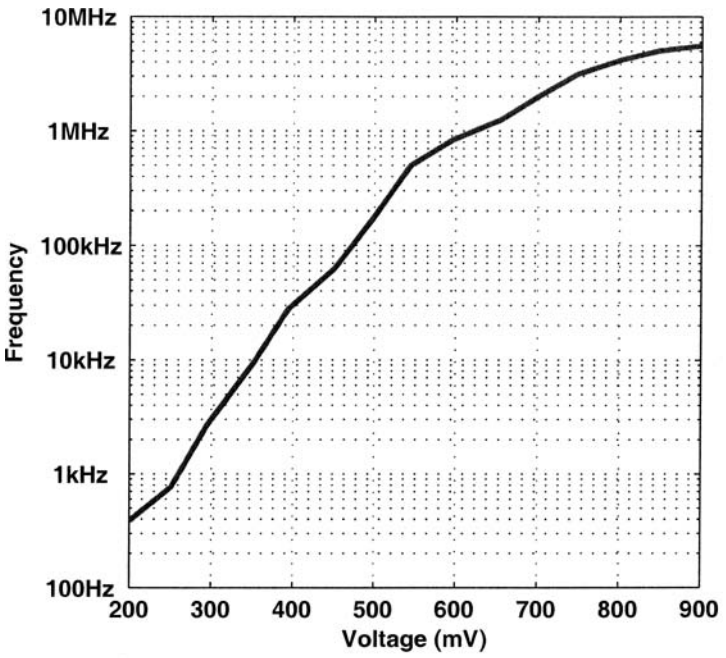

Fig. 9.5. Clock frequency as a function of voltage. (© 2005 IEEE)

9.2 Ultra-Dynamic Voltage Scaling

DVS has become a standard approach for reducing power when performance requirements vary. DVS systems lower the frequency and voltage together to reduce power when lower performance is allowed [191]. In the first true DVS implementation, a critical path replica is used in a feedback loop to adjust the supply voltage to the lowest value that allows the delay to match a given reference frequency [191]. DVS now appears in commercial processors such as, for example, the Intel XScale [192], IBM PowerPC [193], and the Transmeta Crusoe processor [194].

Voltage dithering was proposed as a low overhead implementation of DVS to provide near-optimum power savings using only a few discrete voltage and frequency pairs [68]. The savings are only achievable if the voltage and frequency can change on the same time scale as the altering workload. Previous implementations apply voltage dithering to entire chips and require many microseconds to change operating voltage [68][195]. This section describes a 90nm test chip that demonstrates a proposed concept of Local Voltage Dithering (LVD) and couples LVD with sub-threshold operation to achieve Ultra-Dynamic Voltage Scaling (UDVS) [196]. We provide measurements of

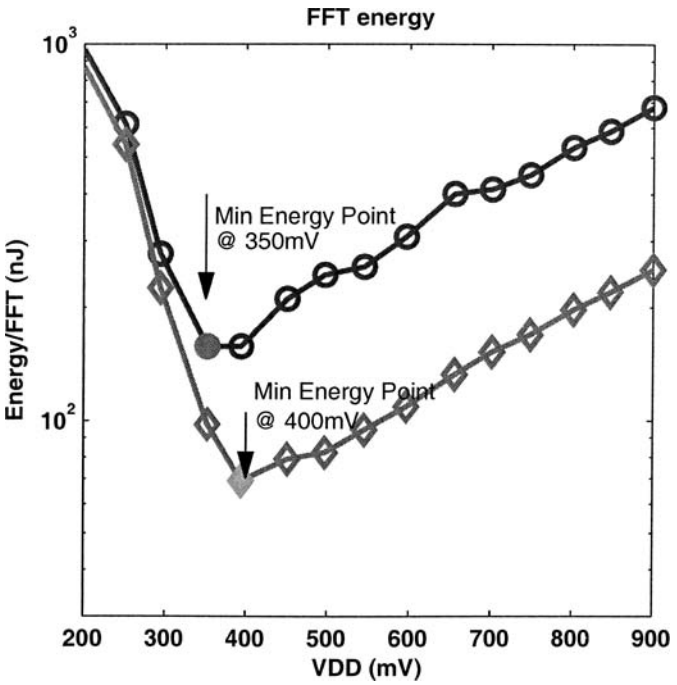

Fig. 9.6. Effect of activity factor on minimum energy point for 8-bit and 16-bit processing.

the effect of temperature on minimum energy operation for the 90nm test chip.

9.2.1 DVS and Local Voltage Dithering

Many signal processing systems process blocks of data that arrive at some regular rate, and sometimes the amount of data to process is less than the maximum amount. This corresponds to a fixed-throughput system whose workload requirements change on a block-to-block basis in a time varying fashion. Examples of this type of application include MPEG video processing and FIR filtering with a variable number of taps [68]. In the video processing example, the maximum workload corresponds to a scene change in the video sequence. In this case, the entire new frame of data requires processing since it is completely different from the previous frames. In the absence of scene changes, new frames of data may not differ significantly from the previous frame, so only a small section of the new frame requires processing. This case represents a reduced workload for the system. The workload of the system measures the amount of processing required for a given block of data, and the rate is simply

the normalized processing frequency [68]. In a system without buffering, the lowest allowable rate equals the workload. If buffering is possible, then there are different strategies involving operation at different rates that can correctly perform the required processing on a block with a given workload by ensuring that the average rate equals the workload. There are many applications where workload varies with time [197], and policies for setting the rate based on incoming data have been explored [198][199][200].

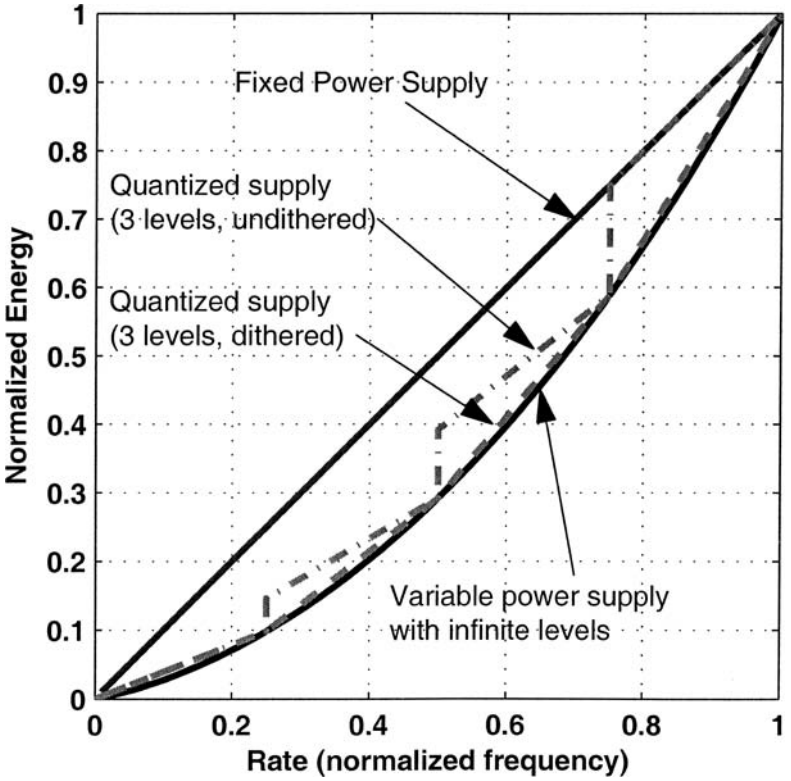

Fig. 9.7. Theoretical energy consumption versus rate for different power supply strategies [68]. (© 2006 IEEE)

Figure 9.7 shows four approaches to power supply management for reducing energy consumption when the workload varies [68]. It plots the required rate of the system versus the normalized energy required to process one generic block of data. The most straightforward method for saving energy when the workload decreases is to operate at the maximum rate until all of the required processing is complete and then to shutdown. This approach only requires

a single power supply voltage (corresponding to full rate operation), and it results in linear energy savings. The fixed power supply curve in Figure 9.7 assumes ideal shutdown (i.e. - no shutdown power). A variable supply voltage with infinite allowable levels provides the optimum curve for reducing energy. This curve in Figure 9.7 corresponds to theoretically ideal DVS according to the model in [68] where velocity saturation is omitted. When velocity saturation occurs, the energy savings for ideal DVS increase because the performance does not decrease as quickly for the same change in V_{DD} [195].

Voltage Dithering

One method that avoids the problem of creating an infinite number of supply voltages is to use quantized supply voltages. In Figure 9.7, three levels of supply voltage quantization are used with two different policies. The undithered policy simply selects the lowest supply voltage for which the rate exceeds the desired rate, operates at that rate and voltage until all of the data in the block is processed, and then shuts down. This results in the stair-step energy characteristic. A better method is called voltage dithering [68]. The basic idea behind voltage dithering is to divide the computation of one block of data between operation at the quantized supply voltage and rate pairs that occur above and below the desired average rate. The energy profile for dithering between quantized voltage supplies linearly connects the quantized rate and energy pairs on the plot. Assuming that the desired rate of operation for a block, R_{BLOCK}, lies between two quantized rates, R_{LOW} and R_{HIGH}, then:

$$E_{BLOCK} = E_{LOW}\left(\frac{R_{HIGH} - R_{BLOCK}}{R_{HIGH} - R_{LOW}}\right) + E_{HIGH}\left(\frac{R_{BLOCK} - R_{LOW}}{R_{HIGH} - R_{LOW}}\right) \tag{9.1}$$

where E_{HIGH} and E_{LOW} are the normalized energies consumed for processing a block at R_{HIGH} and R_{LOW}, respectively.

Figure 9.8 shows an example comparing voltage dithering with fixed and variable supply approaches. This example uses two quantized voltages that provide full rate and half rate operation. For a desired rate of 0.6, the fixed supply approach operates at the highest rate and full power for 0.6 of the full block time (Figure 9.8(b)). Ideal variable voltage operation provides exactly 0.6 rate at the best possible energy (Figure 9.8(d)). Voltage dithering gives an average rate of 0.6 by operating for 20% of the block time at full rate and for 80% of the time at 0.5 rate. The resulting energy consumption is thus averaged between the two quantized points and falls on the connecting line (Figure 9.8(c)). This approach allows a good approximation of the optimum energy profile with less overhead.

Implementations of systems that use voltage dithering apply it monolithically to an entire chip. The system in [68] uses an on-chip variable DC-DC converter to dither the voltage supplied to the entire chip. A chip containing

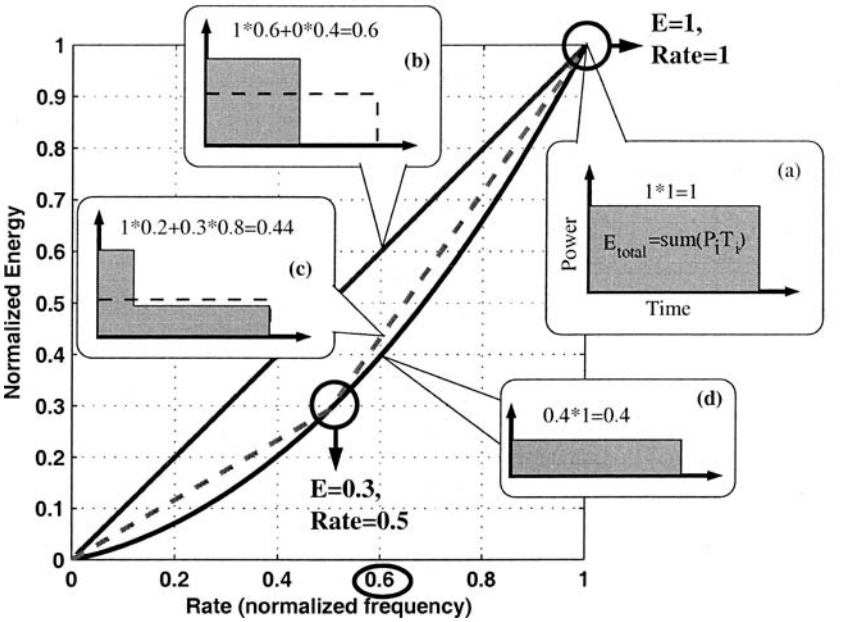

Fig. 9.8. Voltage dithering example for 0.6 rate normalized to full rate (a). Example shows fixed supply (b), voltage dithering (c), and ideal variable supply (d).

header switches was used in [195] to select the voltage supplied to a different chip (off-the-shelf processor). A similar off-chip voltage hopping approach is used in [201] for a zero V_T processor in fully depleted Silicon on Insulator (SOI). These implementations have shown the effectiveness of voltage dithering to save energy for high performance applications with variable workload.

Local Voltage Dithering

Applying voltage dithering at the local level provides several key advantages over previous chip-wide applications. We have proposed local voltage dithering (LVD) to improve upon chip-wide voltage dithering. This section discusses the advantages of LVD and describes a test-chip that demonstrates these improvements.

Previous chip-wide implementations using voltage dithering report that the transition between two different supply voltages takes hundreds of microseconds [68][195]. This prevents the system from achieving any energy savings for faster changes in the workload. Dividing up the power supply grid into local regions reduces the capacitance that must be switched when the voltage supplied to a local block needs to change. This allows for faster changes in

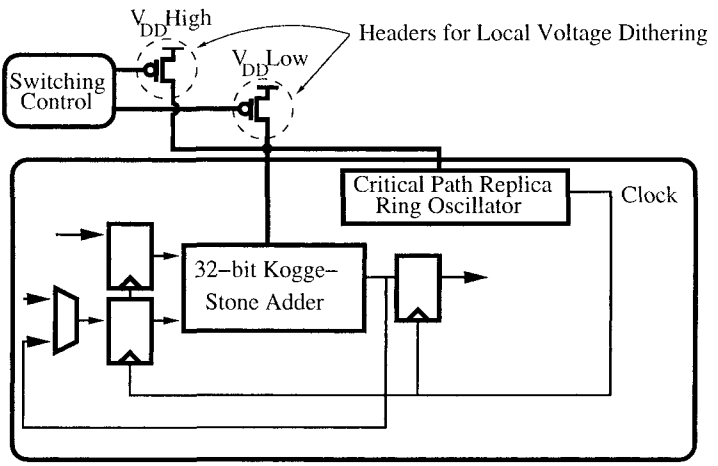

Fig. 9.9. Block diagram of voltage dithered adder and critical path replica using two local header switches for local voltage dithering (LVD). (© 2006 IEEE)

supply voltage with lower transitional energy and permits energy savings for changes in workload on the same timescale.

Chip-wide voltage dithering also restricts the extent to which varying workload may be leveraged because it must account for the highest workload from all of the blocks across the entire chip. For example, suppose a simple chip contains two large blocks. If one block has a workload of 0.9 and the other block has a workload of 0.2, then chip-wide voltage dithering must ensure that the block with the higher workload completes its work. Since both blocks share the dithered voltage supply, they both are forced to operate at the average rate of 0.9. Even if the less active block shuts down (e.g. clock gates) after completing its processing, it still uses more energy than if it could voltage dither based on its own workload. The energy savings that are lost by using chip-wide voltage dithering only increase with more blocks and wider differences between the maximum and minimum workloads. In contrast, LVD lets each block operate according to its own workload.

Our implementation of LVD uses embedded power switches (pMOS header devices) to toggle among a small number of voltage levels at the local block level. One advantage of this implementation approach is that the local dithering switches can be turned off to provide fine-grained power gating essentially for free.

9.2.2 UDVS Test Chip

We have implemented a test chip in 90nm bulk CMOS to demonstrate LVD and UDVS. This section describes the test chip architecture and provides measured results.

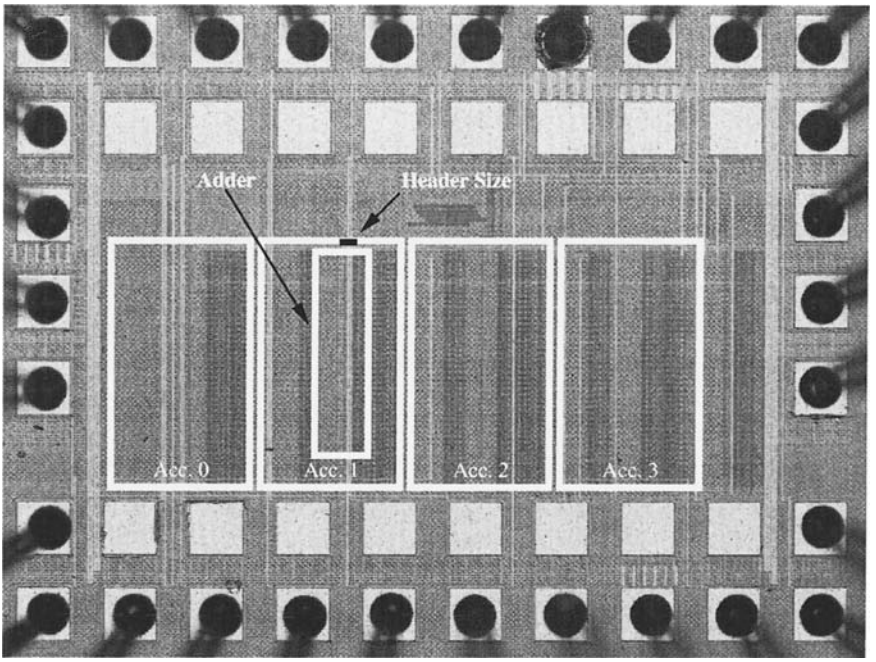

Fig. 9.10. Annotated die photograph showing accumulators with 0, 1, 2 and 3 headers. The size of one header is highlighted for reference. (© 2006 IEEE)

UDVS Test Chip Architecture

Figure 9.9 shows the primary block used for testing LVD on the test chip. The circuit of interest is a 32-bit Kogge-Stone adder that can be configured as an accumulator for testing. In this figure, two pMOS header switches select between a high supply voltage (V_{DDH}) and a low supply voltage (V_{DDL}) for the adder block. Other adders on the chip have different numbers of header devices. A critical path replica ring oscillator shares the same dithered voltage supply as the adder and sets the frequency of the clock based on the selected voltage. The die photo in Figure 9.10 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.

Placing a pMOS header switch in series with the power supply increases the delay of the circuit because of the voltage drop across the on resistance of the header. This effect is well-known and thoroughly analyzed in the context of power gating approaches such as multi-threshold CMOS (MTCMOS). Numerous methods for sizing such header devices are available, and most of them are designed to ensure that the circuit never exceeds some delay penalty relative to the circuit without any headers. The header switches on the test chip are sized to keep the delay penalty less than 10%.

Figure 9.11 shows the architecture, and Figure 9.12 shows the circuit schematics for the adder block on the test chip. The inverse of the propagate and generate signals are calculated in the first stage, and these results are applied to the adder tree. Each reconverging point in the tree has a "dot operator" circuit that calculates the propagate and generate values for that stage. Each stage in our implementation is inverting, so the two flavors of dot operator are shown in the critical path schematic.

Measurements

Since UDVS scales the supply voltage from the full V_{DD} down to the optimum V_{DD} for minimum energy, a UDVS system must consist of circuits that can function in the sub-threshold region. This test chip uses static CMOS circuits to ensure robust sub-threshold operation. The adder blocks on the test chip operate to below 200mV. Figure 9.13 shows an oscilloscope plot of the adder on the 90nm test chip operating in sub-threshold at 300mV, just below the minimum energy voltage.

The minimum energy per operation point measured for the adder appears in Figure 9.14 at $V_{DD} = 330\text{mV}$ ($f = 50\text{kHz}$) and 0.1pJ per addition for 25°C. Figure 9.14 also shows the measured effect of temperature on the total energy per cycle and leakage energy per cycle. An increase in temperature lowers the mobility of MOSFETs and decreases the threshold voltage according to:

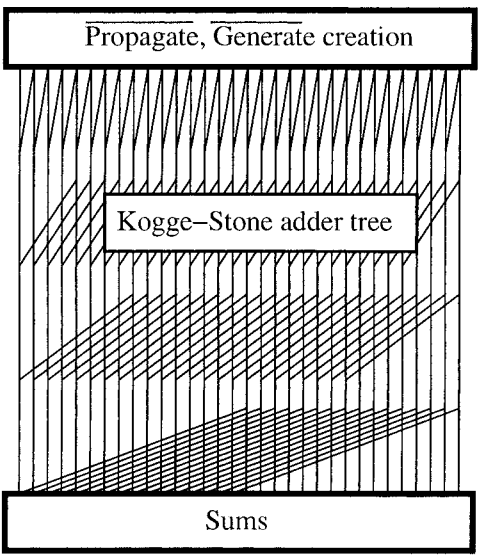

Fig. 9.11. Schematic of adder circuits. Kogge-Stone-based tree with inverting stages of dot operators (at each reconvergence of the tree). (© 2006 IEEE)

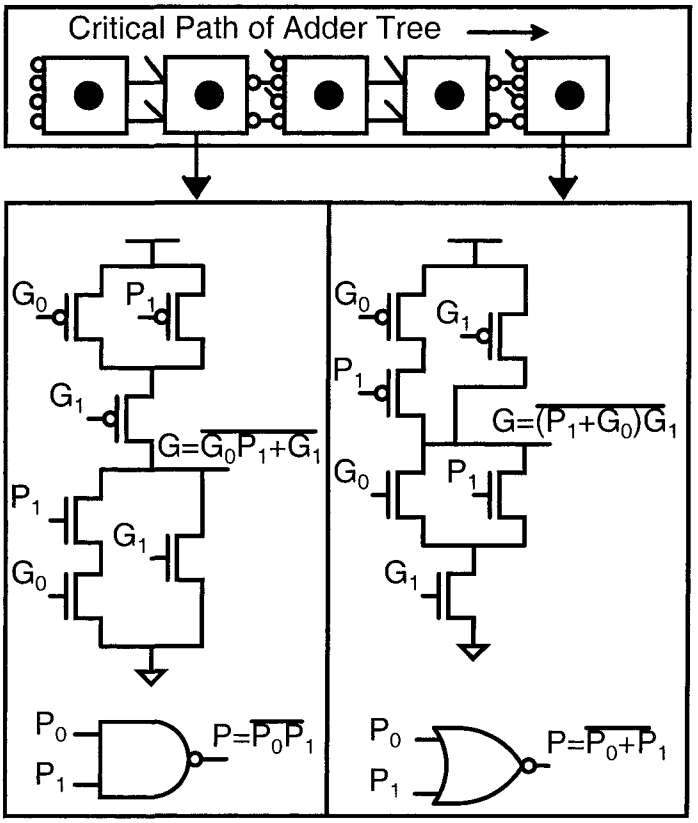

Fig. 9.12. Circuits for Kogge-Stone adder. Inverting stages of dot operators are in series along the critical path. Circuits do not require large stacks of transistors, which degrade sub-threshold operation. (© 2006 IEEE)

$\mu(T) = \mu(T_0)(\frac{T}{T_0})^{-M}$ and $V_T(T) = V_T(T_0) - KT$ [104]. For above-threshold operation, the decreased mobility dominates, and circuits slow down as they heat up. The leakage energy increases quickly with temperature for $V_{DD} > V_T$ because of the exponential dependence on the lower threshold voltage. In the sub-threshold region, however, the increased current also decreases the cycle delay, which causes the higher leakage currents to integrate over a shorter cycle time. As a result, the leakage energy does not change enough with temperature to greatly impact the optimum supply voltage. Figure 9.14 shows that the measured effect of temperature on the minimum energy point is small, validating the model in [99] and the analysis in Section 4.3.2. Figure 9.15 shows the measured frequency of one of the critical path ring oscillators on the test chip versus V_{DD} and temperature, confirming the increase of performance at higher temperatures in the sub-threshold region.

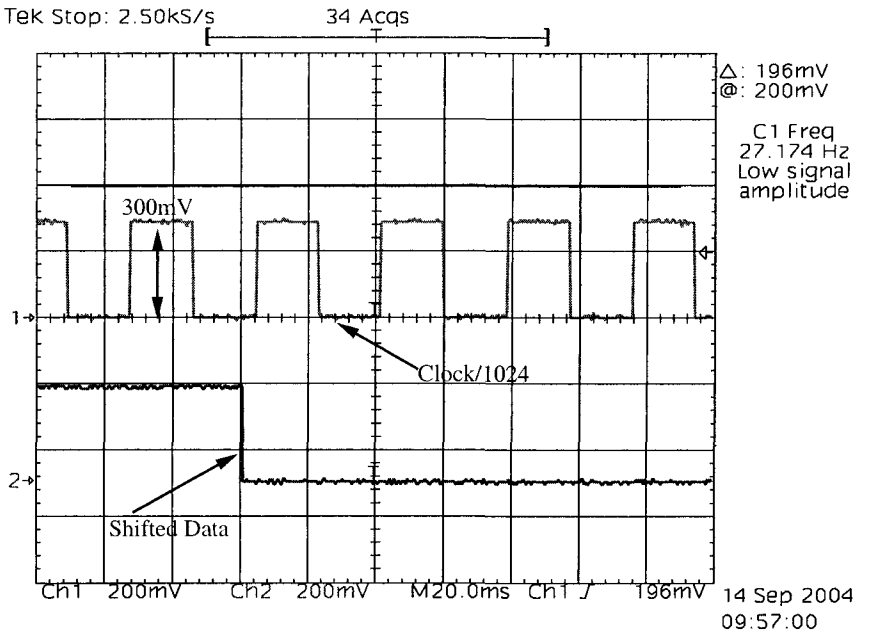

Fig. 9.13. Oscilloscope plot showing the clock and data from the 90nm test chip operating at 300mV, just below the minimum energy point. The adders functioned to 200mV. (© 2006 IEEE)

Figure 9.16 illustrates the savings that LVD provides for the adder block on the test chip when the rate varies. The dotted line shows operation at the highest rate followed by ideal shutdown. The solid line shows the measured energy versus rate for DVS assuming continuous voltage and frequency scaling. Selecting two rates from the curve, 1 and 0.5 in the figure, and operating for the correct fraction of time at each rate results in the dashed line that connects the quantized points, as described previously. 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.

While previously reported chip-wide approaches to voltage dithering have largely ignored the overhead energy of their schemes, we have investigated and measured the time and energy overhead of the LVD switching approach. Figure 9.17 shows the test circuit used to measure the delay overhead of LVD. While the adder runs a long accumulation, V_{DD} dithers to and from the higher rate. The oscilloscope plot in Figure 9.18 shows the divided ring oscillator output and the signal that selects the supply voltage (dither) for a dithering cycle between full and 0.5 rate. When the headers toggle V_{DD}, a counter gates the clock for a specified number of cycles to ensure settling at the new voltage. Checking the accumulated value verifies correct operation

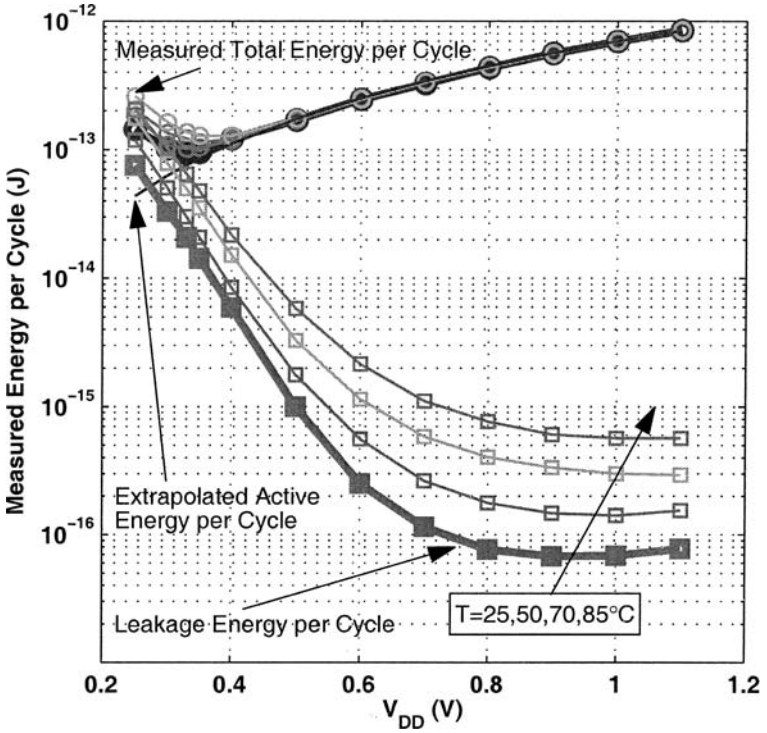

Fig. 9.14. Measured energy per cycle in the sub-threshold region for input activity of one. Minimum energy point occurs at 330mV (50kHz) and 0.1pJ per operation at 25°C. The optimum supply voltage is relatively insensitive to temperature variation. (© 2006 IEEE)

for every cycle. Measurements showed that the correct value was accumulated even with only 1/2 cycle (minimum possible using the test circuit) of clock gating for V_{DDL} above 0.6V, which corresponds to a rate of 0.04. Thus, even conservative settling times for this LVD implementation are on the order of a few cycles. This measurement confirms that LVD can adjust to fast changes in the workload of the local blocks.

In addition to timing overhead, there is energy overhead associated with the LVD approach. The buffer network and control circuits that drive the header switches consume energy every time they toggle the header switches to select a new supply voltage for the adder circuit. We can relate this overhead switching energy to the active switching energy of the adder block to determine its impact on overall energy savings from the LVD approach. To this end, we normalize the effective overhead switched capacitance of the control and buffer circuits, $C_{OVERHEAD}$, to the effective switched capacitance of the adder. The

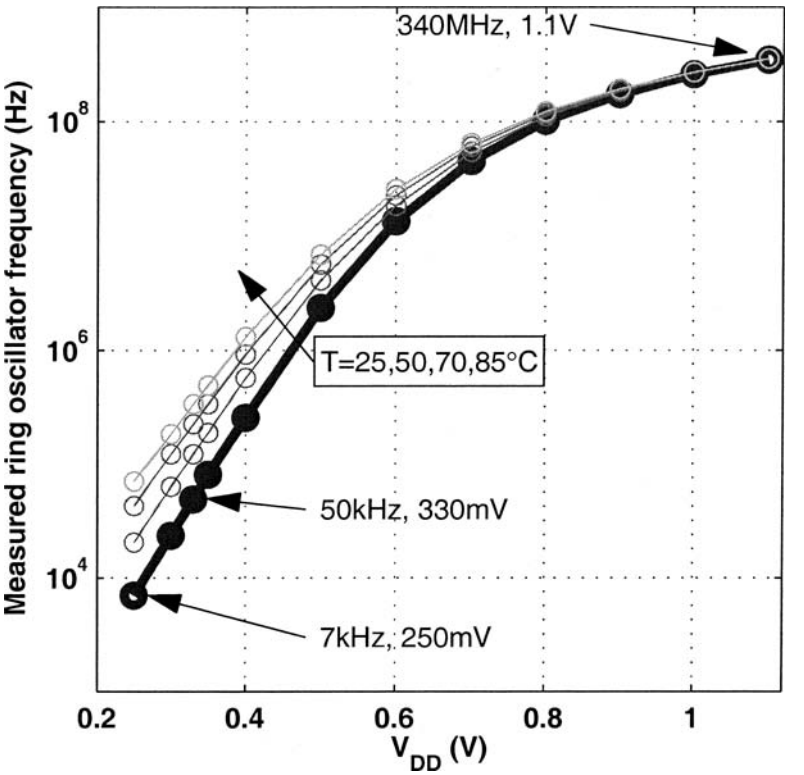

Fig. 9.15. Measured ring oscillator frequency versus V_{DD} and temperature. (© 2006 IEEE)

expression in Equation (9.2) shows the relation that must hold true in order for LVD to provide energy savings for a given transition.

$$NV_{DDH}^2 \geq NV_{DDL}^2 + C_{OVERHEAD}V_{DDH}^2 \tag{9.2}$$

Solving (9.2) for N gives the number of cycles that must occur at V_{DDL} in order to make switching to V_{DDL} worthwhile for saving energy, as shown in (9.3).

$$N \geq \frac{C_{OVERHEAD}V_{DDH}^2}{V_{DDH}^2 - V_{DDL}^2} \tag{9.3}$$

Measurements of the test chip show that $C_{OVERHEAD} = 3.7$ for the adder, so N is only 12 for the adder block with V_{DDH}=1.1V and V_{DDL}=0.9 (rate=0.5). Since the control circuits on the test chip are relatively simple, the overhead energy for more complicated control schemes, such as those that calculate the effective workload, has the effect of increasing N.

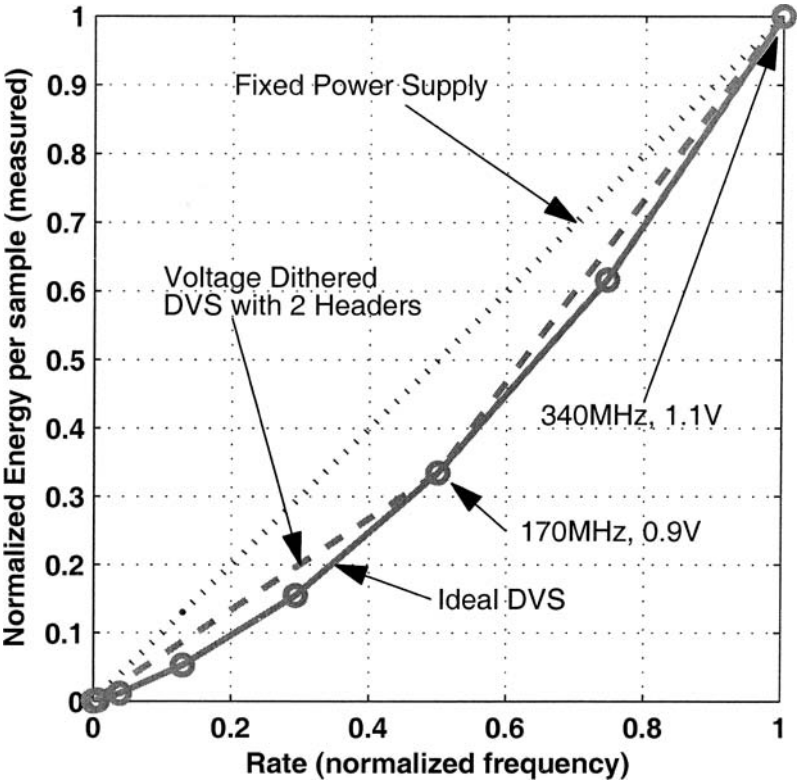

Fig. 9.16. Characterized local voltage dithering using measured results for 32-bit Kogge-Stone adder. (© 2006 IEEE)

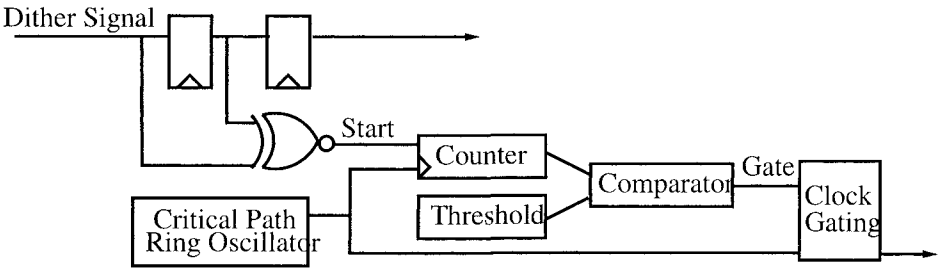

Fig. 9.17. Circuit for measuring timing overhead of LVD that gates the clock at a V_{DD} transition for a given number of cycles. The duration of this clock gating is decreased until the circuit fails. (© 2006 IEEE)

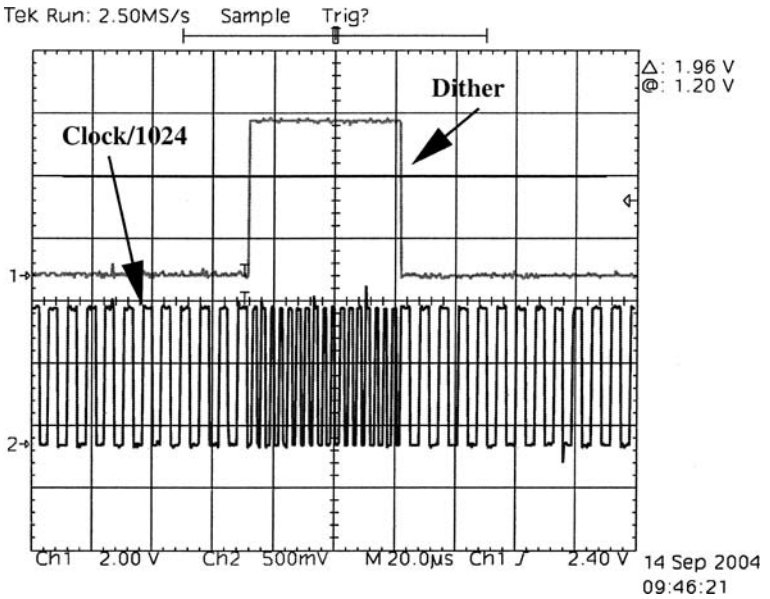

Fig. 9.18. Oscilloscope plot showing the system clock while dithering between rate 0.5 (170MHz) and rate 1 (340MHz). Measurements show correct accumulation at both transitions even no clock gating (see Figure 9.17). (© 2006 IEEE)

9.2.3 UDVS System Considerations

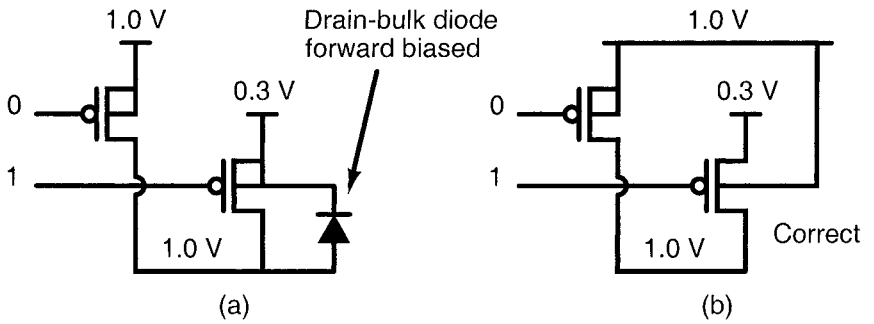

Fig. 9.19. For UDVS, the bulk connections of the pMOS header switches have to connect to the highest supply voltage. (© 2006 IEEE)

The discussion to this point has assumed that the varying rate remains above roughly a few percent. As previously mentioned, sub-threshold operation has proven to minimize energy for low performance applications. While scaling to sub-threshold is rarely advantageous for full processors [101], local blocks or special applications that require brief periods of high performance spend significant amounts of time operating at effective rates that are orders of magnitude below one. Examples of these applications include micro-sensor nodes, medical devices, wake-up circuitry for processors, and local blocks on active processors. When performance is non-critical, energy is minimized by operating at the minimum energy point that occurs because of increased leakage energy at low frequency and then shutting down if there is more timing slack.

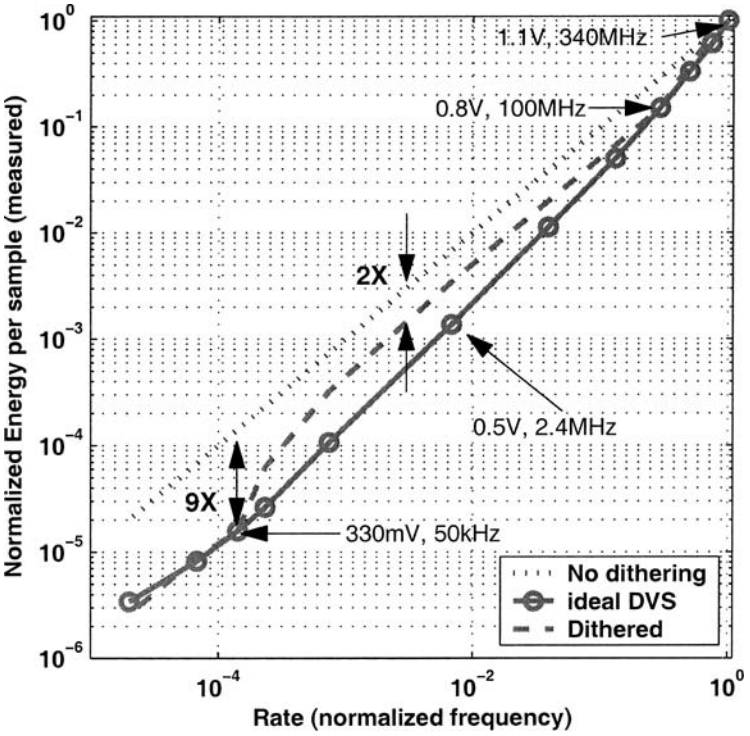

Fig. 9.20. UDVS) using two headers with one variable DC-DC converter or using three headers (c.f. Figure 9.23). (© 2006 IEEE)

Since LVD works well for high speed operation and operating at the minimum energy point is optimal for low performance situations, we propose ultra-dynamic voltage scaling (UDVS) using local power switches [196]. This

approach uses local headers to perform LVD when high performance is necessary and selects a low voltage for sub-threshold operation at the minimum energy point whenever performance is not critical. As with LVD, all of the headers for a given block can turn off when the block is idle to conserve standby power using power gating. Since the power switches in the UDVS approach connect to different voltages that can differ substantially, they must be configured carefully to prevent forward biasing the junction diodes. Figure 9.19 shows that connecting the bulk terminal of the header transistors to the source can forward bias the drain-bulk diode when the V_{DDL} switch is off. One solution to this problem is to tie the bulk of all of the header switches to the highest supply voltage as in Figure 9.19(b).

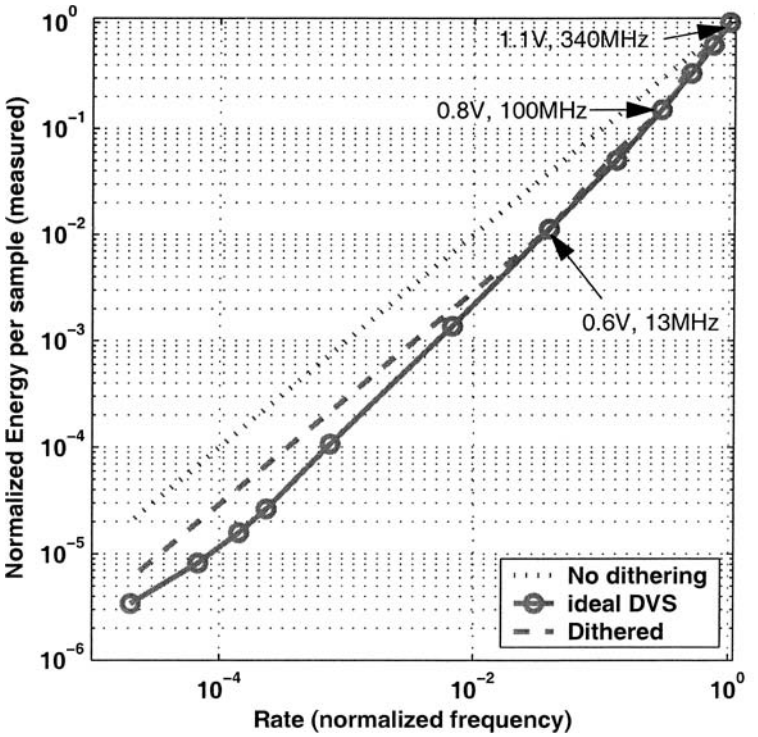

Fig. 9.21. Different choice of dithered voltages for closer fit over the higher range of V_{DD}. (© 2006 IEEE)

Figure 9.20 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

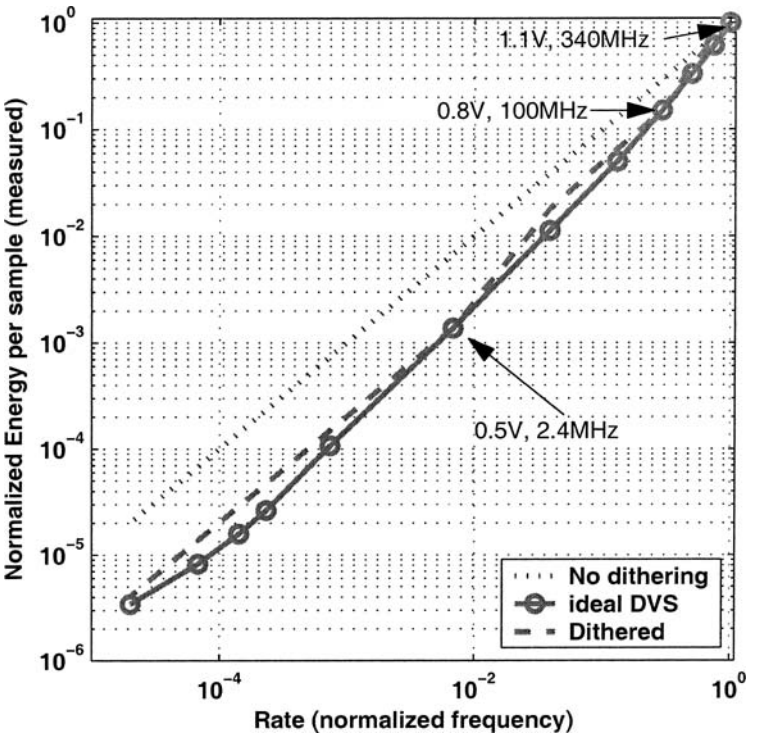

Fig. 9.22. Different choice of dithered voltages for closer fit over the entire range of V_{DD}. (© 2006 IEEE)

minimum energy point at 50kHz and saves 9X 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. It was shown in [202] that significant extra savings are available if the selected dithered rates match to the prominent average rates in the data. This brings the dithered curve closer to the optimum DVS curve for the common cases. Figure 9.21 and Figure 9.22 show two additional examples in which the supply voltages are chosen for different scenarios. For a system whose rate requirements vary evenly over the full range, the voltage choices in Figure 9.21 provide a better match to the ideal energy profile, but the minimum energy per operation is not achievable. If performance constraints prevent a system from ever operating at the minimum energy point, the supply voltage can be adjusted to higher voltages to achieve near optimal energy operation over the range of higher rates (Figure 9.22).

Figure 9.23 shows two options for implementing the power supplies and headers in a UDVS system. The straightforward option is to distribute three

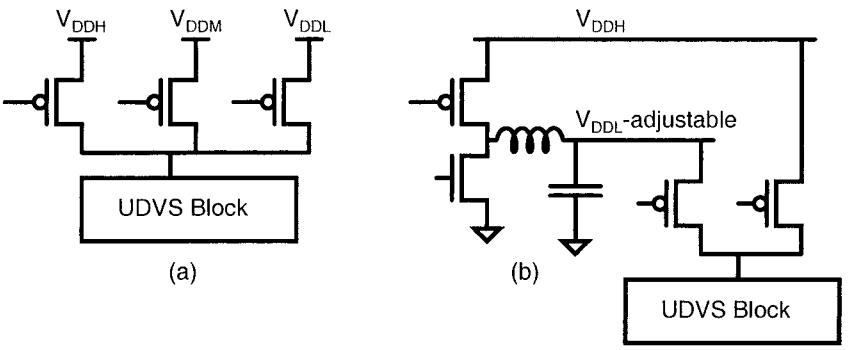

Fig. 9.23. Options for UDVS headers at the system level. (© 2006 IEEE)

supply voltages around the chip and to use three header switches at each block. The voltages V_{DDH}, V_{DDM}, and V_{DDL} can be selected based on the system workload statistics as we described above. The only advantage to using more than three power supplies is to pin the UDVS energy profile to the ideal variable supply profile in more places. The right-hand diagram in Figure 9.23 offers a second option for implementing UDVS. When transitions between high and low performance mode are infrequent and the amount of time in between transitions is long, two header switches may be paired with one adjustable DC-DC converter for the same functionality. For example, during high speed operation, the headers dither between 1.1V and 0.8V. When the rare transition to low speed occurs, the DC-DC converter switches V_{DDL} to 0.35V so that all of the blocks can operate near their minimum energy points.

For applications where some blocks operate in sub-threshold while others are at higher voltages, special interfacing circuits are required at the low voltage region to high voltage region interface. The type of level converters to be used will depend on how the block interfaces to surrounding blocks. Ample previous work on level converter circuits offers many choices for implementing the required interfaces. In a full UDVS system with multiple blocks, each block has its own header devices so that it can voltage dither based on its individual workload. Communication among blocks occurs along a bus, which might be asynchronous to account for different operating frequencies, and level converters interface to the bus as needed.