Standby Power Reduction Techniques for Ultra-Low Power Processors

Yoonmyung Lee, Mingoo Seok, Scott Hanson, David Blaauw, Dennis Sylvester

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI

***Abstract* — Standby power can dominate the power budgets of battery-operated ultra-low power processors, and reducing standby power is the key challenge for further power reduction. State-of-the-art ultra low voltage sensors consume hundreds of nW in wake mode and 100 pW or less in standby mode. Therefore, applying known circuit techniques for further standby power reduction is very challenging. In this paper, we extend known standby power reduction techniques for use in ultra-low power processors. In particular, we propose structures that enable the use of super cut-off voltages throughout the design with minimal power overhead. Different strategies for power gated logic blocks and memory cells are investigated.**

I. Introduction

The size of ultra-low power sensor systems is a critical concern, especially for medical applications requiring implantation. Cost, which is related to system volume, is also an important limitation in sensor systems. Since the size of the power source is restricted in such applications, ultra-low power consumption on the order of nanowatts (nW) and picowatts (pW) is required for these sensor processors.

One of the most promising approaches to achieving ultra-low power consumption is supply voltage scaling into the subthreshold regime [1] to minimize wake mode energy. However, many sensor systems spend much more time in standby mode than wake mode. Previous approaches have neglected the power consumed in this standby mode despite the fact that standby power can dominate the system budget [2]. Recent work [3] has shown that a better balance between wake mode power and standby mode power can be achieved by designing the system with standby power as a primary constraint. Careful technology selection for balancing active and standby power, stacking high-V_{th} transistors in memory cells for less subthreshold leakage, power gating for less standby power and other architectural/circuit techniques were shown to reduce standby power to tens of pW, giving ~1 year lifetime with a $1mm^3$ system size including battery.

However, even with the sleep strategies presented in [3], standby power is still a dominant (>75%) source of total power consumption. Standby power consists of two components. The first component is the power consumed by circuits that are turned off (power gated) during standby mode. The second component is the power consumed by circuits that must retain state and remain turned on (e.g., memory). The ratio between these two types of standby power can vary depending on the complexity of logic and amount of memory required, though the second type dominated the standby power in [3]. Therefore, developing different techniques for reducing each type of standby power is the key challenge for extending the lifetime of ultra-low power applications to the multi-year range.

However, reducing the standby power for circuits that only consume tens of pW is very challenging for several reasons: 1) the power overhead for using any leakage reduction techniques must be a few pW in order to be beneficial, 2) since these systems are typically battery operated, only a single supply voltage is available, 3) any locally generated voltages for power reduction that are greater than power supply voltage (V_{DD}) or less than the ground voltage, should be controlled without level converters or other switches that introduce new leakage paths.

In this paper, we develop standby power reduction techniques that can be applied to ultra low power processors. First, we explore the use of super cut-off MTCMOS for reducing standby power in power gated blocks. Our key contribution is the development of an ultra-efficient charge pump and cut-off circuit designed for low frequency operation (1~10Hz). Next, we investigate leakage paths in memory and propose a leakage reduction strategy that uses a super cut-off voltage to reduce bitline leakage. To support charge pump operation, a sub-pW clock generator with a unique current starving scheme is also introduced.

A test chip is fabricated in a 0.18μm CMOS technology to demonstrate the proposed leakage reduction techniques. This older process was strategically selected due to the availability of devices with high V_{th} and negligible gate leakage. The target V_{DD} of the chip is 0.5V, which is typical for ultra-low power processors. Measured results show that a 4.6Hz charge pump clock is generated with a 0.64pW power overhead and a standby power reduction of 2.3-19.3X is achieved for power gated logic blocks. For memory, standby power is reduced by 29%.

II. Standby Power Reduction for Logic Blocks

Large logic blocks in ultra low power processors, such as the CPU, are often power gated to minimize standby power. For such circuits, using super cut-off is a straightforward and effective method for further reducing standby power [4]. In the super cut-off technique a negative voltage is applied to the power gating NMOS footer or a voltage greater than V_{DD} is applied to the PMOS header. However, the power cost of generating this super cut-off voltage has been shown to be large (50nW in [4]) relative to the sub-nW standby power budget targeted in this work. Consequently, the application of this technique becomes challenging in ultra low power processors. To apply the super cut-off strategy to a block with tens of pW standby power, the generation of the super cut-off voltage must have a power overhead on the order of several pW, or ~1000X lower than the results presented in [4].

As shown in Figure 1, the proposed system includes a charge pump that generates the super cut-off voltage and an output driver to

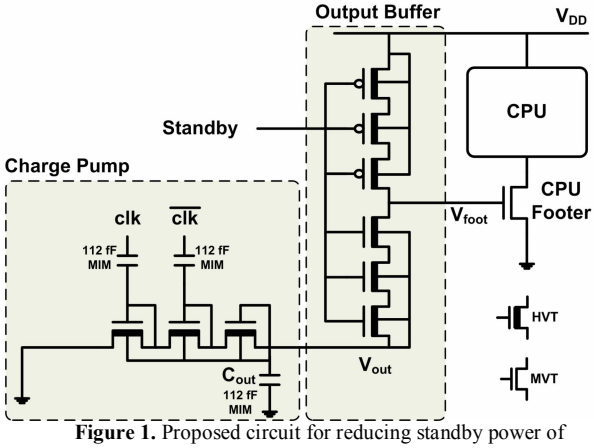


Figure 1. Proposed circuit for reducing standby power of power gated blocks

switch the gate voltage on the footer (V_{foot}) between the super cut-off voltage (V_{out}) in standby mode and V_{DD} in wake mode. The charge pump consists of three high-V_{th} NMOS transistors and three metal-insulator-metal (MIM) capacitors. Two clock signals with opposite phases (to be described further in Section IV) are applied to the pumping capacitors. To ensure maximum power efficiency, the clock must oscillate at the lowest possible frequency, so all leakage paths at V_{out} must be eliminated. Leakage is minimized along the pumping stack by using high-V_{th} devices and by reverse biasing the bodies of the pumping transistors using V_{out}.

To further improve power efficiency, a triple stacked inverter is used for connecting V_{out} to the footer. The PMOS stack minimizes subthreshold leakage during standby mode thereby lessening the pumping overhead and the required pumping frequency, while the NMOS stack plays a critical role when switching from standby mode to wake mode. The long NMOS stack cuts the connection between V_{out} and the gate of the footer to eliminate contention between the PMOS stack and the charge pump. It is also crucial to bias the bodies of the entire NMOS stack with V_{out} to ensure that the NMOS stack is not forward biased during wake mode. The negative voltage developed at V_{out} is preserved during wake mode, which is typically very short (on the order of milliseconds) [2], thus minimizing the time and power overhead of switching back to standby mode.

The carefully designed configuration described in this section allows the charge pump to be operated with low clock frequency (<10 Hz) and sub-pW power while guaranteeing sufficiently low (<-150mV) super cut-off voltage at the output at room temperature (25°C). Measurement results will be discussed in Section V.

III. Standby Power Reduction in Memory

Various SRAM structures, such as the modified-6T [5], 8T [6] and 10T [7] topologies, have been explored for low voltage applications. Despite obvious differences, each of these structures has similar components: a cross-coupled inverter pair, bit-lines, word-lines, access transistors and read buffers. Consequently, we can identify several sources of leakage that are common across all structures. To explore standby power reduction for memory, we study the low-leakage memory cell proposed in [3]. Given the general similarities between various SRAM structures, many of the conclusions in this work may be extended to other cells. As depicted in Figure 2, the memory cell under investigation uses cross-coupled inverters with stacked high-V_{th} transistors to minimize the subthreshold leakage. A separate read buffer with medium-V_{th} transistors is used to boost the read performance and improve cell stability at low voltage.

A. *Leakage reduction for power gated blocks*

Figure 2 shows the most important leakage paths within and between memory cells. Path 1 is the leakage path for circuits that are power gated (i.e., turned off) during standby mode. Only the read buffer is shown in Figure 2, but this category of circuits also includes memory peripherals such as row/column decoders, bit-line drivers and other control logic. Since these circuits are all turned off by a footer, our analysis shows that Path 1 contributes only ~2% of the total standby power. A separate power gating transistor is used to ensure that the current drawn from other power intensive modules, such as the CPU, does not induce read/write errors during wake mode. However, the super cut-off voltage that is generated by the charge pump introduced in Section II can be shared with virtually no power overhead.

B. *Bit-line leakage reduction*

Path 2 in Figure 2 shows the bit-line leakage path in the array structure of the memory. During standby mode, the bit-lines (BL and $\overline{BL}$ in Figure 2) float to some intermediate voltage, V_{BL}, between 0

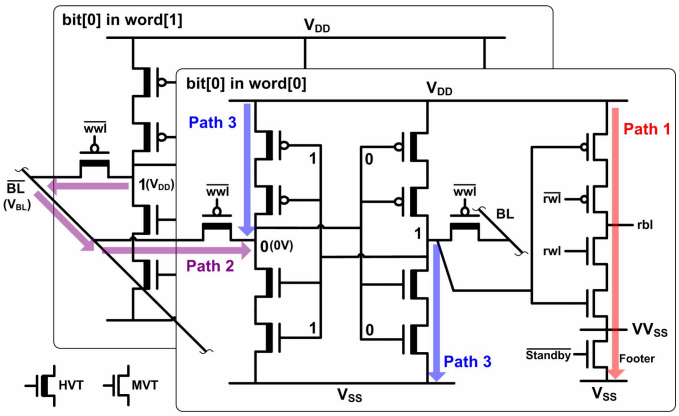


Figure 2. Leakage paths in low leakage memory cell

and V_{DD}. The value of V_{BL} depends on the number of bit cells storing 0's and 1's in the bit-line column. As a result, the transistors that connect the bit-lines and the memory cell (pass transistors) will have a drain-source voltage of V_{BL} or $V_{DD} - V_{BL}$ when the cell stores 0 or 1 in the adjacent node, respectively. This drain-source voltage induces subthreshold leakage on the bit-line, which contributes ~50% of total standby leakage.

In order to reduce the bit-line leakage, a super cut-off voltage (> V_{DD}) can be applied to the gate of the pass transistors during standby mode. This can be achieved by using a charge pump to boost the power supply for the wordline driver connected to the pass transistor control. The basic concept of this strategy is similar to the strategy used with power gated logic blocks, but it raises the following new challenges: 1) a new power supply for the pass transistor control logic must be kept near V_{DD} or higher at all times since low voltage at the gate of pass transistors will turn on the transistors, resulting in data loss, 2) the new power supply should be able to supply enough current to meet the demands of the pass transistor control logic during a memory write operation, and 3) all these criteria should be met with a power budget on the order of pW.

The proposed circuit that meets these criteria is presented in Figure 3. An ultra-low power charge pump similar to the one presented in the previous section is used for boosting the power supply. PMOS transistors are used to generate a positive super cut-off voltage V_{out} (>V_{DD}). The output of this charge pump is tied to the power rail of the wordline drivers. Charge is continuously pumped into the output capacitor (C_{out}) to develop V_{out}. The wordline drivers are structured to always provide full V_{out} in standby mode while also enabling wordline control during the wake mode. However, there can be no direct connection to the power supply at the output node during wake mode because a direct connection to V_{DD} would prevent V_{out} from rising higher than V_{DD} in standby mode. As a result, write operations that lead to a transition at the output of the wordline drivers will consume the charge stored in C_{out}, thereby lowering V_{out}. Therefore, consecutive write operations that occur between pumping cycles (due to the low pumping frequency) may bring V_{out} below V_{DD}. As the voltage reduces, the pass transistors of memory cells will be turned on, resulting in data loss.

To prevent this data loss, a "holder" transistor is introduced. The holder transistor indirectly connects V_{DD} with the output of the charge pump and is turned on during wake mode. When V_{out} drops below V_{DD}, the holder transistor is forward biased and can effectively "hold" V_{out} near V_{DD}. A wide low-V_{th} transistor would be preferable for the holder transistor, but in standby mode, the holder transistor acts as a direct leakage path from the output of the charge pump to V_{DD}, thereby reducing pump efficiency. Thus, a moderately sized (W:0.55µm L:0.35µm) high-V_{th} transistor is chosen to alleviate this

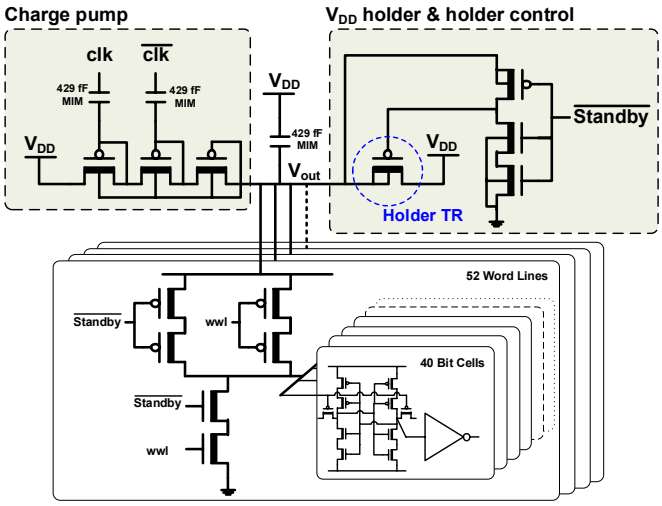


Figure 3. Proposed circuit for bit-line leakage reduction

side effect. Worst case simulations show that this configuration maintains V_{out}>489mV at V_{DD}=0.5V.

C. *Intra-cell leakage*

Finally, Path 3 in Figure 2 shows the intra-cell subthreshold leakage path. In each cell, the primary leakage paths include a single NMOS stack and a single PMOS stack. For example, with a bit value of 1 stored in the front memory cell in Figure 2, the top left PMOS stack and bottom right NMOS stack will leak. Our analysis shows that this leakage amounts to ~48% of total standby power.

In order to suppress intra-cell subthreshold leakage, a reverse body bias can be applied to all transistors or high V_{th} transistors can be used. However, according to our analysis, the standby power of our target memory module was 60.5pW and the overhead of generating enough well bias current to compensate for junction leakage was greater than the projected leakage improvement. Therefore, our memory structure uses high-V_{th} transistors as in [3].

IV. Ultra Low Power Clock Generation

The clock generator is one of the most important elements in our proposed ultra-low leakage system. Without proper design, the clock generator can easily exceed the pW budget allotted. Figure 4 illustrates the proposed clock generator with a unique current starved inverter. In this inverter, current starved transistors are placed next to the output node whereas conventional design places them next to the power and ground rails.

To achieve minimum power, the clock generator is designed for operation at very low frequencies (1~10Hz). Each inverter in the clock generator uses stacked high-V_{th} transistors adjacent to the power and ground rails and current-starved medium-V_{th} transistors in the off-state adjacent to the output node. In this configuration, the on-current of the inverter is determined by the subthreshold leakage of the starved medium-V_{th} transistors, which makes the current consumption very small. When the input is low, the NMOS stack is turned off and a small voltage is developed at the source of the starved NMOS due to stack effect. Thus, a reverse body bias is generated for the starved NMOS, making the off-current smaller and thereby improving the power efficiency over the case where the starved transistors are adjacent to the power and ground. The same effect can be observed in the PMOS stack. Our analysis shows that the 10%-90% rise/fall time can be reduced by 19.6% with our proposed design, making the clock generator more stable and robust.

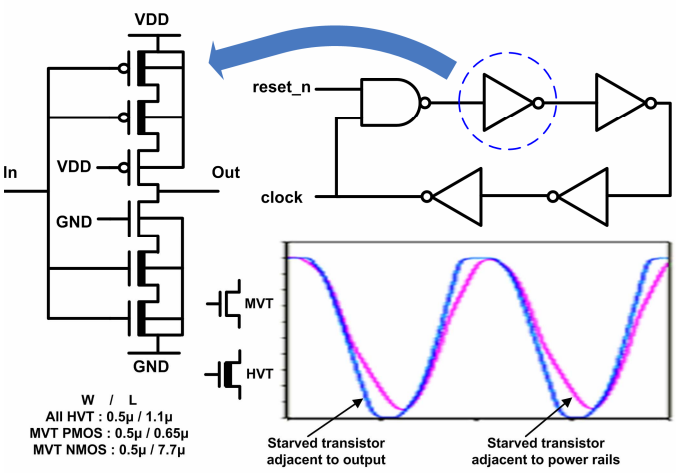


Figure 4. pW clock generator with current starved transistors and output waveform comparison between different starved transistor placement schemes

V. Measurement Results

A. *Measurement of standby power for power gated blocks*

A large CPU block with 23,472 transistors has been tested using 4 different medium-V_{th} footer sizes at room temperature (25°C). Figure 5 shows the generated super cut-off voltage and charge pump power consumption as functions of the charge pump clock frequency. The charge pump clock was supplied externally in this specific experiment to give maximum tunability. Strong super cut-off voltages (<-150mV) are generated with low pumping frequency (<10Hz) and sub-pW power consumption. The leakage reduction achieved using super cut-off MTCMOS is shown in Figure 6. With a footer width of 17.16µm, the CPU block consumes 15.4pW in standby mode without super cut-off MTCMOS. For low pumping frequencies (<10Hz), increasing the pumping frequency reduces total standby power since the super cut-off voltage reduces. However, as frequency exceeds 10Hz, the charge pump overhead becomes dominant and increases total power consumption. Total standby power reaches a minimum of 0.8pW at 10Hz, a 19.3X reduction over normal operation.

Figure 7 shows the standby power reduction for different footer sizes. Despite different footer sizes, the standby power converges to ~1pW for all cases at an optimal pumping frequency of 10Hz. Therefore, the power gain is largest (19.3X) with the widest footer and smallest (2.3X) with the narrowest, which suggests that this power reduction technique may also enable active power reduction by allowing more freedom when choosing the size of the power gating transistor.

The size of the power gating transistor is constrained by the standby mode power budget and wake mode current demand. In wake mode, a wider power gating transistor is preferred to minimize the voltage drop across the power gating transistor. However, since the standby power of a circuit block is determined by the size of the power gating transistor, narrow width is preferred for minimum standby power. Energy consumption in standby mode dominates wake mode energy consumption for ultra-low power processors, so a power gating transistor with very narrow width is typically used (a footer width of only 0.66µm was used in [3]). The voltage drop across such a narrow power gating transistor effectively reduces V_{DD} for the logic, making the circuit block slower, less robust and less energy efficient. In light of our measured results, a wider power gating transistor can be used with a minor standby power penalty and significant wake mode energy reduction (estimated at 23% by eliminating 116mV out of 500mV).

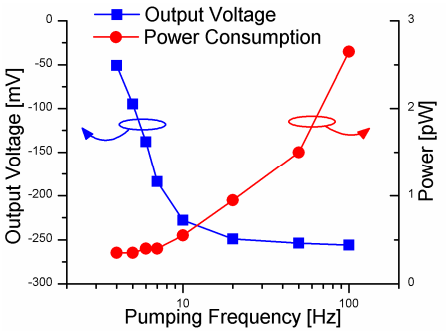


Figure 5. Generated super cut-off voltage and charge pump power for CPU leakage reduction

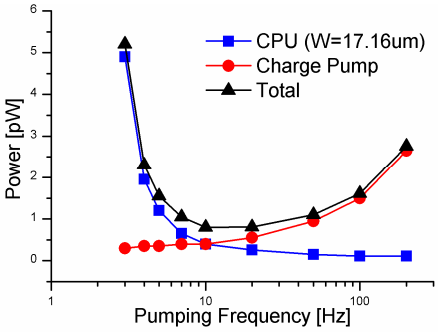


Figure 6. CPU and charge pump power in standby mode

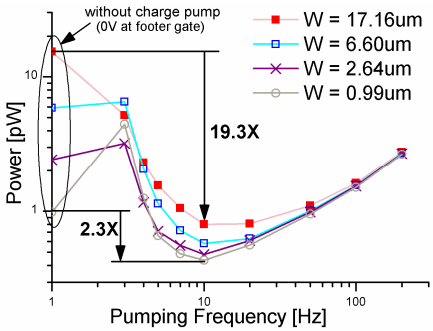


Figure 7. Total standby power of CPU and charge pump with various footer size

B. *Measurement of standby power for memory*

A memory with 2,720 bit cells has been tested at room temperature (25°C). Figure 8 shows the generated super cut-off voltage and charge pump power consumption as functions of charge pump clock frequency. The power overhead for the charge pump is significantly higher than for the previous section due to the larger number of leakage paths such as the pass-transistor controllers and the holder transistor. At the power optimal pumping frequency of 20 Hz, the charge pump overhead is below 5% of original memory standby power. Total standby power is shown in Figure 9. At a pumping frequency of 20 Hz, standby power is reduced by 29.1% compared to normal operation. Note that power actually increases at low frequencies since the output of the charge pump can fall below V_{DD} (0.5V) in this region and cause increased leakage across pass transistors.

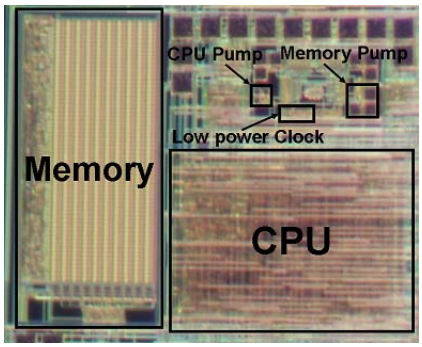
C. *Low power clock generator*

Testing of the low power clock proposed in Section IV shows an average oscillating frequency of 4.6 Hz with a power consumption of only 0.64pW. Simple calculations suggest that, at the optimal frequency for the two previously described charge pumps (10Hz, 20Hz), clock power can be maintained below 3pW. Since the power characteristic in Figure 6 is flat near the minimum, applying the memory-optimal clock frequency of 20Hz to the CPU charge pump results in a negligible power penalty of only 1.3%. This result suggests that a single clock generator can be shared between the memory and CPU.

Measurements at temperatures ranging from 0-80°C reveal that the low power clock tracks the power optimal frequency well. Figure 10 shows the power optimal charge pump clock frequency for CPU and generated frequency, both normalized at 40°C. Over this temperature range, discrepancies between the optimal frequency and the generated frequency result in a maximum power penalty of only 14% compared to the optimal operation point.

VI. CONCLUSION

Super cut-off circuit techniques for reducing the standby power of ultra-low power processors have been presented along with a supporting low power clock generator. A standby power reduction of 2.3-19.3X is achieved for power gated logic blocks, while standby power is reduced by 29.1% for memory using the proposed techniques.



Dimensions	
CPU Pump	960 μm^2
Memory Pump	1,365 μm^2
Low power Clock Gen.	658 μm^2

Figure 11. Die Micrograph and Dimensions

REFERENCES

[1] A. Wang, A. Chandrakasan, "A 180mV FFT processor using subthreshold circuit techniques," Int. Solid-State Circuits Conf., 2004, pp. 292-293

[2] M. Seok, *et al.*, "Analysis and Optimization of Sleep modes in Subthreshold Circut Design," ACM/Design Automation Conference, 2007

[3] M. Seok, *et al.*, "The Phoenix Processor: a 30pW platform for sensor applications," Symposium on VLSI Circuits, June 2008, in press

[4] H. Kawaguchi, *et al.*, "A super cut-off CMOS (SCCMOS) scheme for 0.5-V supply voltage with picoampere stand-by current," JSSC, Oct. 2000, pp.1498-1501

[5] B. Zhai, *et al.*, "A sub-200mV 6T SRAM in 130nm CMOS," Int. Solid-State Circuits Conf., 2007, pp. 332-333

[6] L. Chang, *et al.*, "Stable SRAM cell design for 32 nm node and beyond", Symposium on VLSI Technology, Jun. 2005, pp. 128-129

[7] B. Calhoun, A. Chandrakasan, "A 256kb Sub-threshold SRAM in 65nm CMOS," Int. Solid-State Circuits Conf., 2006, pp. 2592-2601

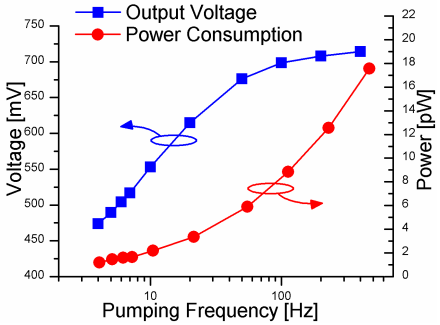


Figure 8. Generated super cut-off voltage and pump power for Memory leakage reduction

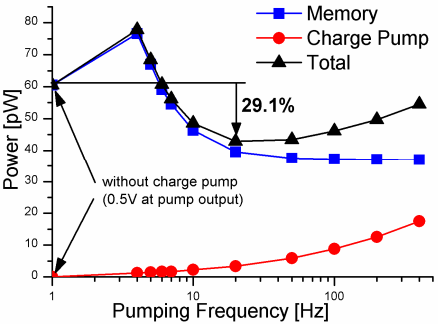


Figure 9. Memory and charge pump power in standby mode

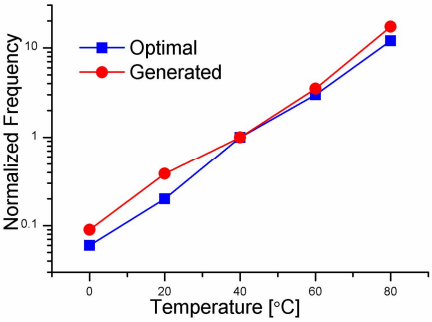


Figure 10. Optimal and generated clock frequency normalized at 40°C