Extending Energy-Saving Voltage Scaling in Ultra Low Voltage Integrated Circuit Designs

Mingoo Seok[1], Dongsuk Jeon[2], Chaitali Chakrabati[3], David Blaauw[2], Dennis Sylvester[2]
Columbia University[1], University of Michigan[2], Arizona State University[3]
mgseok@ee.columbia.edu

***Abstract* — in this paper, we propose several design approaches to extend useful voltage scaling (i.e. voltage scaling with net energy savings) beyond the conventional limit, which is imposed by the rapid increase of leakage energy overhead in ultra low voltage regimes. We are able to achieve such extra voltage scaling and thus energy savings without compromising performance and variability through minimizing the ratio of leakage to dynamic energy in a circuit. Novel design approaches in pipeline, clocking and architecture optimization are investigated; and applied during the design of a 16b 1024pt complex FFT core. The measurement results from the prototyped FFT core in a 65nm CMOS show the energy consumption of 15.8nF/FFF with the clock frequency of 30MHz and the throughput of 240Msamples/s at the supply voltage of 270mV, which exhibits 2.4× higher energy efficiency and >10× higher throughput than the previous low power FFT designs. Measurement of 60 dies shows modest frequency and energy σ/μ spreads of 7% and 2%, respectively.**

Keywords - ultra low voltage, ultra low power, super-pipelining, two-phase latch-based design, less-buffered clock networks, energy-optimal FFT architecture, FFT core, useful voltage scaling

I. INTRODUCTION

Voltage scaling techniques have been one of the promising methods to minimize energy consumption of CMOS integrated circuits. As the supply voltage scales, quadratic to exponential energy savings in switch, subthreshold leakage, and gate leakage energy can be achieved. Particularly, for the applications that have a relaxed performance requirement, we can reduce the supply voltage to near or below the threshold voltage (V_{th}), referred to as the ultra low voltage (ULV) regimes, to maximize energy efficiency.

However, such useful voltage scaling, i.e. a voltage scaling which reduces energy consumption, stops due to leakage energy overhead. Zhai [1] and Calhoun [2] showed that energy consumption starts to increase if we scale the supply voltage too low since the increasingly slow circuits accumulate more and more leakage energy, which eventually offsets the quadratic savings of dynamic energy. We define the supply voltage at which the total energy consumption starts to increase as energy optimal voltage or V_{opt}. The energy consumption at V_{opt} is referred to as E_{opt}. The V_{opt} often lies at 0.35-0.45V for circuits in modern sub-micron CMOS technologies.

The V_{opt} and E_{opt} pose the conventional limit of energy efficiency through ULV operations. To improve energy efficiency beyond this point, it is necessary to minimize leakage energy overhead and make design more dynamic-energy dominated. Generally, reducing leakage power or shortening clock period can reduce leakage energy overhead; however one must make sure that such efforts do not come with other overheads in switch energy consumption, performance, variability and robustness such that we can achieve net benefits.

In this paper, we propose several circuit and architecture level design approaches for reducing leakage energy and thus extending useful voltage scaling. First, we explore pipeline methodology in ULV design space and propose so-called "super-pipelining", where we employ much denser pipelining, i.e. as small as 17 fan-out-of-4 (FO4) delays per stage, than conventional ULV design practices to reduce leakage energy consumption via shorter cycle time [15]. In addition, two-phase latch-based design is employed to reduce cycle time against process variations [3][15]. The cycle borrowing window of latch-based design re-enables averaging effects of delay variations through the long chain of logic gates, which can be lost in a flop-based dense pipeline scheme. We also explore the design of less-buffered clock distribution for smaller skew variations, which can reduce the cycle time as smaller margins for skew become available [16]. Finally, we investigate the energy-optimal FFT architecture with the focus of leakage energy minimization; and propose a modified R4MDC architecture, which exhibits higher utilization of transistors [4][11].

To verify the effectiveness of the proposed techniques, we apply them in designing a 1024-pt complex FFT core in a 65nm CMOS technology. The prototyped design achieves the energy efficiency of 15.8nJ/FFT while operating at 30MHz at the ultra low supply voltage of 0.27V. Moderate frequency and energy σ/μ spreads of 7% and 2%, respectively, have been observed with 60 prototyped chips.

II. FFT CORE DESIGN

A. Super-pipelining

ULV designs have typically employed relaxed pipelining schemes (i.e. many FO4 delays per stage) due to two major benefits. First, the sequential overhead of both registers and clock distributions becomes much smaller with the relaxed pipeline. Given the large energy consumption in clock distribution and flip-flops [17], this can greatly reduce overall energy consumption. Along with the energy benefit, long paths per stage help to mitigate delay variability through averaging of process variations over many gates. In this respects, a much relaxed pipeline in the range of 50-300 FO4 delays per stage, is often the preferred design choice for the recent ULV designs [6][7][8].

Contrary to such conventional pipeline practices in ULV design, we propose to use significantly shorter pipeline stages and more pipeline registers, which counter-intuitively improves energy efficiency and delay *simultaneously*. This improvement is achieved by leakage energy reduction with the shorter clock period since leakage energy consumption is the integration of leakage power over a clock cycle. This reduced leakage energy consumption extends the useful voltage scaling, which is limited due to the rapid increase of leakage energy overhead. The extra voltage scaling can result in extra switch energy savings. Therefore, by increasing the number of pipeline stages, we can reduce both dynamic and leakage energy consumption, which results in the savings in total energy consumption per operation.

978-1-4673-0145-9/12/$31.00 ©2012 IEEE

However, more pipelining also increases the sequential energy overhead including clock distributions. Therefore, the benefit of pipelining on total energy consumption saturates as the number of pipeline stages becomes larger than a certain point. In this respect, it is important to know the energy-optimal pipeline for given circuit topologies in ULV operations.

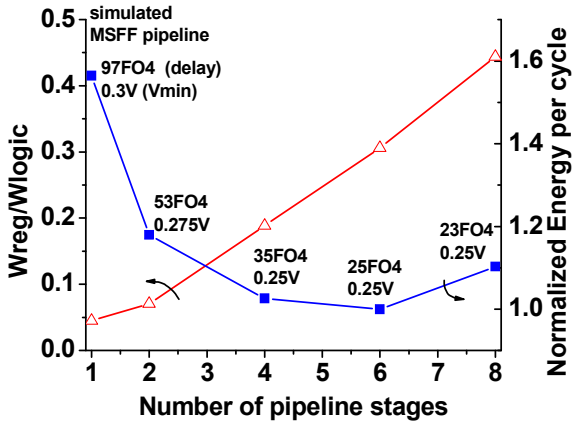


Figure 1. Total register width and energy consumption with different pipeline stages in carry-save multipliers [3]

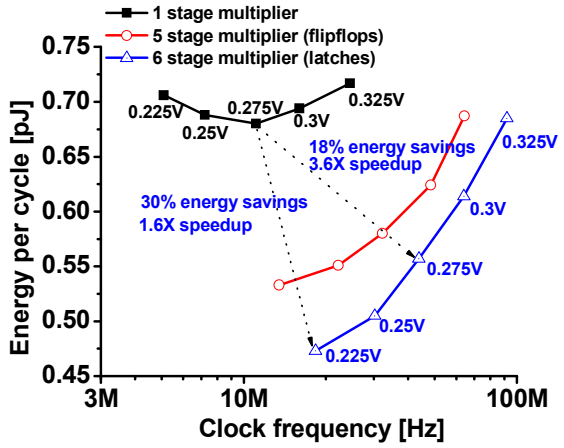


Figure 2. Measurement results of three multipliers [3]

As an important building block for FFT cores and other signal-processing cores, we investigate the optimal number of pipeline stages for a 16b carry save array multiplier. The baseline multiplier takes two 16b inputs and generates 18b outputs. The 14 least significant bits (LSB) are discarded based on our precision analysis in FFT cores. A ripple carry adder (RCA) performs final carry accumulation.

We investigate the energy consumption per multiplication as we increase the number of pipeline stages of multipliers. In this experiment, a master-slave flip-flop (MSFF) is used as pipeline registers. As shown in Figure 1, more pipeline stages initially reduce energy consumption per operation and then saturate due to sequential overheads when the multipliers are pipelined in 4-6 stages. At this point, the total width of the MSFF employed in the multiplier become about 20-30% of the total width of the logic gates. (i.e. W_{reg}/W_{logic}= 0.2~0.3) The energy saving voltage scaling extended from 0.3V with non-pipelined designs to 0.25V with 4-8 stage pipelined designs. We also find that the delay per stage at the energy-optimal pipeline reduces by 4× from 97 with the non-pipelined design to 25 FO4 delays with the 6 stage design.

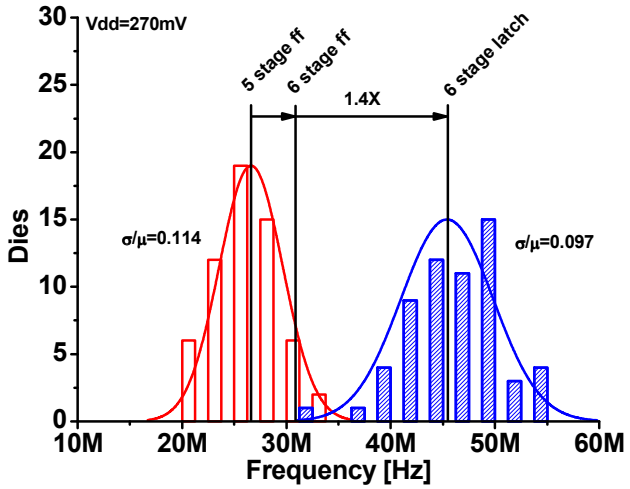


Figure 3. Measured performance distribution of two pipelined multipliers from 60 dies [4]

B. *Two-Phase Latch-based Design*

We also investigate the use of two-phase latches for pipeline registers instead of the conventional MSFF to utilize cycle borrowing ability. As briefly pointed out, one advantage of long pipeline stage is to mitigate delay variability through averaging of process variations over many gates. In the drastically reduced pipeline stages in the super-pipelining scheme, such benefits may disappear and the design can suffer from delay variability. In the densely pipelined designs, contrary to MSFFs, two-phase latch based design can still maintain the advantage by removing hard boundaries in the pipeline and re-establishing the averaging action through long paths that are present in un-pipelined designs.

Guided by the super-pipelining and latch-based techniques, we design a 6-stage multiplier in a 65nm CMOS technology. Additionally, we perform several circuit-level optimizations such as sharing local clock buffers and merging logic-register to further reduce the cycle time. The final accumulation adder is also replaced with a faster variable length carry skip adder, which exhibit better delay and energy tradeoff than a 2-stage ripple carry adder. The resulting pipeline stage delay is only 17 FO4 including sequential overhead, which is 3-17× shorter than the conventional practices in ULV designs [6][7][8][12].

We fabricated 5-stage MSFF-based and 6-stage latch-based pipelined multipliers (the optimal numbers of stages for each sequential element) along with an unpipelined design serving as the baseline, all in a 65nm CMOS technology. Energy and performance measurements are shown in Figure 2. The proposed 6-stge latch-based design consumes only 0.47 pJ/multiplication at 225 mV and achieves 30% energy savings with 1.6× higher performance at its own energy-optimal point V_{opt}, compared to the baseline. Alternatively, it achieves 18% energy savings with 3.6× better performance at a fixed operating voltage of 275 mV compared to the unpipelined design.

We also compare the measured maximum clock frequency of 60 MSFF-based and latch-based designs. Expected performance improvement in moving from 5 to 6 pipeline stages in a MSFF-based design are obtained from simulation and used to scale 5-stage results for an iso-stage comparison. As shown in Figure 3, the latch-based design recovers averaging effects through its time borrowing capabilities and achieves better variation tolerance,

978-1-4673-0145-9/12/$31.00 ©2012 IEEE

translating to 1.4× higher average performance than MSFF-based design although the delay spread (σ/μ) of the latch-based design moderately improves due to global variations across the dies.

C. FIFO Design

In the pipelined FFT architecture (will be discussed in Section II.E), FIFOs in the commutators can contribute as much as 29% of the total FFT energy consumption. To address this, we replace the address decoder with a cyclic address generator for reduced energy and use logic-based readout paths for improved performance, as shown in Figure 4. A single latch is used to store one bit instead of double latches (e.g. MSFF). Simulation results show that the proposed FIFO design consumes 12% lower energy while improving performance by 20% over a memory with MUX-based readout.

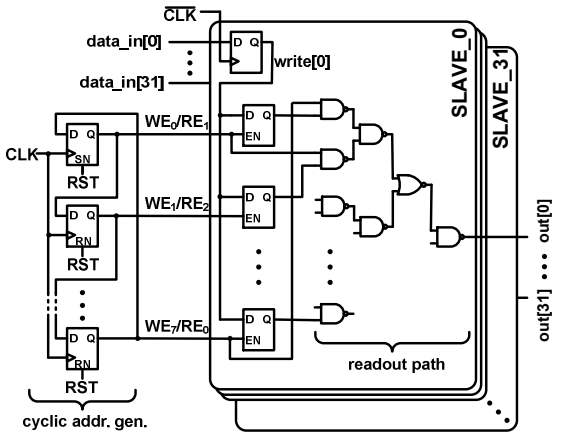


Figure 4. A proposed 8-word FIFO design with a cyclic address generator and logic-based readout paths [15]

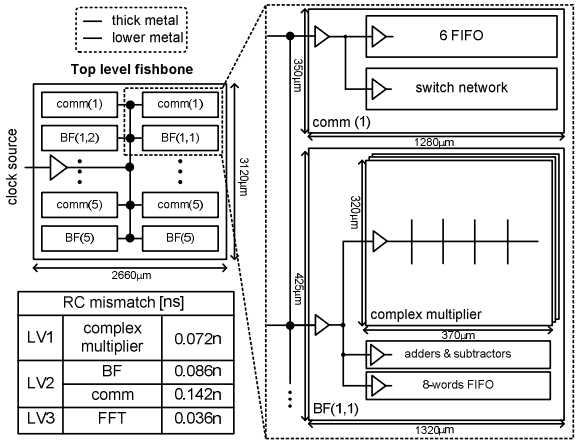
D. Robust Clock Network Design

Another challenge in ULV design is robust clock distribution due to the severe delay variability. This becomes particularly challenging with the super-pipelining scheme as the number of sinks substantially increases. Unmanaged skew, slew and their variability can mandate a pessimistic margin in clock cycle, which undermine energy efficiency through leakage energy waste.

In the nominal voltage regime, a good amount of clock buffer is employed to improve skew and slew as the gate delay is small compared to the RC delay of global clock networks. In the ULV regime, however, exponentially increased gate delay dominates while RC delay no longer contributes appreciably to clock path delay mismatches. In contrast, buffer mismatch significantly impacts clock distribution delays and adding more buffers leads to higher clock skew. Mismatch can be effectively suppressed by employing only a small number of large buffers, since they are robust to random process variations [5].

Based on this observation, we employ a 3-level clock distribution network for the FFT core. In order to exploit the small RC delay in ULV regime, we assign a single buffer (W_{nfet}=33.54μm) for an entire complex multiplier, which includes three 16b multipliers, five adders, a Look-Up Table (LUT), and a controller. Each complex multiplier has a footprint of $0.37\times0.32mm^2$ and 750 sinks. As shown in Figure 5, we perform similar clock buffer assignments for other blocks from the first level to the second and third level. In addition, we seek to minimize RC mismatch using a balanced tree scheme such as [9] in the clock network. For example, the third level has a fishbone-shaped clock network. We also selectively use thick metals for further minimizing RC mismatch without adding buffers.

Monte Carlo simulations are performed on the extracted clock network across supply voltage considering random process variation. As shown in Figure 6, SPICE simulations show that +2σ skew is limited to 0.68×FO4 delays at 0.27V, which is only 2% of the measured clock cycle (30MHz), showing the proposed clock network operates robustly under process variations. We also observe excellent skew and slew variability (σ/μ) of 0.18 and 0.08, respectively, at 0.27V. The amount of RC mismatch (Figure 5) in each clock tree level is at most 0.14×FO4 delays, or 0.43% of the clock cycle. These results confirm that the proposed buffered clock network can effectively improve skew and slew variations.



RC mismatch [ns]		
LV1	complex multiplier	0.072n
LV2	BF	0.086n
	comm	0.142n
LV3	FFT	0.036n

Figure 5. RC-matched 3-level distribution network. The maximum RC mismatch values are in a table [15]

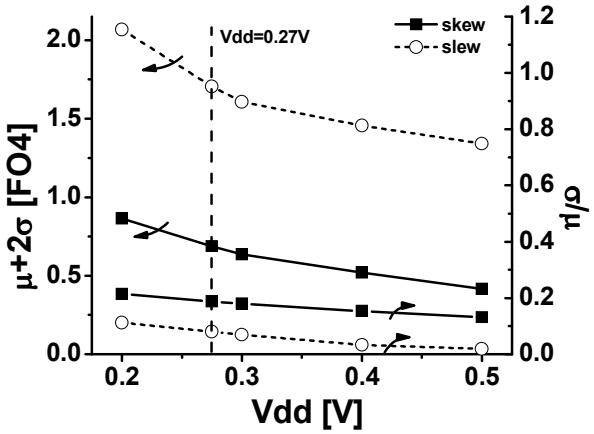


Figure 6. Simulated clock skew and slew variability [16]

E. FFT Architecture Optimization

We also explore FFT architectures to minimize leakage energy overhead as it can allow larger potential energy savings by extending useful voltage scaling. In a traditional memory-based FFT, most memory cells idle while a single butterfly unit processes data word by word over many clock cycles. These idling cells waste leakage energy, harming energy efficiency and voltage scalability. On the other hand, conventional pipeline architectures such as MDC (Multi-path Delay Commutator), have high memory utilization but low butterfly unit activity [10].

Therefore, as shown in Figure 7,we propose to modify MDC to accept 4 inputs concurrently with a new commutator configuration, enabling full utilization of both butterflies and memory elements. Additionally, we use two of the modified MDC lanes to double throughput and halve memory counts per lane, reducing leakage

978-1-4673-0145-9/12/$31.00 ©2012 IEEE

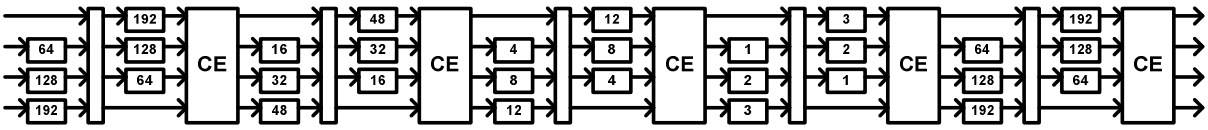


Figure 7. Modified R4MDC architecture (only one lane shown, CE: computational elements, block with a number: FIFO with word depth) [11]

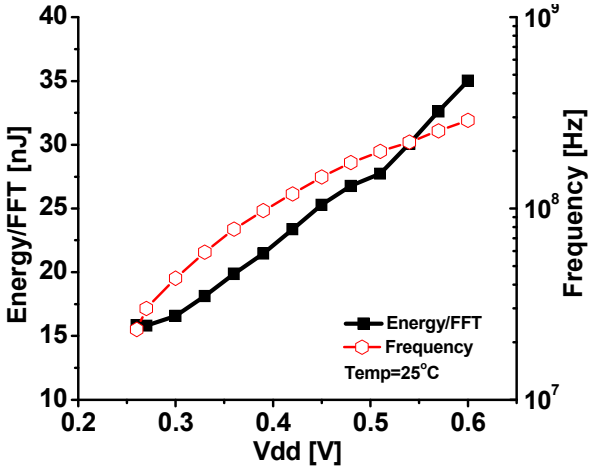


Figure 8. Measured (a) energy consumption and (b) performance of the FFT core

energy consumption from commutators. These modifications improve energy efficiency and throughput by 2.8× and 6.2×, respectively, compared to a radix-4 memory-based FFT core [11].

III. Measurement Results

The FFT core is designed using the described circuit and architectural techniques and fabricated in a 65nm CMOS technology. Figure 8 provides measured energy and performance results for the proposed FFT core. The core consumes 15.8 nJ/FFT at a measured maximum clock frequency of 30 MHz at 270 mV, yielding 240 Msamples/s. The energy efficiency and throughput are improved by 2.4X and 10~100×, respectively, compared to the typical ULV FFT designs [12][13][14]. At 600 mV the proposed design consumes 35.0 nJ/FFT at a clock frequency of 290 MHz, and this energy efficiency is 2× better than the high performance design in [14] at the same throughput. It is functional across a wide temperature range from -20 to 80°C. The average energy consumption and clock frequency at Vdd = 300mV are 17.1nJ/FFT and 41MHz, respectively, as measured across 60 die. It also shows modest frequency and energy spreads of only 7% and 2%, respectively, in terms of σ/μ. The total area of the FFT core is 8.3 mm^2 (2.66×3.12mm).

IV. Conclusions

In this paper, we investigate various circuit and architecture level techniques to minimize leakage overhead and extend useful voltage scaling in order to achieve the energy efficiency beyond the conventional voltage scaling can obtain. We propose super-pipelining techniques and two-phase latch based design to reduce cycle time while improving delay variability and performance. The less-buffered clock network design with a +2sigma skew of 0.68FO4 is proposed to eliminate a pessimistic margin for skew and slew variability in ULV regimes. The energy-optimal FFT architectures are explored to reduce leakage waste and extend useful voltage scaling. The effectiveness of these techniques is verified through a ultra-low energy FFT core demonstration, which achieves a significantly higher energy efficiency (2.4×) and performance (10-100×), compared to the previous state-of-the-art designs.

Acknowledgment

The authors acknowledge the supports from Army Research Laboratory and the chip fabrication from STMicroelectronics

References

[1] B. Zhai, et al., , "Theoretical and Practical Limits of Dynamic Voltage Scaling," in *ACM/IEEE Design Automation Conf.*, pp. 868-873, May 2005

[2] B.H. Calhoun, et al., "Characterization and Modeling Minimum Energy Operation for Subthreshold Circuits," *ACM/IEEE International Symposium on Low Power Electronics and Design*, pp.90-95, 2004

[3] M. Seok, et al., "Pipeline Strategy for Improving Optimal Energy Efficiency in Ultra-Low Voltage Design" *ACM/IEEE Design Automation Conference,* pp.990-995, 2011

[4] D. Jeon et al., "A Super-Pipelined Energy Efficient Subthreshold 240MS/s FFT Core in 65nm CMOS," *IEEE Journal of Solid-State Circuits,* vol.47, no.1, pp.23-34, 2012, invited

[5] J. Konwg, et al., "Variation-driven Device Sizing for Minimum Energy Sub-threshold Circuits," *ACM/IEEE International Symposium on Low Power Electronics and Design*, pp.8-13, 2006

[6] M. Seok, "The Phoenix Processor: A 30pW Platform for Sensor Applications," *IEEE Symposium on VLSI Circuits*, pp.188-189, 2008

[7] J. Kwong, et al., "An Energy-Efficient Biomedical Signal Processing Platform," *IEEE Journal of Solid-State Circuits*, vol.46, no.7, pp.1742-1753, Jul., 2011

[8] S.R. Sridhara, et al., "Microwatt Embedded Processor Platform for Medical System-on-Chip Applications," *IEEE Symposium on VLSI Circuits*, pp.15-16, 2010

[9] T.-H. Chao, et al., "Zero Skew Clock Routing with Minimum Wire Length," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol.39, no.11, pp. 799-814, Nov., 1992

[10] E. E. Swartzlander, et al., "A radix 4 delay commutator for fast Fourier transform processor implementation," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 5, pp. 702-709, Oct. 1984.

[11] D. Jeon, "Energy-Optimized High Performance FFT Processor," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.1701-1704, 2011

[12] A. Wang et al., "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 310-319, Jan. 2005

[13] C.-H. Yang, et al., "A 5.8mW 3GPP-LTE Compliant 8×8 MIMO Sphere Decoder Chip with Soft-Outputs," *IEEE Symposium on VLSI Circuits*, Jun. 2010, pp. 209-210

[14] Y. Chen, et al., "A 2.4-Gsample/s DVFS FFT Processor for MIMO OFDM Communication Systems," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 5, pp. 1260-1273, May. 2008

[15] M. Seok, "A 0.27V, 30MHz, 17.7nJ/transform 1024-pt Complex FFT Core with super-pipelining," *IEEE International Solid-State Circuits Conferences*, pp. 342-344, 2011

[16] M. Seok, "Robust Clock Network Design Methodology for Ultra-Low Voltage Operations," *IEEE Journal on Emerging and Special Topics on Circuits and Systems*, vol.1. no.2, pp.120-130, 2011, invited

[17] N. Magen et al., "Interconnect-Power dissipation in a microprocessor," *International Workshop on System level interconnect prediciton*, pp. 7.13, 2004