# Misleading Energy and Performance Claims in Sub/Near Threshold Digital Systems

Yu Pu[1], Xin Zhang[1], Jim Huang[1], Atsushi Muramatsu[2], Masahiro Nomura[2], Koji Hirairi[2],
Hidehiro Takata[2], Taro Sakurabayashi[2], Shinji Miyano[2], Makoto Takamiya[1], Takayasu Sakurai[1]
University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan[1]
Semiconductor Technology Academic Research Center, Japan[2]
Email: {ypu, zhangxin, jimhuang, mtaka, tsakurai}@iis.u-tokyo.ac.jp;
{muramatsu.atsushi, nomura.masahiro, hirairi.koji, takata.hidehiro, sakurabayashi.taro, miyano.shinji}@starc.or.jp

***Abstract*— Many of us in the field of ultra-low-$V_{dd}$ processors experience difficulty in assessing the sub/near threshold circuit techniques proposed by earlier papers. This paper investigates five major pitfalls which are often not appreciated by researchers when claiming that their circuits outperform others by working at a lower $V_{dd}$ with a higher energy-efficiency. These pitfalls include: *i)* overlook the impacts of different technologies and different $V_{th}$ definitions, *ii)* only emphasize energy reduction but ignore severe throughput degradation, or expect impractical pipelining depth and parallelism degree to compensate this throughput degradation, *iii)* unrealistically assume that memory's $V_{dd}$ and energy could scale as well as standard cells, *iv)* use the highest temperature as the worst timing corner as in the super-threshold, but in fact negative temperature becomes much more detrimental in the sub/near threshold regime, *v)* pursue just-in-need $V_{dd}$ to compensate effects of PVT, but without considering the high energy loss on DC-DC converters. Therefore, the actual energy benefit from using a sub/near threshold $V_{dd}$ can be greatly overestimated. This work provides some design guidelines and silicon evidence to ultra-low-$V_{dd}$ systems. The outlined pitfalls also shed light on future directions in this field.**

## I. Introduction

The latest IEEE standards, such as WiMAX and 4G wireless communication, require SoCs on portable systems to handle far more complex multimedia applications than ever before. High computational-performance and high energy-efficiency are thus extremely crucial to the digital processors embedded in these SoCs. The semiconductor industry expects the energy of consumer electronics to be reduced by an order of magnitude in the next 5 years. An emerging trend for lowering energy is to scale the supply voltage $V_{dd}$ to the sub/near threshold region, which brings not only quadratic dynamic energy savings, but also super-linearly reduced leakage current. However, we recognize that it is quite difficult to assess the sub/near threshold circuit techniques proposed by earlier papers, as the papers often do not appreciate five major pitfalls which may cause impractically favorable results. These pitfalls include: *i)* overlook the impacts of different process technologies and different $V_{th}$ definitions, *ii)* only emphasize energy reduction but ignore severe throughput degradation, or expect impractical pipelining depth and parallelism degree to compensate for this throughput degradation, *iii)* unrealistically assume that memory's $V_{dd}$ and energy could scale as well as standard cells, *iv)* assume that a higher temperature results in a worse speed. While this is true in the super-threshold, in fact negative temperature is much more detrimental in the sub/near threshold region, *v)* pursue just-in-need $V_{dd}$ to compensate effects of PVT on performance and also in this way minimize $V_{dd}$ guardbandings, but without considering the high energy loss when converting a nominal battery voltage to a sub/near threshold voltage.

To investigate the severity of these pitfalls, representative digital circuits and DC-DC converters are fabricated in a 65nm CMOS process. We conclude that the actual energy benefit from using a sub/near threshold $V_{dd}$ can be greatly overestimated by some earlier papers. This paper provides some guidelines and silicon evidence to the design of ultra-low-$V_{dd}$ systems. The readers please be aware that it is not the intention of this paper to suggest that some earlier works were misleading. In fact, because sub/near threshold design is a very new field, researchers need more effort to setup widely-agreed, community-defined metrics. The outlined pitfalls also shed light on future directions in this field.

## II. Pitfall 1: Overlook the impacts of different technologies and different $V_{th}$ definitions

TABLE I
A Performance Comparison

| Processors | sub/near threshold | nominal supply |
|---|---|---|
| 45nm Accelerator [1] | 80MHz at 0.4V | 2.3GHz at 1.1V |
| 65nm SubJPEG [2] | 2.5MHz at 0.4V | 300MHz at 1.2V |
| 65nm MSP430 [3] | 0.1MHz at 0.4V | 300MHz at 1.2V [a] |

[a]The maximum frequency is not released. 300MHz at the nominal 1.2V is our estimation for the MIT's MSP430-like DSP.

Table I compares the performance of three recent works: (*a*) Intel's 45nm CMOS 300mV 4-Way sub-word parallel accelerator [1]; (*b*) *SubJPEG* [2], NXP's 65nm ASIC JPEG co-processor; (*c*) MIT's 65nm TI-MSP430-like DSP processor with embedded DC-DC converter [3]. Table I clearly shows that the performance gap of (*a*), (*b*) and (*c*) at aggressively scaled $V_{dd}$s is much larger than at nominal $V_{dd}$s. In fact, depending on the difference of $V_{th}$s, the on-current $I_{on}$ ratio between different technologies can be up to many orders of magnitude in the sub/near threshold region, in contrast to only a few times at the nominal $V_{dd}$. We also note that, there lacks a clear and widely-agreed definition of $V_{th}$ in sub/near threshold design. Since $V_{th}$ is a non-existing and artificial parameter, it has many different interpretations. Many previous works cite the $V_{th}$ values from their foundries. However, for different foundries the definitions of $V_{th}$ can be arbitrary and quite different. As an example, for the 65nm low-power(LP) CMOS process technology which is used throughout our research, the foundry's $V_{th}$ is about 0.3V, which is defined as the $V_{gs}$ when $I_{ds} \approx 2\mu A$ at 1.2V. However, the actual watershed between the exponential region and the linear region of this process is around 0.6V according to the simulation. Therefore, it is quite difficult to fairly compare the sub/near threshold circuit techniques proposed by earlier papers, unless we know their $V_{th}$ definitions. Some "state-of-the-art" circuits which are claimed to have functioned "at an even lower $V_{dd}$ than others" or "well-below the $V_{th}$" are not really because of circuit

978-1-4244-8192-7/10/$26.00 ©2010 IEEE

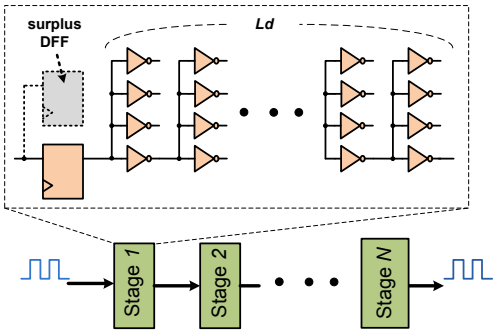



Fig. 1. Diagram of pipelined circuits with different *Ld* and *N*

technique improvements, but just because they rely on processes with lower $V_{th}$s or use different $V_{th}$ definitions.

We hereby call for widely-agreed and community-defined metrics. This is similar to a well-known problem in the field of parallel computer architecture: to get more impressive performance than others, some architectures quote 8 or 16-bit results, not 32-bit results, so later in the famous Roza graph [4] all processors' performance must be scaled to 32-bit operations, in order to give a fair performance comparison.

## III. Pitfall 2: Overlook Severe Throughput Degradation

When $V_{dd}$ scales to the sub/near threshold region, the rapidly diminished driving current causes a severe throughput degradation, hence discouraging aggressive $V_{dd}$ scaling from being applied to medium/high throughput consumer electronics. Most of the existing sub/near threshold prototype chips focused on pursuing optimum energy points or improving functioning yield for KHz range sensor applications, e.g., [3] [5] [6] [7]. Only [1] [2] [8] considered throughput as a serious issue. Many researchers believe this throughput degradation is not a problem as it may be easily compensated by deep pipelining and massive parallelism while maintaining an ultra-low energy/operation. Therefore, one purpose of this research is to investigate the impacts of pipelining depth and parallelism degree on the throughput and the energy in a sub/near threshold system.

To understand the lower bound of energy reduction by $V_{dd}$ scaling, we assume embarrassing pipelining and parallelism are used, meaning that no performance penalty due to data and control dependencies would be incurred from using deeper pipelining and more parallelism. This assumption is ideal for general-purpose CPUs, because it completely neglects the overhead such as hazard control, branch prediction, parallel programming software issues for multi-threading cores. Instead, this assumption suits better the computation-oriented streaming processors, e.g., single-instruction multiple-data (SIMD) processor which exploits the inherent data-level parallelism in many algorithms. To represent a high-performance processor, our baseline circuit consists of 9 pipeline stages and each pipeline has 24×FO4 delay, i.e., ($L_{baseline}$=24, $N_{baseline}$=9). Choosing 24×FO4 delay per stage is suggested by [9], because Intel's high-end Pentium-4 has an about 20×FO4 delay [10], the 24×FO4 delay is fairly representative for processors with slightly shallower pipelining than the Pentium-4. As $N_{baseline}$ increases, the effort to handle pipelining hazards grows rapidly. A processor with less than 12×FO4 delay is not likely to happen because the instruction stalls due to pipelining control overhead can easily offset the increase in instruction throughput achieved from having more pipelining stages.

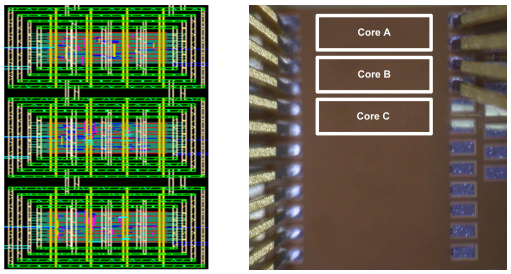
When the baseline circuit changes its pipelining depth from $N_{baseline}$ to *N*, the corresponding logic depth becomes:

$$L_d = N_{baseline} L_{baseline} / N \quad (1)$$

$T_{pipeline}$, the delay of each pipeline stage, is described as following:

$$T_{pipeline} = L_d T_{FO4} + T_{DFF} \quad (2)$$

We are aware that transistor-level simulations are not capable of characterizing energy (especially various and complex mechanisms of leakage current) at scaled $V_{dd}$s with high confidence, so we fabricated three cores: *i)* Core A, the baseline core, ($L_d$=24, $N$=9) *ii)* Core B, ($L_d$=18, $N$=12) *iii)* Core C, ($L_d$=12, $N$=18). The circuit schematics are illustrated in Figure 1. Intuitively, the number of DFFs grows linearly with the increased *N*. However, [11] argued that in reality the numder of DFFs follows $N^\rho$, where $\rho$ is a factor indicating a super-linearly increased number of DFFs with the pipelining depth *N* on datapaths (e.g., pipelining a multiplier). Following them, we choose $\rho$= 1.2 in this research. Therefore, some surplus DFFs are inserted intentionally in Core B and Core C, as also shown in Figure 1. The cores' layout view without I/O cells and pads, and the die photo are shown in Figure 2.




(a) (b)

Fig. 2. Core A, Core B, Core C: (a) layout view (b) die photo

The energy per clock cycle can be modeled as following:

$$E_{cycle} = \alpha E_{switch}(V_{dd}) + I_{leakage}(V_{dd}) T_{cycle} \quad (3)$$

where $\alpha$ is the average switching activity factor of all the internal nodes, $E_{switch}(V_{dd})$ is the $V_{dd}$ dependent switching energy per cycle. $I_{leakage}(V_{dd})$ is the total leakage current which is also dependent on $V_{dd}$, and $T_{cycle}$ is the fastest achievable operating clock cycle time. Among these parameters, $E_{switch}(V_{dd})$ is insensitive to process variations (PV), but $I_{leakage}(V_{dd})$ and $T_{cycle}$ are strongly influenced by PV. The measured $E_{switch}(V_{dd})$ and $I_{leakage}(V_{dd})$ are shown in Figure 3 and Figure 4. Since low-power (LP) process is used in our implementation, the leakage energy is quite small compared to the switching energy.

To probe the internal signals on the wafer, the prototype chip uses analog I/O cells and pads. Unfortunately, these I/O cells and pads present very heavy loadings to the output signals, hence preventing us from measuring the achievable $T_{cycle}$, in spite of the fact that all the three cores can reach 2GHz at 1.2V $V_{dd}$. In the following analysis simulation results are used to resolve this issue. Our simulation is based on the recently released PSP model from Philips, which claims superior accuracy over the BSIM4 when modeling $I_{on}$ at low $V_{dd}$. We do not expect the $T_{cycle}$ difference between silicon measurement and simulations to change our conclusions significantly.

In Figure 5, assuming $\alpha$ = 30%, the energy-efficiency of each core at different $V_{dd}$s are normalized to the energy-efficiency of the baseline core (Core A) at 1.2V $V_{dd}$. For multimedia processors $\alpha$ = 30%

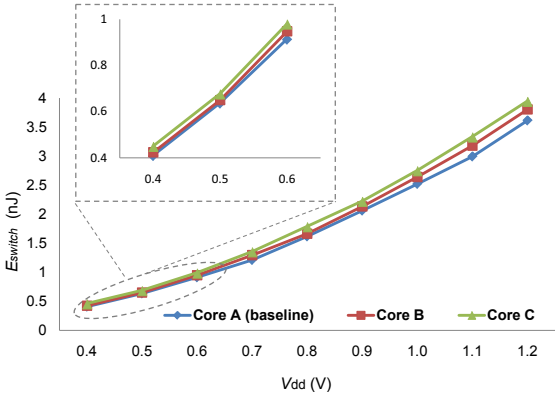



Fig. 3. Measured switching energy $E_{switch}$ at different $V_{dd}$s

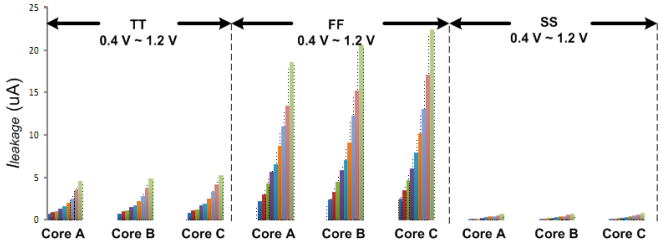



Fig. 4. Measured $I_{leakage}$ at different $V_{dd}$s and process conditions

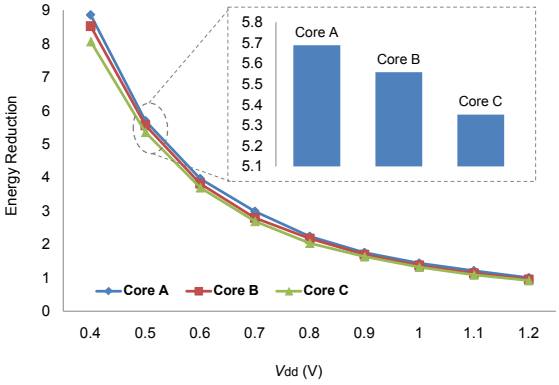



Fig. 5. Normalized energy-efficiency of core (A),(B),(C) at different $V_{dd}$s ($\alpha$= 30%)

is quite common, and for general purpose processors $\alpha$ is typically smaller due to very low activity on memories and caches [12]. From Figure 5, it is clear that the lower the $V_{dd}$, the higher the energy-efficiency. However, to compenstate for throughput degradation, the associated parallelism degree from simulations is shown in Figure 6. Once $V_{dd}$ scales to below 0.5V, the number of needed parallel units increases sharply. A larger parallel degree implies a larger silicon area and an increased layout difficulty (e.g., global/semi-global routing). It also means a further degraded yield since the fabrication defects increase in an exponential way when the silicon area increases, i.e., $Yield \propto e^{-Area}$. Assume the imperfect dice for circuits having only one unit is 100 p.p.m (part per million), then the imperfect dice for circuits having 100 parallel units quickly goes to 10000 p.p.m. In addition, depending on the targeted applications, there is also a

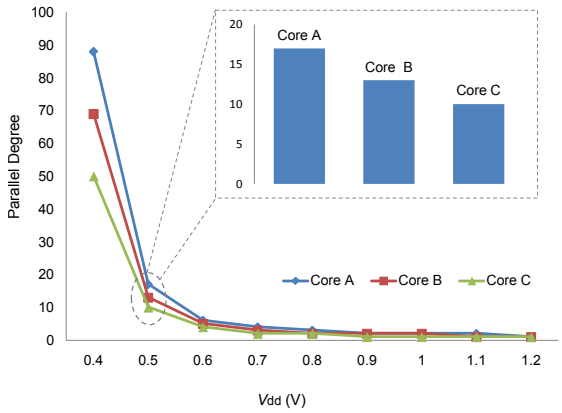



Fig. 6. The associated parallel degree of core (A),(B),(C) at different $V_{dd}$s from simulation

critical parallelism degree according to the famous Amdahl's Law [13] [14]. The speedup from having more parallelism than this critical degree will saturate. For example, only 32× speedup is observed on a 128-core system [15]. Therefore, making a good trade-off between the parallelism degree and the energy-efficiency becomes extremely cruicial at the sub/near threshold region. At 0.5V $V_{dd}$, Core C, which has the deepest pipelining, still needs 10 units performing in parallel to satisfy the throughput, meanwhile gaining about 5× energy reduction. Recall that the control overhead caused by pipelining and parallelism has been completely ignored, so this 5× reduction can be far beyond what we could reach in a real processor. While many papers claim a near 10× higher energy-efficiency, they usually do not appreciate these difficulties to maintain a high throughput.

## IV. Pitfall 3: Assume Memory's $V_{dd}$ and Energy Could Scale as Well as Standard Cells

When analyzing the energy of an ultra-low-$V_{dd}$ processor, some researchers assume that memory's $V_{dd}$ and energy could scale as well as standard cell based logic in their performance modeling tools. However, current practice is that memory is not as resilient as standard cells to $V_{dd}$ scaling. Commercial 6-T SRAMs which achieve high density fail at 2/3 of the nominal supply due to static noise margin (SNM) degradation [16] [17]. The data stored in bit-cell is susceptible to a very small injected bitline noise and may flip. Some recent SRAM chips can scale reliably to below 0.4V [18] [19] [20] [21] [22]. To improve SNM, these works either add extra devices within bit-cell or enlarge bit-cell's size. However, they also introduce considerable area and performance overheads compared to commercial SRAMs. As an example, to build a SRAM block with the same dimension, the MIT 10-T SRAM realization is 66.7% bigger and 3× slower (at 0.7V and above) than SRAMs that are synthesized with a commercial low-power memory generator. The low density, low speed and high energy consumption in the super-threshold mode are big limitations for these proposed works [23] to be applied in medium/high speed consumer electronics.

SRAMs bitcells' energy accounts for more than 80% of SRAMs' total energy, while other components such as sense-amplifiers, word-line/bitline drivers, address decoders and writing circuits only occupy less than 20% [24]. Therefore, SRAMs bitcells' energy is the main concern in our analysis. We approximate the delays of logic cell and SRAM bitcell by Equations (4) and (5). Equation (6) and (7) are also our approximation for switching energy of logic cell and

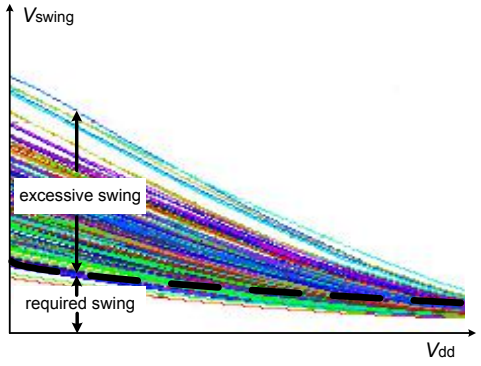



Fig. 7. Illustration: the actual and required bitline swings when $V_{dd}$ scales

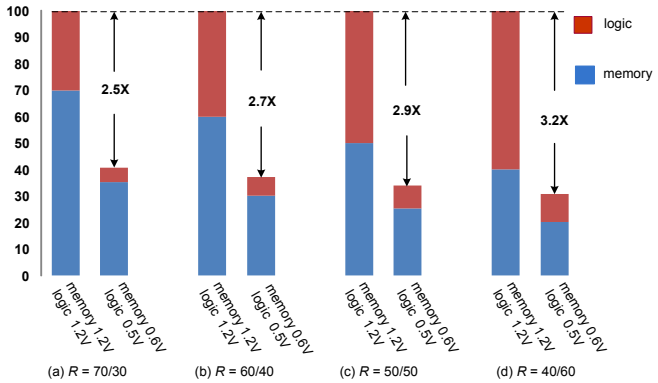



Fig. 8. The normalized energy vs. $R$ from simulation. $R$ signifies $(Memory\ energy)$ / $(Logic\ energy)$ at the non-scaled 1.2V.

SRAM bitcell in one clock cycle. $C_{load}$ is the loading capacitances including both gate and interconnection wire capacitances. $I_{drive}$ is the average charging/discharging currents. $C_{bitline}$ and $I_{bitline}$ are the loading capacitance and the average current on a SRAM bitline. $V_{swing}$ is the bitline swing, which must exceed a minimum magnitude (e.g., 10% nominal $V_{dd}$) required by sense-amplifiers to make correct decisions.

$$STDCell_{delay} \propto C_{load}V_{dd}/I_{drive} \tag{4}$$

$$SRAMBitcell_{delay} \propto C_{bitline}V_{swing}/I_{bitline} \tag{5}$$

$$E_{STDCell} \propto C_{load}V_{dd}V_{dd} \tag{6}$$

$$E_{SRAMBitcell} \propto C_{bitline}V_{dd}V_{swing} \tag{7}$$

$V_{swing}$ cannot scale with $V_{dd}$. Contrarily, according to our simulation, the actual $V_{swing}$ may even increase because of process variations. The required minimum $V_{swing}$ also increases slightly due to degradation on sense-amplifiers' sensitivity, as illustrated in Figure 7. From Equation (4) to Equation (7) we can conclude that : 1) when $V_{dd}$ scales, SRAMs' speed deteriorates faster than standard cells. In other words, if both SRAMs and logic scale to the same $V_{dd}$, the SRAMs will become the performance bottleneck. 2) SRAMs' energy can only scale (sub-)quadratically with $V_{dd}$, in contrast to standard cells' quadratic relationship with $V_{dd}$. These conclusions can interpret the simulation results provided in [16], and they hold true for both differential and single-ended SRAMs. Our SRAM models in Equation (5) and (7) are quite simple, so they can be used as a rule-of-thumb for quick estimation.

For processors operating at the non-scaled $V_{dd}$, it sometimes happens that the memory's energy dominates the total processor's energy, e.g., 70% of video processors' energy is consumed by frame memory [23], 95% of NoC (Network-on-Chip)'s energy is consumed by memory hierarchy and foreground memory [25]. Even for embedded microprocessors, the energy contribution of instruction and data memories can be up to 40% [26]. We introduce a factor $R$, which represents the ratio between memory's energy and logic's energy in application benchmarks at the non-scaled 1.2V. Let us assume that the $V_{dd}$ of logic scales to 0.5V and the $V_{dd}$ of memory scales to 0.6V, otherwise memory's speed becomes the system bottleneck. Figure 8 depicts the normalized energy when $R$ changes from 70/30 to 40/60. From this figure, a 3× higher energy-efficiency can be a realistic expectation for processors consisting of both memory and logic.

## V. Pitfall 4: Assume highest temperature as the worst timing case

Temperature directly affects digital gate's delay. For commercial standard cell library which operates at nominal $V_{dd}$, the worst timing corner case happens at the highest temperature (e.g., 125 °C). We notice that some sub/near threshold works still report the performance at high temperature as the worst timing case. However, an inversed worst case happens at ultra-low $V_{dd}$s, i.e., low temperature (e.g., −25 °C) becomes worst case corner instead of traditional high temperature.

The driving current of an nMOSFET is modeled as:

$$I_{on} \propto \begin{cases} \mu(T)e^{\frac{(V_{gs}-V_{th}(T))}{S(T)}} & (V_{gs} < V_{th}) \\ \mu(T)(V_{gs}-V_{th}(T))^{\beta} & (V_{gs} \geq V_{th}) \end{cases} \tag{8}$$

where $\mu(T)$ is the carrier mobility that is intrinsic to the process, $\beta$ is the velocity saturation effect factor and $S(T)$ is the sub-threshold swing. $\mu(T)$, $S(T)$ and $V_{th}(T)$ are temperature dependent parameters.

As temperature dereases, the carrier mobility $\mu(T)$ increases, which tends to increase the driving current. Meanwhile, $V_{th}(T)$ increases and sub-threshold swing $S(T)$ reduces, which tend to reduce the driving current. In the super-threshold region, both $\mu(T)$ and $V_{th}(T)$ have near linear effects on $I_{on}$. The influence from $\mu(T)$ is slightly stronger, so the combined effect results in an increased $I_{on}$ and hence digital gates run faster. In the sub-threshold region, $V_{th}(T)$ and $S(T)$ have an exponential and dominant influence on $I_{on}$. As a result, the combined effect shows a significantly reduced $I_{on}$ hence digital gates run much slower. Figure 9 shows the simulated gate delay at varied temperature, which is normalized to the delay at 125 °C temperature. As seen, from 125 °C to −25 °C, the gate which operates at nominal $V_{dd}$ gains speedup by 12%, however, the gate operating in the sub-threshold region becomes up to 16× slower. This means that negative temperature can be even as detrimental as process variations to the sub/near threshold designs! It is thus of great importantance to have standard cell libraries characterized at low temperatures and use the lowest temperature case as the worst corner in sub/near threshold designs.

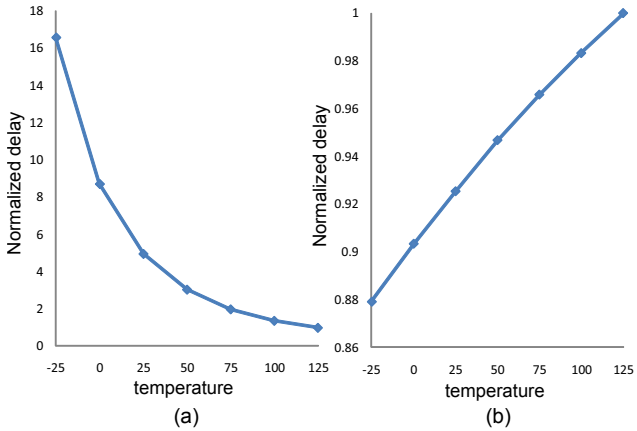



Fig. 9. Simulated gate delay at different temperatures, with normalization to delay at 125 °C (a) at 0.5V $V_{dd}$ (b) at 1.2V nominal $V_{dd}$

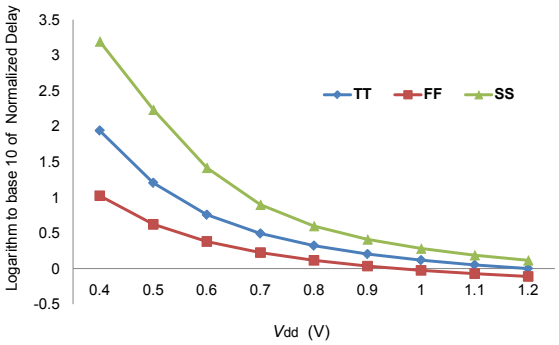



Fig. 10. Simulated delay of Core C at different $V_{dd}$s under TT, SS, FF conditions, with normalization to its delay at 1.2V under TT condition

## VI. Pitfall 5: PVT compensation using just-in-need $V_{dd}$ without considering loss on DC-DC

The $T_{cycle}$ in a processor is limited by the slowest pipelining stage. Temperature changes and process variations (including die-to-die (D2D), within-die (WID) and random variations) result in a very wide spread of delay in the sub/near threshold region. According to the Monte-Carlo simulation, for the three cores introduced in Section III, 50mV $V_{dd}$ guardbanding is enough for compensating the effects of WID and random variations in the sub/near threshold region. In addition, logic paths with longer logic depth $L_d$ has smaller delay variability because the total delay attempts to average out timing variability from each individual gate. However, targeting at the worst case, i.e., the SS corner, more guardbandings must be reserved. Taking Core C as an example, its delays at the scaled $V_{dd}$s under different process conditions are normalized to its delay at the 1.2V $V_{dd}$ under TT condition, and the logarithm results are plotted in Figure 10. As seen, at the SS corner, the $V_{dd}$ should be increased to around $V_{dd} + 100mV$ to meet the speed at $V_{dd}$ under TT condition, so the energy-efficiency will decrease. If temperature change is also considered, the guardbandings should be increased further.

To reduce the guardbandings hence the energy, many papers propose to use just-in-need $V_{dd}$ with delay sensing schemes. Some of them also propose to use bulk-biasing with variable $V_{p-well}$ and $V_{n-well}$, e.g., [27]. With these approaches, the most appropriate voltages can be employed for each individual die, e.g., a lower $V_{dd}$ can be used for faster chips while a higher $V_{dd}$ can be used for slower chips. However, researchers often do not care the energy overhead to generate these ultra-low $V_{dd}$ power supplies. Ideal power supplies with 100% efficiency were assumed. This assumption can be accepted for conventional $V_{dd}$ scaling from nominal $V_{dd}$ to 2/3 nominal $V_{dd}$ because DC-DC efficiency can be easily above 90% within this region. However, the MIT's MSP430-like DSP [3], which is so far the only sub/near threshold $V_{dd}$ system that has an embedded DC-DC converter, observed that the low efficiency for converting voltages from a off-chip battery level to a sub/near threshold level is a major bottleneck limiting system energy reduction.

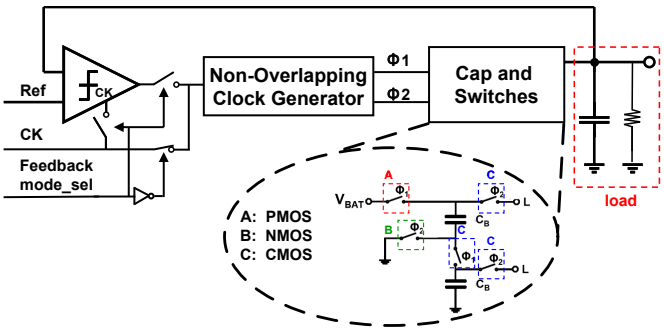



Fig. 11. Diagram of $T_6$ DC-DC converter

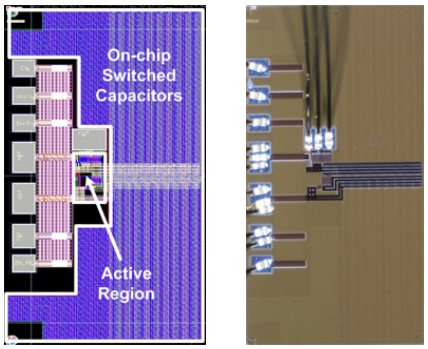
In our study we exploited the MIT's switched capacitor DC-DC converter architecture [28]. This architecture is the most suitable for ultra-low-$V_{dd}$ conversion. Other two well-known architectures, i.e., Low-Dropout (LDO) and Buck-Boost, are not good options in our case, because (1) LDO cannot perform well when the gap between input and output voltages is large; (2) Buck-Boost needs big inductor, so it is difficult to be integrated fully on-chip unless some special techniques such as bondwire spriral inductor [29] or 3-D stacked process can be used.

The theoretical maximum efficiency that can be achieved by this architecture is:

$$\eta = 1 - \Delta V / V_{NL} \tag{9}$$

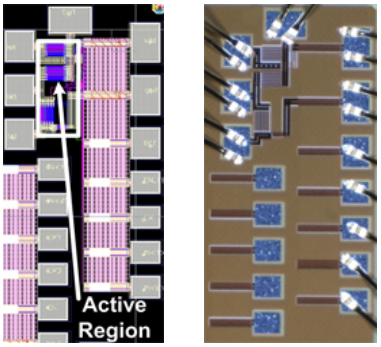
The $V_{NL}$ is the no-load output voltage and the $\Delta V$ is the difference between the actual ouput voltage $V_{out}$ and $V_{NL}$, i.e., $\Delta V = V_{NL} - V_{out}$. Therefore, the higher the $\Delta V$, the lower the maximum achievable conversion efficiency. The difference between the MIT's and our designs is that we use only the $T_6$ topology since we need a narrow output voltage band (0.4V∼0.6V), whereas the MIT's design switches among $T_4$∼$T_{12}$ topologies to cover also super-threshold output voltage range. The $T_4$, $T_6$, $T_8$, $T_9$ and $T_{12}$ are different topologies to generate $V_{NL}$s of $V_{Battery}/3$, $V_{Battery}/2$, $2V_{Battery}/3$, $3V_{Battery}/4$ and $V_{Battery}$. The diagram of $T_6$ topology is shown in Figure 11.

Although we had employed the state-of-the-art DC-DC architecture, we still saw a low efficiency for converting an off-chip battery voltage to the sub/near threshold region. This is because the resistances of power switch transistors become exceedingly large at a very low $V_{dd}$ such that a high $\Delta V$ is unavoidable. In addition to the fundamental energy loss on power switch transistors, the overheads from bottom-plate parastic capacitors, control circuits and gate-drive loss also decrease the efficiency. We fabricated two DC-DC converters, one with on-chip capacitors and the other with off-chip capacitors. The layout views and die photos of the two DC-DC converters are shown in Figure 12 and Figure 13. The measured




(a) (b)

Fig. 12. Switched capacitor DC-DC converters with on-chip capacitors (a)layout view (b)die photo




(a) (b)

Fig. 13. Switched capacitor DC-DC converters with off-chip capacitors (a)layout view (b)die photo

conversion efficiency from 1.2V and 1.5V battery voltages to the sub/near threshold region with varied output current loads is plotted in Figure 14. This figure shows that: 1) $V_{out}$ can never reach $V_{NL}$ due to the fundamental conduction loss on power switches. 2) When $V_{out}$ increases, the conduction loss reduces linearly, so the conversion efficiency gradually grows and eventually reaches a peak. After that the efficiency quickly decreases because the energy overhead from control circuits dominates the total energy. 3) The energy loss due to bottom-plate parasitic capacitors for DC-DC with on-chip capacitors is quite severe, which can be more than 10%.

The achievable peak conversion efficiency from a battery voltage to a near-threshold voltage is around 70%, which is close to the measurement results from the MIT group. Therefore, unlike operating processors near the non-scaled voltage where DC-DC efficiency can be constantly over 90%, the DC-DC energy cost becomes a big challenge for sub/near threshold designs. Although we will continue improving the efficiency, the headroom for further improvement is already rather limited by the intrinsic energy losses on the power switch transistors and the bottom-plate parasitic capacitors.

Taking the DC-DC converter's efficiency also into account, the achievable energy gain from using a sub/near threshold $V_{dd}$ decreases by 30% immediately. In addition, when exploiting a just-in-need $V_{dd}$ to reduce guardbandings, one should also pay attention to the linearly decreased DC-DC efficiency as $V_{dd}$ decreases. The growing DC-DC energy loss largely counteracts the energy savings obtained from lowering the $V_{dd}$ of processors. Since there are cases where the energy consumption of SoCs is dominated by SRAMs whose energy

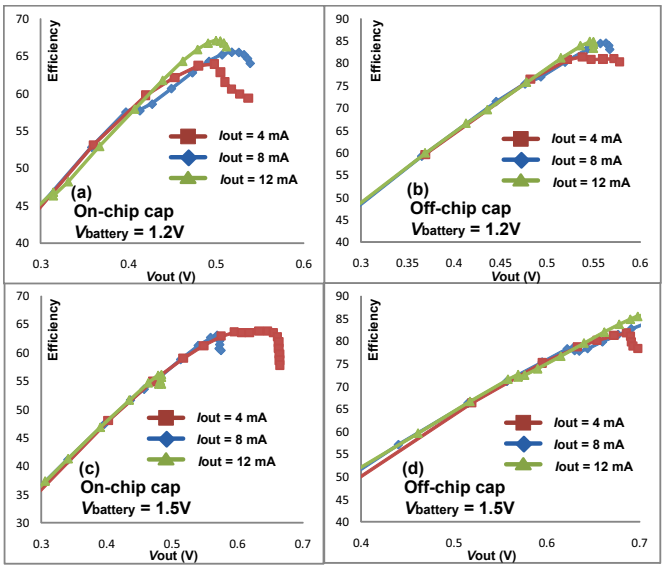



Fig. 14. Efficiency of DC-DC converters: (a) with on-chip capacitor, $V_{Battery}$ = 1.2V (b) with off-chip capacitor, $V_{Battery}$ = 1.2V (c) with on-chip capacitor, $V_{Battery}$ = 1.5V (d) with off-chip capacitor, $V_{Battery}$ = 1.5V

consumption is almost linearly dependent on $V_{dd}$, it may not be of use to further lower the $V_{dd}$. Besides, if the overhead of complicated PVT sensing schemes is also included, the energy benefit from using a just-in-need $V_{dd}$ in a real system with embedded DC-DC converters can be greatly exaggerated in some earlier papers.

## VII. Conclusion

Scaling $V_{dd}$ to the sub/near threshold is the way to ultra-low energy processors. However, many of us doing research in this field recognize that it is necessary to have reflections on existing sub/near threshold circuit techniques. This paper has carefully investigated five pitfalls often not appreciated by researchers. These pitfalls lead to an overestimation of the actual energy benefit from using a sub/near threshold $V_{dd}$ in many earlier papers. To make the dream of 10× energy reduction for consumer electronics come true, more efforts should be made other than aggressive $V_{dd}$ scaling alone. The future directions in this field are, but not limited to:

- Currently conventional super-threshold processes are being used to demonstrate subthreshold circuit ideas. Devices that are optimized for sub/near threshold regime with larger $I_{on}$ and lower $V_{th}$ variability) are preferred. For example, compared to using conventional bulk CMOS processes, fully depleted silicon-on-insulator (FDSOI) CMOS technology shows lower RDF and smaller subthreshold swings. For instance, a RISC processor in FDSOI process which operates at ultra-low-$V_{dd}$ with greatly enhanced computational energy efficiency has been introduced in [30].
- For logic-centric circuits whose energy almost depends quadratically on $V_{dd}$, it is worthwhile exploring a lower $V_{dd}$ even though the linear energy loss on the DC-DC converter is taken into account. To recover throughput, architecture-level solutions (e.g., efficient parallelism) can be helpful but must be used with cautious.
- Make memory's energy super-linearly dependent on $V_{dd}$, being different from its present form where the energy depends linearly

on $V_{dd}$. Otherwise $V_{dd}$ scaling cannot be very effective for those memory-centric SoCs whose energy is dominated by memories.
- Explore more energy-efficient DC-DC converters whose conversion efficiency can be almost independent of output $V_{dd}$. In this paper we used on-chip switched capacitor DC-DC for a fair comparison. Following which, buck-boost DC-DC with 3-D stacked inductors is marked on our roadmap.

## ACKNOWLEDGMENT

This work was carried out as a part of the Extremely Low Power (ELP) project supported by the New Energy and Industrial Technology Development Organization (NEDO).

## REFERENCES

[1] H. Kaul, M. A. Anders, S. K. Mathew, S. K. Hsu, A. Agarwal, R. K. Krishnamurthy, and S. Borkar. A 300mV 494GOPS/W Reconfigurable Dual-Supply 4-Way SIMD Vector Processing Accelerator in 45nm CMOS. In *IEEE ISSCC*, pages 260–263, 2009.

[2] Y. Pu, J. Pineda de Gyvez, H. Corporaal, and Y. Ha. An ultra-low-energy multi-standard JPEG co-processor in 65nm CMOS with sub/near threshold supply voltage. *IEEE JSSC*, 45(3):668–680, 2010.

[3] J. Kwong, Y. Ramadass, N. Verma, and A. Chandrakasan. A 65 nm Sub-$V_t$ Microcontroller With Integrated SRAM and Switched Capacitor DC-DC Converter. *IEEE JSSC*, 44(1):115–126, 2009.

[4] E. Roza. Systems-on-chip: what are the limits? *Electronics and Communication Engineering Journal*, 13(6):249–255, 2001.

[5] M. Seok, S. Hanson, Y.S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw. The Phoenix Processor: A 30pW platform for sensor applications. In *2008 IEEE Symposium on VLSI Circuits*, pages 188–189, 2008.

[6] A. Wang and A. Chandrakasan. A 180-mV subthreshold FFT processor using a minimum energy design methodology. *IEEE JSSC*, 40(1):310–319, 2005.

[7] B. Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, D. Blaauw, and T. Austin. A 2.60 pJ/Inst subthreshold sensor processor for optimal energy efficiency. In *IEEE Symposium on VLSI Circuits*, pages 154–155, 2006.

[8] V. Sze, R. Blaquez, M. Bhardwaj, and A. Chandrakasan. An energy efficient sub-threshold baseband processor architecture for pulsed ultra-wideband communications. In *IEEE ICASSP*, pages 14–19, 2006.

[9] N. S. Kim, T. Kgil, K. Bowman, V. De, and T. Mudge. Total power-optimal pipelining and parallel processing under process variations in nanometer technology. In *IEEE ICCAD*, pages 535–540, 2005.

[10] E. Sprangle and D. Carmean. Increasing processor performance by implementing deeper pipelines. In *IEEE ISCA*, pages 25–34, 2002.

[11] V. Srinivasan, D. Brooks, M. Gschwind, P. Bose, V. Zyuban, P. Strenski, and P. Emma. Optimizing pipelines for power and performance. In *IEEE MICRO*, pages 333–344, 2002.

[12] R. Gonzalez, B. M. Gorden, and M. A. Horowitz. Supply and threshold voltage scaling for low power CMOS. *IEEE JSSC*, 32(8):1210–1216, 1997.

[13] G. M. Amdahl. Validity of the single-processor approach to achieving large scale computing capabilities. In *IEEE AFIPS*, pages 483–485, 1967.

[14] G. M. Amdahl. Can we still keep the faith? In *IEEE ICCAD*, 2007.

[15] J. T. Deutsch and A. R. Newton. multiprocessor implementation of relaxation based electrical circuit simulation. In *IEEE DAC*, 1984.

[16] G. K. Chen, D. Blaauw, T. Mudge, D. Sylvester, and N. S. Kim. Yield-driven near-threshold SRAM Design. In *IEEE ICCAD*, 2007.

[17] Jan Rabaey. *Low Power Design Essentials*. Integrated Circuits and Systems. Springer, 2009.

[18] B. Calhoun and A. Chandrakasan. A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation. *IEEE JSSC*, 42(3):680–688, 2007.

[19] T. H. Kim, J. Liu, J. Keane, and C. H. Kim. A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme. In *IEEE ISSCC*, pages 330–331, 2007.

[20] N. Verma and A. Chandrakasan. A 256 kb 65 nm 8T subthreshold SRAM employing Sense-amplifier Redundancy. *IEEE JSSC*, 43(1):141–149, 2008.

[21] B. Zhai, S. Hanson, and D. Sylvester D. Blaauw. A variation-tolerant sub-200mV 6T subthreshold SRAM. *IEEE JSSC*, 44(10):2338–2348, 2008.

[22] J. Chen, L. Clark, and T. Chen. An ultra-low-power memory with a subthreshold power supply voltage. *IEEE JSSC*, 41(10):2344–2353, 2006.

[23] Y. He, Y. Pu, Z. Ye, S. Moreno Londono, R. Kleihorst, A. Abbo, and H. Corporaal. Xetal-pro: an ultra-low energy and high throughput simd processor. In *IEEE DAC*, pages 543–548, 2010.

[24] M. Do, M. Drazdziulis, P. Larsson-Edefors, and L. Bengtsson. Parameterizable architecture-level SRAM power model using circuit-simulation backend for leakage calibration. In *IEEE ISQED*, pages 557–563, 2006.

[25] A. Lambrechts, P. Raghavan, A. Leroy, G. Talavera, T. V. Aa, M. Jayapala, F. Catthoor, D. Verkest, G. Deconinck, H. Corporaal, F. Robert, and J. Carrabina. Power breakdown analysis for a heterogeneous NoC platform running a video application. In *IEEE ASAP*, pages 179–184, 2005.

[26] K. Keutzer D. Chinnery. Closing the power gap between ASIC and custom: an ASIC perspective. In *IEEE DAC*, pages 275 – 280, 2005.

[27] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye. Adaptive performance compensation with in-situ timing error prediction for subthreshold circuits. In *IEEE CICC*, pages 215–218, 2009.

[28] Y. Ramadass and A. Chandrakasan. Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip. In *IEEE PESC*, pages 2353–2359, 2007.

[29] M. Wens and M. Steyaert. A fully-integrated 0.18$\mu$ m CMOS DC-DC step-down converter, using a bondwire spiral inductor. In *IEEE CICC*, pages 17–20, 2008.

[30] H. Kawaguchi, K. Kanda, K. Nose, S. Hattori, D. D. Antono, D. Yamada, T. Miyazaki, K. Inagaki, T. Hiramoto, and T. Sakurai. A 0.5-V, 400-MHz, $V_{dd}$-hopping processor with zero-$V_{th}$ FD-SOI technology. In *IEEE ISSCC*, pages 106–197, 2003.