Adaptively Provisioned Signal Line Drivers for Variation Robustness

Timothy Normand Miller
Binghamton University (SUNY), Comp. Sci.
millerti@binghamton.edu

André Pouliot
Cégep de Trois-Rivières, C2T3
andre.pouliot@cegeptr.qc.ca

ABSTRACT

We present a mechanism to mitigate process variation in line drivers used extensively for long signal lines such as bit lines in SRAMs. Our objective is to reduce the substantial leakage power (P_L) expended by over-sized drivers. We achieve this by providing multiple amplifiers per driver, wired in parallel. After self-test identifies the optimal combination of amplifiers, they are connected to the circuit statically (e.g. antifuse) or dynamically (e.g. enable transistors). We have performed Monte Carlo simulations over wide ranges of supply voltage and process variation to learn how to build optimal drivers that minimize P_L despite high delay variation. We improve P_L by up to 36% for fixed-voltage circuits and up to 54% for variable-voltage circuits.

1. INTRODUCTION

We are motivated by the demand for ultra-low power circuits that require line drivers whose efficiency and reliability are severely impacted by both technology scaling and very low supply voltage (i.e. near-threshold). We propose a methodology to reduce the impact of process variation on drivers of long signal lines, as found in static RAM circuits and other heavily-loaded signal lines where amplification is required. Although relatively small in number compared to other components, line drivers suffer from much higher leakage current, because they must be much wider in order to charge or discharge the substantial capacitance in a signal line of any nontrivial length. Due to process variation, they must be further oversized to ensure that the worst case is fast enough. As a result, nearly all driver transistors are faster and leakier than they need to be. By mitigating variation, drivers can be made much smaller, reducing the size guard band and their leakage power contribution.

2. RELATED WORK

In [3], SRAM power is reduced by a combination of techniques, including a single-line filter cache (line buffer), finer grained banks, and bit line segmentation into multiple independent segments. Many solutions approach the delay and power problem using smaller granularity banks, to shorten wires, e.g. [10]. Other techniques involve lowering supply voltage and mitigating reliability problems using redundant signaling logic [8] and forward error correction [7, 4, 2].

Some research explores placement and sizing requirements of line drivers to minimize delay and power [5] on medium-length lines. Others consider the incorporation of a sleep mechanism to lower transistor leakage in advanced process technologies [9]. Each of these approaches helps to reduce power in SRAMs by tackling different sources of power loss.

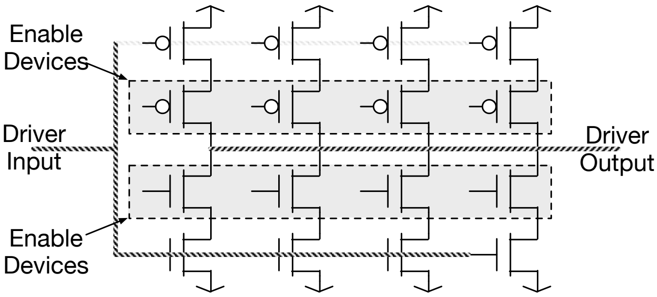


Figure 1: Four-amplifier line driver circuit.

3. DRIVER CIRCUITS

In some cases long wires can be boosted by a series of small repeaters, but this is not true for SRAM bit lines. A single wire connects a write/precharge driver, access transistors for every cell in a column, and one input port of a sense-amp. High capacitive load requires bit line drivers to be very large, fast, and power-hungry to meet timing. For both single and repeater drivers, variation is a major challenge.

3.1 Reference Driver

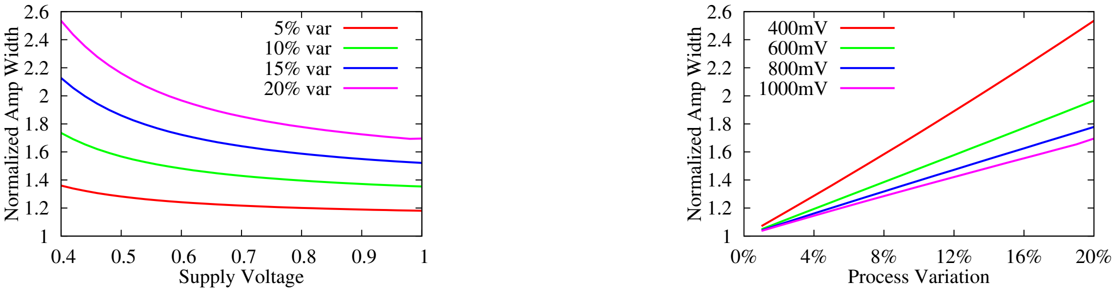
The reference circuit is a CMOS inverter at nominal width (W_R) for drivers used in 22nm technology SRAMs, with *no* process variation, and we use this to set the target switching delay (D_R) at a given supply voltage. In circuits with variation, a larger width ($W > W_R$, by necessity) is chosen so that at least 99% (2.576σ) of drivers of width W have delay $D < D_R$ and leakage power is minimized. In general, tristating may be required, but we assume that is unchanged compared to a standard design.

3.2 Multi-Amplifier Drivers

Figure 1 shows an example of a four-amplifier driver circuit (8 driver transistors), where enable devices connect individual amplifiers to the circuit. With dynamic configuration, amplifier combinations can be altered at run time to support different supply voltages and other variations in operating conditions. Additionally, the enable devices (if fast enough) could serve double-duty for tri-stating, as an alternative to a pass transistor or comparable mechanism. In any case, the enable devices must be sized and doped so that their impedance and delay do not substantially impact the performance of the driver transistors. Alternatively, fuses or antifuses can be employed to configure drivers statically. Effects of these extra components are not addressed here.

Table 1 lists the solutions tested in simulation. We size

W-MOS @ ASPLOS 2014, Salt Lake City, Utah USA



(a) Amp width as a function of voltage for four degrees of variation. (b) Amp width as a function of variation for four different voltages.

Figure 2: Single-amplifier transistor widths, normalized to the ideal reference. Baseline to evaluate multi-amplifier solutions.

Scenario	Description
1-Amp	A single amplifier of width W_s
2-Amp Iso	Two amplifiers of width W_d
3-Amp Iso	Three amplifiers of width W_t
4-Amp Iso	Four amplifiers of width W_q
4-Amp Opt	Four amplifiers of different widths $\langle W_{0,1,2,3} \rangle$

Table 1: For iso-size, all amplifiers have the same width; for "opt," widths are independent. In each case, optimization is performed to find the width(s) that meet yield at minimum mean leakage power, across ranges of V_{dd} and variation.

amplifiers in combinations of 1 to 4. For four amplifiers, we consider two scenarios, one where all have the same size and another where they can be of different sizes. Transistors are sized so that 99% of the drivers simulated will meet the reference delay target with some combination of one or more amplifiers, while minimizing average leakage power.

3.3 Sizing Algorithms

We initially considered sizing based on geometric progression (e.g. $\langle W, \frac{1}{2}W, \frac{1}{4}W, \frac{1}{8}W \rangle$), where smaller amplifiers are enabled as necessary to boost weak larger ones. We also tried a yield-based strategy, where one amplifier is sized to account for 25% of yield, the second for the next 25%, and so forth. Both of these solutions turned out to be much worse than when all amplifiers were of the *same* size. This is because smaller drivers are disfavored due to being subject to more severe effects of process variation.

Optimization starts with large W and tries ever-smaller values in small steps. The search stops when the yield drops below 99%, but the width with the lowest power is kept, even if the yield is higher. This can happen because the search space is not monotonic, owing to process variation, which increases in severity for smaller channel area. For iso-size, the same width W is swept for all amplifiers.

For optimal sizing, amplifiers can have different widths. Each amplifier's size is swept in turn while holding all others constant. The cycle repeats until convergence. In general, optimal widths turn out very close to the same value, and power is not much less than for iso-size.

To size for variable voltage, we must ensure that yield is at least 99%, regardless of V_{dd}. Width W is swept from high to low just as above. For each W, voltages from 400mV to 1V are considered in steps of 100mV, where target delay D_R is still a function of V_{dd}. Yield is computed for each voltage, and search stops when *any* yield drops below 99%. The configuration with the lowest leakage is kept, even if the yield is higher than 99%, and leakage is computed as the geometric mean across all voltages.

4. EXPERIMENTAL METHODOLOGY

We use delay and leakage power formulas from Markovic [6]. Reference drivers are sized for small SRAM mat structures under ideal conditions, and target delay D_R is computed as the maximum of rise and fall times.

Yield is estimated by Monte Carlo simulation. Initially, 1000 sets of 8 gaussian-distributed random numbers $R_{i,j}$ are computed where σ=1 and μ=0. Mean V_{th} is 210mV, and the degree of process variation v is specified in terms of σ/μ. σV_{th} increases in proportion to the square root of the reduction in channel area [1]. For transistor i, j with variation v at nominal width W_R and scaled width W, threshold voltage is computed as $V_{\mathrm{th}}(i, j) = \mu + \mu v R_{i,j} \sqrt{W_R/W}$. Given V_{th} values for transistors in a driver, we compute power by summing over transistor leakage and delay from summing currents. Yield is the proportion of the population whose delay D is less than the reference D_R.

5. EXPERIMENTAL RESULTS

The widths shown are those necessary to meet 99% yield across the entire population of simulated drivers, and leakage power figures are averaged across the entire population, for the optimal combinations of amplifiers for each individual driver. Both supply voltage and severity of process variation have a substantial effect on both power and sizing requirements.

5.1 Sizing Single-Amplifier Drivers

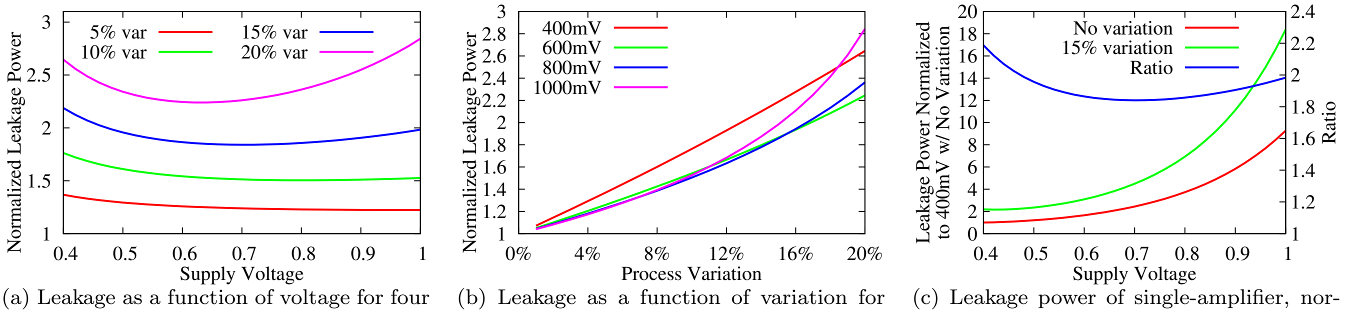
To establish a baseline, we consider how drivers must be sized for traditional solutions. A single amplifier width is chosen so that yield is met, given a particular supply voltage (V_{dd}) and severity of variation. Figure 2a shows how width must scale as a function of V_{dd}, at four different variation points; conversely, Figure 2b shows how width must scale as a function of variation, at four different voltage points.

As V_{dd} increases, required width reduces asymptotically. As V_{dd} and V_{th} diverge, the relative effect of V_{th} variation approaches zero. Required size is linear with variation because delay (in particular of the slowest driver) scales about linearly with V_{th}. Applying polynomial regression, width relative to reference is approximated by Equations 1 and 2.

$$\mathrm{slope} = -30.3V_{\mathrm{dd}}^3 + 77.61V_{\mathrm{dd}}^2 - 68.443V_{\mathrm{dd}} + 24.473 \quad (1)$$

$$W/W_R = \mathrm{slope} \cdot \mathrm{variation} + 1 \quad (2)$$

Figures 3a and 3b show corresponding graphs for leakage power (P_L). In Figure 3b, we see that at low V_{dd}, P_L scales nearly linearly with the degree of variation. This is because at low V_{dd}, the effect of V_{th} variation on delay is severe, requiring much wider drivers to compensate. By narrowing

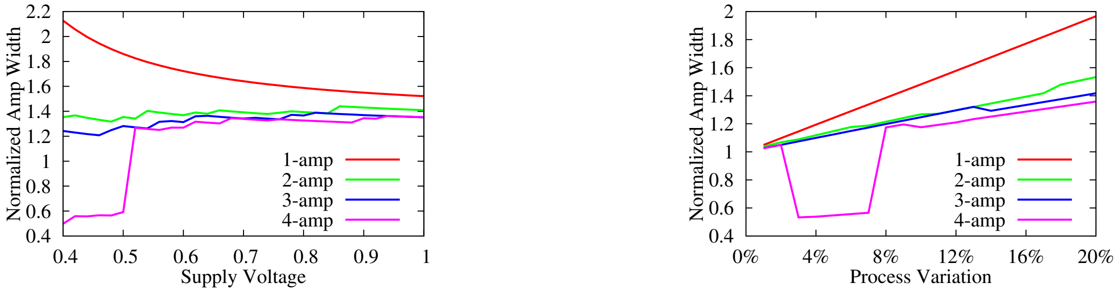


(a) Leakage as a function of voltage for four different degrees of process variation.

(b) Leakage as a function of variation for four different voltages.

(c) Leakage power of single-amplifier, normalized to the ideal reference at 400mV.

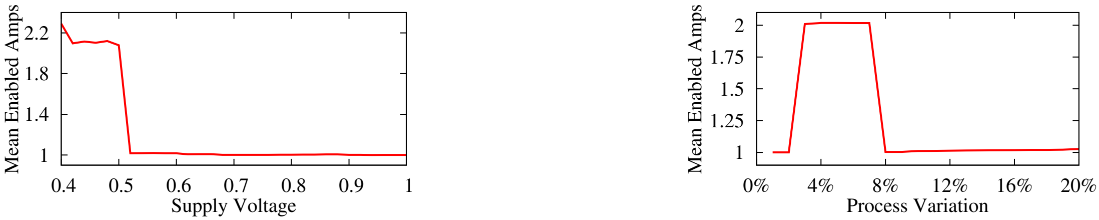
Figure 3: Single-amplifier leakage power, normalized to the ideal reference. Baseline to evaluate multi-amplifier solutions.



(a) Amp width as a function of voltage. Variation is fixed at 15%.

(b) Amp width as a function of variation. Voltage is fixed at 600mV.

Figure 4: Amp widths for multi-amplifier solutions, normalized to the ideal reference width. Amplifiers are iso-size, meaning that the width shown applies to all amplifiers in the driver.



(a) Variation is fixed at 15%.

(b) Voltage is fixed at 600mV.

Figure 5: Average number of amplifiers used simultaneously for iso-size quad-amplifier drivers.

σV_{th}, P_L scales about linearly with variation. At higher V_{dd}, we can use much smaller drivers, but that increases σV_{th}, and the P_L of very low V_{th} transistors dominates, because leakage is an exponential function of $(\textsc{dibl})V_{\mathrm{dd}} - V_{\mathrm{th}}$.

As for Figure 3a, the leakage power *relative to the reference* is functionally constant. Curvature occurs because leakage is a function of both V_{dd} and optimal driver width, which also changes with voltage. We clarify with Figure 3c, where curves are relative to the reference at fixed $V_{\mathrm{dd}} = 400\mathrm{mV}$. This more absolute graph shows that both reference and variation-affected leakage power increase with supply voltage, and variation-affected leakage is greater.

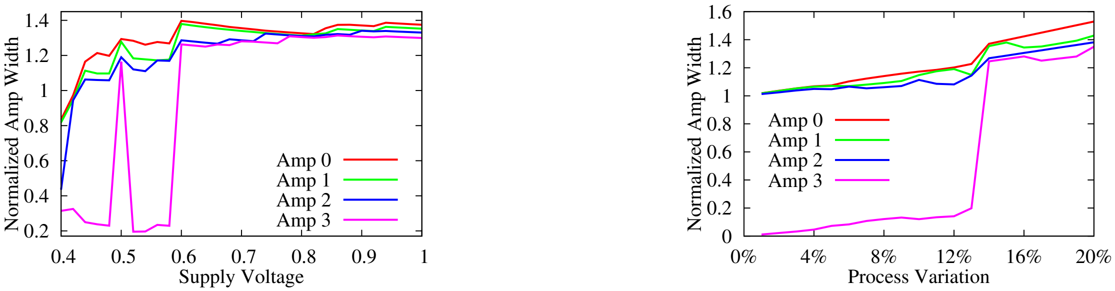
5.2 Sizing Multi-Amplifier Drivers

For drivers with 2 to 4 amplifiers, one width is selected for all amplifiers for each data point (iso-size). In Figure 4a variation is constant at 15%, and we sweep voltage. The 1-amp curve is the same as in Figure 2a. As amplifiers are added, width can be smaller, because there are more options, increasing the probability that one (or a combination) of the amplifiers can drive the signal line with sufficient speed.

The graph is jagged because amplifier subsets are not a smooth function of variation or V_{dd}. This is particularly pronounced for four amplifiers: At low V_{dd}, the optimizer finds it beneficial to use amplifiers in combination; at higher V_{dd}, it is typically more efficient to use only one, selecting which one meets timing at lowest power. This pattern corresponds to Figure 5a, which shows the average number of amplifiers used simultaneously, as a function of V_{dd}. When amplifiers are used in conjunction, their drive currents sum, allowing smaller amplifiers to meet the performance target.

Figure 4b complements the above, where V_{dd} is constant at 600mV, and we sweep degree of variation. The 1-amp curve is the same as in Figure 2b. Amplifiers are iso-size here as well. Corresponding to Figure 5b, we see that the quad-amplifier solution prefers to use multiple smaller amplifiers at lower variation. When variation is low, enabling multiple amplifiers has an averaging effect. At higher variation (particularly at low V_{dd}), delay variation in small amplifiers becomes too severe, and larger transistors are preferred.

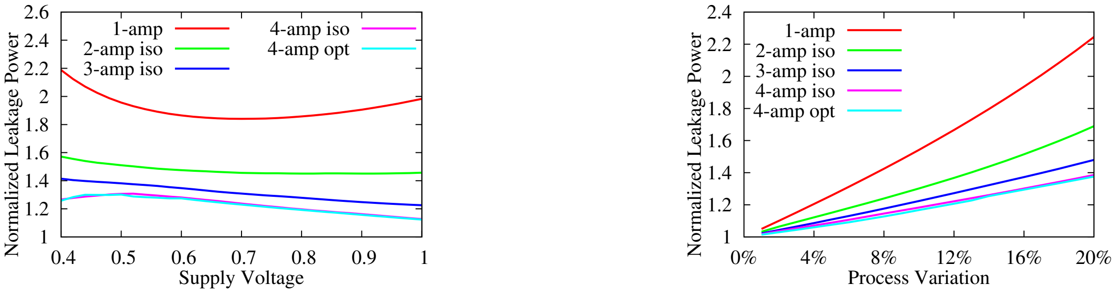
Whether the optimizer chooses small amplifier sizes to be used in combination or larger amplifier sizes to be used singly depends on the severity of variation and averaging effect of using multiple drivers. Ideally amplifier sizes are chosen so that N amplifiers out of M are enabled in the typical case, where $N < M$, allowing freedom to choose which subset of N meets timing at minimum power. For extremely slow amplifiers, N+1 can be enabled, while for extremely fast amplifiers



(a) Amp width as a function of voltage. Variation is fixed at 15%.

(b) Amp width as a function of variation. Voltage is fixed at 600mV.

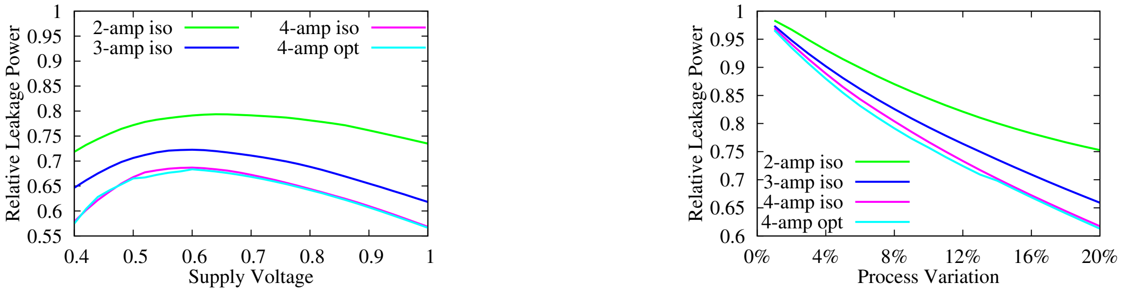
Figure 6: Amp widths for the independently-sized (optimal) four-amp solution, normalized to reference width. Each line indicates the transistor width of one of four amplifiers wired in combination to meet yield at minimum power.



(a) Leakage as a function of voltage. Variation is fixed at 15%.

(b) Leakage as a function of variation. Voltage is fixed at 600mV.

Figure 7: Leakage power for single- and multi-amplifier solutions, normalized to ideal reference. Lower is better.



(a) Leakage as a function of voltage. Variation is fixed at 15%.

(b) Leakage as a function of variation. Voltage is fixed at 600mV.

Figure 8: Leakage power for multi-amplifier solutions, relative to the single-amplifier solution. Lower is better.

$N-1$ may be selected. Delay variation (a function of size, voltage, and process variation) dictates the optimal N. As process variation increases or voltage is lowered, amplifiers must be increased in size to keep delay variation down. For $M = 4$, viable options for N are 1, 2, and 3. $N = 4$ is ruled out because there is no $N+1$ option for extremely slow amplifiers, and the amplifiers are too small, suffering from too much delay variation. $N=3$ is not chosen because the amplifiers are still too small; the variability is too high necessitating over-sized amplifiers to compensate. For low to moderate variation, $N = 2$ is chosen because amplifiers are large enough to minimize delay variation, there is a very high probability of finding two out of four that will meet timing at minimal leakage, and the penalty of enabling an additional amplifier is much less compared to $N=1$. With extremely high process variation or extremely low voltage, large amplifiers must be used to minimize delay variation, making $N=1$ the optimal choice.

5.3 Optimal Sizing of Quad-Amplifier Drivers

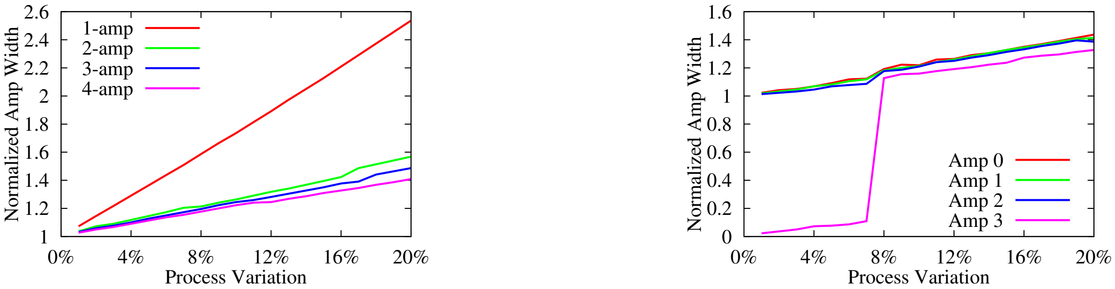
Section 3.3 describes the algorithm to select independent widths for quad-amplifier solutions. In Figures 6a and 6b, we see that generally three of the four amplifiers are the same size (about the same as for iso-size), while the fourth may be much smaller. (The instability is due to the granular search space and relatively small population size without perfectly even distribution.) At lower V_{dd}, the optimizer prefers the variation-averaging effects of using multiple amplifiers simultaneously, leading to the use of smaller amplifiers in general. As V_{dd} increases, the averaging effect becomes less productive, and it is preferred to use only one available amplifier, sometimes with support from a tiny fourth amplifier. As variation increases, delay variation in small amplifiers becomes too severe, and only larger ones are preferred.

	2-Amp Iso	3-Amp Iso	4-Amp Iso	4-Amp Optimal
Min P_L	0.72	0.62	0.57	0.56
Mean P_L	0.77	0.69	0.64	0.64
Max P_L	0.79	0.72	0.69	0.68

Table 2: Leakage power (P_L) across supply voltages for multi-amplifier solutions, relative to the single-amplifier baseline. Variation is 15%. Lower values are better.

5.4 Power Efficiency of Proposed Solutions

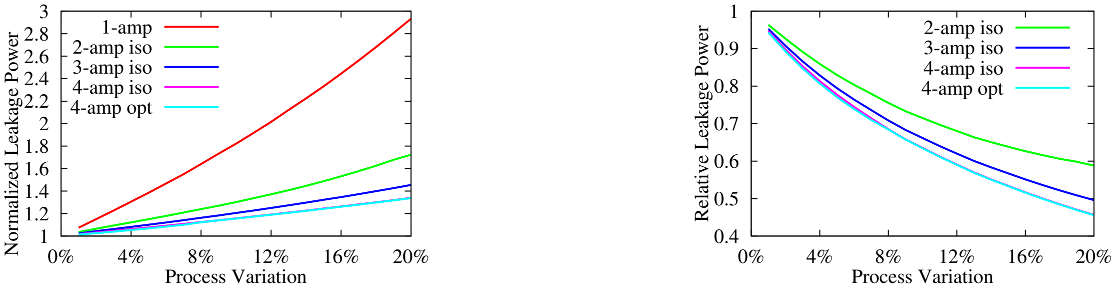
Figures 7a, 7b, 8a, and 8b and Tables 2 and 3 summarize the over-all leakage power (P_L) improvements provided by



(a) Multi-amplifier iso-size solutions

(b) Independently-sized (optimal) four-amp solution

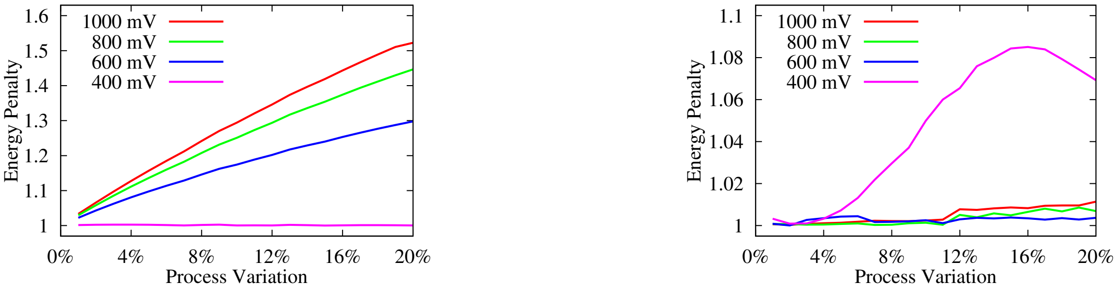
Figure 9: Amp width as a function of variation, normalized to the ideal reference width. *Optimized for all voltages.*



(a) Leakage normalized to the ideal reference.

(b) Leakage relative to the single-amplifier solution.

Figure 10: Leakage as a function of variation, for single- and multi-amplifier solutions *optimized for all voltages.*



(a) Single-amplifier

(b) Four-amplifier iso-size

Figure 11: Energy overhead of using amplifiers sized for variable voltage relative to amplifiers optimized for specific voltages.

	2-Amp Iso	3-Amp Iso	4-Amp Iso	4-Amp Optimal
$P_L(v)$	$0.98-1.21v$	$0.97-1.61v$	$0.96-1.80v$	$0.95-1.78v$

Table 3: Leakage power (P_L) as a function (by linear regression) of degree of process variation v for multi-amplifier driver solutions, relative to the single-amplifier baseline. $V_{\mathrm{dd}} = 600\mathrm{mV}$. Lower values are better.

the multi-amplifier solutions. Compared to the reference, relative P_L is effectively constant for this voltage range and nearly linear with respect to variation. Of particular note, the optimal quad-amplifier solution is only an incremental improvement over iso-size. While finding the best iso-size width is a quick optimization, finding the independently-sized optimal solution is much costlier and likely not worth the effort for such an insignificant gain. At 15% variation, four-amplifier solutions can reduce line driver P_L by an average of 36%. At 600mV, P_L is reduced by 18% at 10% variation, 27% at 15% variation, and 36% at 20% variation.

5.5 Sizing for Variable Voltage

Experiments above optimize drivers for specific voltages and degrees of variation. Here, we consider drivers sized to work well across the entire range of V_{dd}, where widths are chosen so that 99% yield is met regardless of voltage. Looking at Figure 4, we can infer that amplifier width will be heavily influenced by the yield at the bottom of the voltage range. Although one configuration of widths is chosen for each degree of variation, the specific combination of amplifiers enabled for each individual driver to meet timing is selected dynamically and may therefore vary with voltage.

Figure 9a (compare to 4b) shows how width must scale with the degree of process variation, for iso-size solutions. Widths scale about linearly with respect to variation. Figure 9b (compare to 6b) shows widths for the independently-sized (optimal) 4-amplifier solution. For low variation, three of the four amplifiers are about the same width, while a tiny fourth is sometimes used to provide a boost. At higher variation, the impact of variation on smaller amplifiers becomes to severe, and all amplifiers are chosen to be of similar size. The step in Figure 9b at 8% variation corresponds to a change in Figure 12; there we see that below 8% variation the tendency to use the tiny fourth amplifier increases with higher voltage. This occurs because widths are a compromise across voltages; the impact of voltage on the speed of ideal versus up-sized variation-affected amplifiers is slightly different, necessitating more boost (in terms of additional enabled amplifiers) at higher voltages to meet timing.

Figures 10a and 10b (compare to 7b and 8b) show how

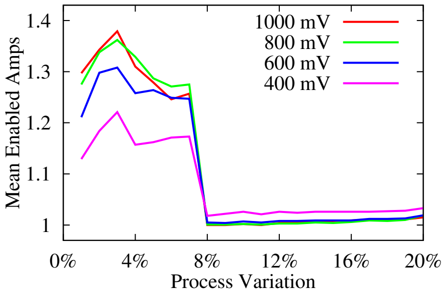


Figure 12: Average number of amplifiers used simultaneously for optimal-size quad-amplifier drivers. *Sizes optimized for all voltages.*

leakage power scales with variation, for all solutions. (Leakage in this figure is the geometric mean of leakage energy, Power × D_R, across all voltages.) Once again, the difference between the two four-amplifier solutions (iso-size and optimal) is negligible. Compared to a single-amplifier solution, using four amplifiers reduce leakage by 37% for 10% variation, 47% at 15% variation, and 54% at 20% variation.

Figure 11 shows the power overhead of using drivers sized for variable voltage. For each voltage and degree of variation, the figures plot the leakage power for the variable-voltage sizing divided by the leakage power for a driver optimized for the specific voltage and variation. For single-amplifier solutions (Figure 11a), the overhead is severe, more than 50% at 1V. On the other hand, the overhead is always less than 1% for four-amplifier solutions, which makes this a viable solution for circuits that support DVFS.

6. FUTURE WORK

There are several next steps we intend to explore. First, it is important to validate current findings using SPICE simulations. This will allow more in-depth analysis of switching characteristics. Each amplifier in a driver switches at a different speed, and this will have an impact on over-all rise and fall times as well as have the potential to cause glitches. We must also more fully characterize the effects of the enable transistors and consider multiple approaches to tri-stating.

The same basic techniques developed here can be applied to longer transmission lines. Besides compensating for variation, selectable combinations of amplifiers can be switched to compensate for noise and other sources of error. When errors occur, we can increase drive strength. This can be done incrementally, stepping up drive strength until errors are manageable. Further research will reveal how to directly compute appropriate drive strength as a function of signal noise, for better responsiveness than the incremental approach. SPICE analysis will also reveal whether amplifier combinations can be changed mid-transmission or if it is necessary to switch on packet boundaries.

Other open issues include area analysis (of driver and enable transistors), an approach to area optimization, and an analysis of multi-amplifier drivers in complete SRAM designs. And finally, we need to develop theory and implementations for control logic necessary for dynamic reconfiguration.

7. CONCLUSIONS

We have explored a new approach to mitigating process variation in line drivers, which are in the critical path and major consumers of power in many important circuits, particularly SRAMs. Severe process variation is a characteristic of deep submicron and low-voltage designs, requiring substantial size and voltage guard bands, which waste a great deal of energy in terms of leakage power. Our method builds drivers from multiple amplifiers, wired in parallel, where one or more is enabled as necessary to meet a delay target at a given process yield. This facilitates the use of smaller transistors that dissipate less leakage power. Our analysis explores sizing requirements across a space of supply voltages and severities of process variation, where we find as much as 36% leakage power savings over conventional designs with fixed-voltage. We also find that four-amplifier drivers can be built for variable voltage with little loss of efficiency and reduce leakage as much as 54% compared to a single-amplifier variable-voltage design.

8. REFERENCES

[1] A. Asenov. Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFETs: A 3-D "atomistic" simulation study. *Electron Devices, IEEE Transactions on*, 45(12):2505–2513, 1998.

[2] Z. Chishti, A. R. Alameldeen, C. Wilkerson, W. Wu, and S.-L. Lu. Improving cache lifetime reliability at ultra-low voltages. In *International Symposium on Microarchitecture (MICRO)*, December 2009.

[3] K. Ghose and M. B. Kamble. Reducing power in superscalar processor caches using subbanking, multiple line buffers and bit-line segmentation. In *Low Power Electronics and Design, 1999. Proceedings. 1999 International Symposium on*, pages 70–75. IEEE, 1999.

[4] H. Y. Hsiao, D. Bossen, and R. Chien. Orthogonal latin square codes. *"IBM Journal of Research and Development"*, 14(4):390–394, July 1970.

[5] R. Li, D. Zhou, J. Liu, and X. Zeng. Power-optimal simultaneous buffer insertion/sizing and wire sizing. In *Proceedings of the 2003 IEEE/ACM international conference on Computer-aided design*, page 581. IEEE Computer Society, 2003.

[6] D. Markovic, C. Wang, L. Alarcon, T.-T. Liu, and J. Rabaey. Ultralow-power design in near-threshold region. *"Proceedings of the IEEE"*, 98(2):237–252, February 2010.

[7] T. Miller, J. Dinan, R. Thomas, B. Adcock, and R. Teodorescu. Parichute: Generalized turbocode-based error correction for near-threshold caches. In *International Symposium on Microarchitecture (MICRO)*, 2010.

[8] N. Verma and A. P. Chandrakasan. A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy. *Solid-State Circuits, IEEE Journal of*, 43(1):141–149, 2008.

[9] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y. Ng, L. Wei, Y. Zhang, K. Zhang, and M. Bohr. A 4.0 GHz 291Mb voltage-scalable SRAM design in 32nm high-κ; metal-gate CMOS with integrated power management. In *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pages 456–457,457a, 2009.

[10] B.-D. Yang and L.-S. Kim. A low-power SRAM using hierarchical bit line and local sense amplifiers. *Solid-State Circuits, IEEE Journal of*, 40(6):1366–1376, 2005.