

Multi-Dimensional Circuit and Micro-Architecture Level Optimization

Zhenyu(Jerry) Qi[†], Matthew Ziegler[‡], Stephen V. Kosonocky[‡], Jan M. Rabaey[§], Mircea R. Stan[†]

[†] Charles L. Brown ECE Department, University of Virginia, Charlottesville, VA 22904, USA

[‡] IBM, T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

[§] Berkeley Wireless Research Center, University of California at Berkeley, Berkeley, CA 94704, USA

{jerry, mircea}@virginia.edu, {mziegler, stevekos}@us.ibm.com, jan@eecs.berkeley.edu

Abstract

This paper studies multi-dimensional optimization at both circuit and micro-architecture levels. By formulating and solving the optimization problem with conflicting design objectives and multiple tunable knobs, it is revealed that the 'sensitivity balance' strategy proposed in recent works for performance-energy optimization is a special case of a general multi-dimensional optimization framework. The results derived in this paper help the understanding of efficient trade-off among multiple design objectives with multiple knobs. The example of an industrial control logic implemented in PLA shows 22% energy saving and 70% area reduction at the expense of 4% delay increase.

1 Introduction

Integrated circuit (IC) design has seen major changes over the past few decades. For a very long time circuit speed and area are the only concerns in optimization [6, 4, 14]. As technology scales down, power consumption shows itself as a non-negligible, or even a major consideration that must be included across the design paradigm [2, 7]. The trade-off relationship between energy and delay is well known and the product of the two (*EDP*, or energy-delay product) is often adopted as a figure of merit. However, besides the question whether *ED* or *EDⁿ* ($n > 1$) is more appropriate, none of these provides insight for designers as how or whether designers can further improve the design, thus they cannot be directly used in optimization.

Energy performance optimization frameworks are proposed in recent works [10, 13, 1, 18, 17, 11]. Although they target different design levels and different knobs are chosen, their results share some very interesting similarity in that all get to the conclusion that optimal design points are reached when sensitivities for different knobs are 'balanced'. The intuition is that if the sensitivity of one knob is larger than others, it's always possible to improve the design by changing the knob with larger sensitivities, keeping energy/delay constant while reducing the other. Here 'knob' means a tunable circuit parameter. These works are reviewed in section 2.

A major drawback of all methods mentioned heretofore is that their limitation to two dimensional optimization, i.e., performance and energy trade-off. However circuit design

is constantly becoming more complex, with more design objectives and constraints, such as area, delay, energy, leakage power, throughput, etc. On the other hand, technology improvement provides designers with greater freedom by offering more knobs like multiple voltage supply domains, multiple threshold voltages, multiple clock islands, and even post-process tuning. Therefore two dimensional optimization is insufficient in fully exploring this convoluted design space. To answer this challenge, in this paper we will reveal the underlying theoretical principle which explains the similarity of existing works. The idea of 'sensitivity balance' is demystified when it is shown to be a special case of our multi-dimensional result, which provides both designers and CAD tools a general approach for multi-objective-multi-knob design. The same framework applies to both circuit and micro-architectural level optimization.

The rest of this paper is organized as follows. The next section provides some background on sensitivity balance in energy-performance optimization. Section 3 reveals the optimization principle in multi-dimensions and discusses its application in different scenarios. Section 4 demonstrates experimental results that validate our analysis. Potential pitfalls in sensitivity based optimization frameworks are pointed out in section 5 and section 6 concludes the paper.

2 Sensitivity Balance

Works in [10, 13, 1, 18, 17, 11] all target performance-energy optimization. However, instead of following the traditional method to look for the point with the minimum *EDP*(energy-delay product) or *EDⁿ* ($n > 1$), they define *optimalpoints* as those on the energy-delay *Pareto* curve. Points on the *Pareto* curve all have such property that improvement of one objective must entail deterioration of the other. *Pareto* curve allows designers to pick different optimal design points that meet with specified timing or energy budgets with the confidence that there's no room for improvement for the other objective.

The concept of *hardware intensity* η and *voltage intensity* θ are introduced in [17, 18]:

$$\eta = -\frac{D}{E} \frac{\partial E / \partial \eta}{\partial D / \partial \eta} \Big|_v, \quad \theta = -\frac{D}{E} \frac{\partial E / \partial v}{\partial D / \partial v} \Big|_\eta \quad (1)$$

By solving the problem of minimizing energy $E(\eta, \theta)$ subject to the constant delay constraint $E(\eta, \theta) = D$, the authors arrive at

$$\frac{\partial D/\partial \eta}{\partial E/\partial \eta} = \frac{\partial D/\partial v}{\partial E/\partial v} \quad (2)$$

Rearranging and plugging in (1), we end up with

$$\eta = \theta(v) \quad (3)$$

or hardware intensity should be equal to voltage intensity. This means at optimal design points, any marginal gain of energy (delay) by tuning knobs should result in a same amount of loss in delay (energy).

The authors in [10] defined the sensitivity of knob x as

$$S_x(X) = \frac{\partial E/\partial x}{\partial D/\partial x} \Big|_{x=X} \quad (4)$$

S_x represents the amount of energy that can be traded for delay by tuning variable x . Notice that while the objectives chosen in all those works are exactly the same, i.e., energy(E) and delay(D), the knobs are η , θ in [17, 18] and V_{dd} (supply voltages), V_{th} (threshold voltages) and W (transistor sizes) in [10, 13, 1] Yet all these works propose to balance sensitivities of all knobs.

The intuition behind sensitivity balance is that if the sensitivity of one knob is larger than others, it's always possible to improve the design by changing the knob with larger sensitivities, keeping energy/delay constant while reducing the other. A natural question arises as whether/how this solution changes in higher dimensions. Intuitively a Pareto curve evolves into a Pareto surface in three-dimensions and a hyper-surface in higher dimensions. The marginal gain and loss among objectives become complicated around optimal design points, and it's even more difficult for designers to tell how much room there exists to improve their designs. Simply trading off one objective for the other as in the two-dimensional case wouldn't work as efficiently. In the next section, we move into the multi-dimensions and show that eq. (3) is actually a special case of a much more general result.

3 Multi-dimensional Optimization

3.1 Typical Multi-dimensional Case

In multi-dimensional case, a typical generalized circuit or micro-architecture optimization problem can be formulated as follows:

$$\begin{aligned} \text{minimize:} & \quad Y_0(x_0, x_1, x_2, \dots, x_{n-1}) & (5) \\ \text{subject to:} & \quad Y_1(x_0, x_1, x_2, \dots, x_{n-1}) = C_1 \\ & \quad Y_2(x_0, x_1, x_2, \dots, x_{n-1}) = C_2 \\ & \quad \dots \quad \dots \\ & \quad Y_{n-1}(x_0, x_1, x_2, \dots, x_{n-1}) = C_{n-1} \end{aligned}$$

where Y_0 is the minimization objective, Y_1, Y_2, \dots, Y_{n-1} are objectives that either need to be minimized or meet certain budgets (C_1, C_2, \dots, C_{n-1}) to meet with. They can be delay, energy, area... etc. x_0, x_1, \dots, x_{n-1} are circuit or micro-architecture parameters that can be tuned by designers and treated as *knobs*, including but not limited to supply voltage V_{dd} , the threshold voltage V_{th} and transistor sizing W .

Micro-architecture level optimizations often involve discrete knobs such as pipeline stages, levels of parallelism, cache sizes and associativity numbers. A discrete variable called *architectural complexity* ξ is introduced in [17] for handling discrete knobs. However, rounding optimization results based on continuous knobs into integers can still give reasonably optimized results as long as the design space is

well shaped, which is usually true for circuits. For simplicity of discussion in this work we only consider circuit level knobs such as V_{dd} , V_{th} and W . Also in the above formulation the minimization objective Y_0 is arbitrarily chosen from all the objectives, and it will be clear later that all these objectives are equivalent no matter they serve as the minimization goal or constraints.

To solve the above problem in eq. (5), we can follow the Lagrange's Theorem [3] for optimization to introduce the Lagrange multipliers $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$ and solve the following equations:

$$\begin{aligned} \frac{\partial}{\partial x_0}(Y_0 + \lambda_1 Y_1 + \lambda_2 Y_2 + \dots + \lambda_{n-1} Y_{n-1}) &= 0 \\ \frac{\partial}{\partial x_1}(Y_0 + \lambda_1 Y_1 + \lambda_2 Y_2 + \dots + \lambda_{n-1} Y_{n-1}) &= 0 \\ &\dots \quad \dots \\ \frac{\partial}{\partial x_{n-1}}(Y_0 + \lambda_1 Y_1 + \lambda_2 Y_2 + \dots + \lambda_{n-1} Y_{n-1}) &= 0 \end{aligned}$$

Rearranging this equation array as a linear system of $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$

$$\begin{bmatrix} \frac{\partial Y_1}{\partial x_0} & \dots & \frac{\partial Y_1}{\partial x_0} & \dots & \frac{\partial Y_{n-1}}{\partial x_0} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{\partial Y_1}{\partial x_{n-1}} & \dots & \frac{\partial Y_1}{\partial x_{n-1}} & \dots & \frac{\partial Y_{n-1}}{\partial x_{n-1}} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \dots \\ \lambda_{n-1} \end{bmatrix} + \begin{bmatrix} \frac{\partial Y_0}{\partial x_0} \\ \dots \\ \frac{\partial Y_{n-1}}{\partial x_{n-1}} \end{bmatrix} = 0 \quad (6)$$

Notice that there are n equations with $(n-1)$ variables. If any nontrivial solution exists for $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$, the augmented coefficient matrix has to be linearly dependent:

$$\begin{vmatrix} \frac{\partial Y_0}{\partial x_0} & \dots & \frac{\partial Y_i}{\partial x_0} & \dots & \frac{\partial Y_{n-1}}{\partial x_0} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{\partial Y_0}{\partial x_{n-1}} & \dots & \frac{\partial Y_i}{\partial x_{n-1}} & \dots & \frac{\partial Y_{n-1}}{\partial x_{n-1}} \end{vmatrix} = 0 \quad (7)$$

This reveals the property of optimal points in the general multi-dimensional optimization scenario. Notice that in multi-dimensional cases, if one can balance all sensitivities of objectives with respect to all knobs, equation (7) still holds, but trying to balance the all these sensitivities as in two dimensional cases [10, 13, 1, 18, 17] is apparently an overkill. Equation (7) is an important and general result for multi-dimensional optimization for digital systems, since little assumption is made about the objectives and knobs. An important observation is that although Y_0 and Y_i ($i > 0$) are in the optimization goal and constraints respectively in the problem formulation (5), they are totally interchangeable in eq. (7). This is noticed previously in [10, 18, 4] too and obviously holds in multi-dimensions. Finally, in order to see how the multi-dimensional result reverts back to the two dimensional case, consider the following problem:

$$\text{minimize:} \quad E(V_{dd}, W) \quad (8)$$

$$\text{subject to:} \quad D(V_{dd}, W) = D_0, \quad D_0 \geq 0 \quad (9)$$

where E is energy, D is delay or latency. Simply applying eq. (7) we get

$$\begin{vmatrix} \frac{\partial E}{\partial V_{dd}} & \frac{\partial D}{\partial V_{dd}} \\ \frac{\partial E}{\partial W} & \frac{\partial D}{\partial W} \end{vmatrix} = 0 \quad \text{or,} \quad \frac{\frac{\partial E}{\partial V_{dd}}}{\frac{\partial E}{\partial W}} = \frac{\frac{\partial D}{\partial V_{dd}}}{\frac{\partial D}{\partial W}} \quad (10)$$

which is equivalent to the optimization scheme in previous work [10, 13, 1]. If hardware intensity η and voltage intensity θ are chosen as knobs, we get the exactly the same result as eq. (3) in [18, 17].

3.2 Extension into Different Number of Knobs and Objectives

In the problem formulation (5), the number of knobs equals the number of objectives. This is rarely true in practice. We can always have more knobs to tune with less objectives. For example, while both works in [10, 18] have energy and delay as objectives, there are three knobs V_{th} , V_{dd} and W in [10] and three knobs η , θ and ξ in [18]. For simplicity, here we only discuss the cases where the number of knobs equals the number of objectives plus one and similar conclusions can be drawn with even more knobs. In this case, the problem is changed a little bit:

$$\begin{aligned} \text{minimize:} \quad & Y_0(x_0, x_1, x_2, \dots, x_n) \\ \text{subject to:} \quad & Y_1(x_0, x_1, x_2, \dots, x_n) = C_1 \\ & Y_2(x_0, x_1, x_2, \dots, x_n) = C_2 \\ & \dots \\ & Y_{n-1}(x_0, x_1, x_2, \dots, x_n) = C_{n-1} \end{aligned} \quad (11)$$

where x_n is the extra knob. Following the same derivation in section 3.1, we need to solve

$$\begin{bmatrix} \frac{\partial Y_1}{\partial x_0} & \dots & \frac{\partial Y_1}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial Y_{n-1}}{\partial x_0} & \dots & \frac{\partial Y_{n-1}}{\partial x_n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \dots \\ \lambda_{n-1} \end{bmatrix} + \begin{bmatrix} \frac{\partial Y_0}{\partial x_0} \\ \dots \\ \frac{\partial Y_0}{\partial x_n} \end{bmatrix} = 0 \quad (12)$$

Now we end up with $(n + 1)$ equations with $(n - 1)$ variables. If any non-trivial solution exists, the coefficient matrix picked from any n rows in the augmented matrix has to be linearly dependent. Now let's consider the scenario in [10].

$$\begin{aligned} \text{minimize:} \quad & E(V_{dd}, W, V_{th}) \\ \text{subject to:} \quad & D(V_{dd}, W, V_{th}) = D_0, \quad D_0 \geq 0 \end{aligned} \quad (13)$$

Following the above discussion, we should have

$$\left| \begin{array}{cc} \frac{\partial E}{\partial V_{dd}} & \frac{\partial D}{\partial V_{dd}} \\ \frac{\partial E}{\partial W} & \frac{\partial D}{\partial W} \end{array} \right| = 0, \quad \text{and} \quad \left| \begin{array}{cc} \frac{\partial E}{\partial V_{th}} & \frac{\partial D}{\partial V_{th}} \\ \frac{\partial E}{\partial W} & \frac{\partial D}{\partial W} \end{array} \right| = 0 \quad (15)$$

Which leads to

$$\frac{\frac{\partial E}{\partial V_{dd}}}{\frac{\partial D}{\partial V_{dd}}} = \frac{\frac{\partial E}{\partial W}}{\frac{\partial D}{\partial W}}, \quad \text{and} \quad \frac{\frac{\partial E}{\partial V_{th}}}{\frac{\partial D}{\partial V_{th}}} = \frac{\frac{\partial E}{\partial W}}{\frac{\partial D}{\partial W}} \quad (16)$$

$$\text{or} \quad \frac{\frac{\partial E}{\partial V_{dd}}}{\frac{\partial D}{\partial V_{dd}}} = \frac{\frac{\partial E}{\partial V_{th}}}{\frac{\partial D}{\partial V_{th}}} = \frac{\frac{\partial E}{\partial W}}{\frac{\partial D}{\partial W}} \quad (17)$$

This explains why the principle of 'sensitivity balancing' is still valid in two dimensional cases with multiple knobs. However, with the increase of dimension we will have to watch for several equations like (7) before we can assert an optimal point.

On the other hand, if the objective number is larger than than the number of knobs, the optimization problem (5) is no longer valid since all knobs can be readily determined from those constraints.

3.3 Extension into Inequality Constraints

In the problem formulation (5), all constraints are equations. While this is useful when objective budgets are specified, or one wants to obtain particular points on the Pareto curve or surface. Oftentimes the constraints are inequalities which only specify upper or lower bounds of circuit parameters to be obtained. Since lower bounds can be converted

into upper bounds simply by multiplying minus one, we can assume all inequality constraints are upper bounds, and the optimization problem (5) becomes:

$$\begin{aligned} \text{minimize:} \quad & Y_0(x_0, x_1, x_2, \dots, x_{n-1}) \\ \text{subject to:} \quad & Y_1(x_0, x_1, x_2, \dots, x_{n-1}) = C_1 \\ & Y_2(x_0, x_1, x_2, \dots, x_{n-1}) = C_2 \\ & \dots \\ & Y_{m-1}(x_0, x_1, x_2, \dots, x_{n-1}) = C_{m-1} \\ & Y_m(x_0, x_1, x_2, \dots, x_{n-1}) \leq C_m \\ & \dots \\ & Y_{n-1}(x_0, x_1, x_2, \dots, x_{n-1}) \leq C_{n-1} \end{aligned} \quad (18)$$

The Karush-Kuhn-Tucker Theorem [3] can be readily applied in this kind of optimization problems. Due to the limited space here we won't elaborate the derivation here. The major idea is the same as shown in section 3.1. If the objectives are equal to or less than the number of knobs, conclusions very similar to (7) and (15) can be drawn. However, in cases with inequality constraints, it's still a legitimate optimization problem even when the number of constraints is larger than the number of knobs, in which case the optimization framework proposed here can no longer be used directly.

Finally, all the conditions derived from Lagrange or Karush-Kuhn-Tucker Theorem are necessary but not sufficient conditions. Moreover, local optimal points may be mistakenly identified as points on the Pareto curve/surface. However, it was observed that circuit performance are usually well behaved in feasible design space [12].

4 Experimental Results

In order to validate our results on general multi-dimensional optimization, we tested two circuits. In our experiments we try to choose various knobs to demonstrate the versatility of our multi-dimensional optimization result (7) as well as some special cases. Since the results are used to demonstrate the theory, we limit the dimension within three in order to illustrate the ideas by figures. While formula for power, delay and leakage are provided in [10], we follow [4] and choose simulation based sensitivity computation as a more general and accurate approach. Nonetheless, in practice designers can always apply available formula to get an initial point followed by a simulation based optimization. Good initial points guide optimization in feasible regions and facilitate faster convergence.

4.1 An Inverter Chain

The first circuit is a five stage inverter chain designed in Berkeley PTM 70nm technology shown in fig. 1. First we choose transistor sizing and the supply voltage as knobs and watch for energy-delay trade-off, which is a two dimensional case. However instead of tuning each transistor size as done in [1], we fix the ratio of $W2 : W3 = 1 : 4$ and use $W2$ as one knob, while keeping $W1, W4, W5$ fixed. This imitates the real case when the optimization is applied hierarchically - the low level parameters are either chosen by designers based on expertise or by transistor sizing tools, and the block is inserted into a higher level block for optimization, and so on. Fig. 2 shows the data points when

varying the two knobs. The color of each data point corresponds to the determinant in (7) at that point. The gray scale color is composed such that the darker color a point has, the smaller determinant that design point is associated with. Referring to eq. (10), smaller determinants indicate the sensitivities of the two knobs are more balanced, and vice versa. As discussed previously, the choice of knobs won't change the optimization framework.

By removing the fixed ratio of $W2$ and $W3$ and introducing area (A) as an extra objective, we move onto three-dimensional cases. By sweeping of all the three knobs respectively, we obtain a three dimensional plot showing the pattern of determinants, as in Fig. 3. Indeed this circuit is pretty friendly to optimization. A simple gradient based scheme should be able to converge onto the Pareto surface without much trouble.

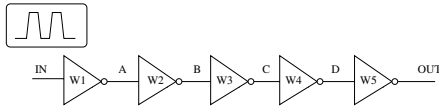


Figure 1. A five stage inverter chain

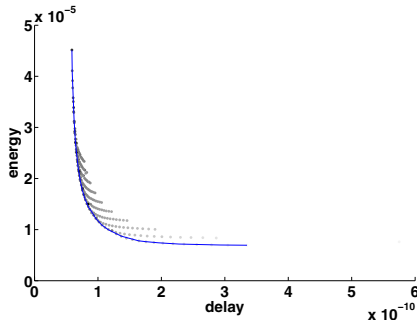


Figure 2. A two dimensional case: V_{dd} and W as knobs and E and D as objectives.

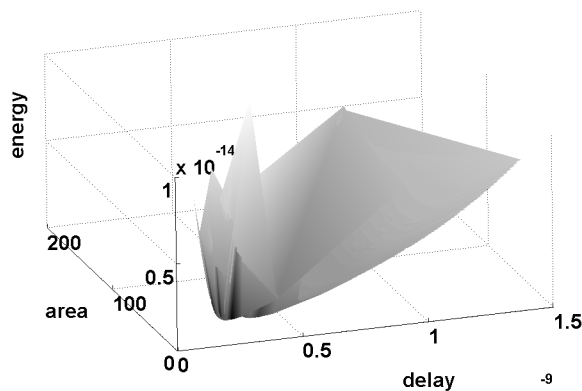


Figure 3. A three dimensional case: V_{dd} , $W2$ and $W3$ as knobs and E , D and A as objectives.

Now we examine a very common optimization case which turns out to be quite interesting. This time we again fix the ratio of $W2 : W3 = 1 : 4$, but introduce the thresh-

old voltage V_{th} as one knob. This three dimensional problem has W , V_{dd} and V_{th} as knobs and delay (D), energy (E) and area (A) as objectives. With this setting, the left hand side of eq. (7) becomes

$$\begin{vmatrix} \frac{\partial E}{\partial V_{dd}} & \frac{\partial D}{\partial V_{dd}} & 0 \\ \frac{\partial E}{\partial V_{th}} & \frac{\partial D}{\partial V_{th}} & 0 \\ \frac{\partial E}{\partial W} & \frac{\partial D}{\partial W} & 1 \end{vmatrix} \equiv \begin{vmatrix} \frac{\partial E}{\partial V_{dd}} & \frac{\partial D}{\partial V_{dd}} \\ \frac{\partial E}{\partial V_{th}} & \frac{\partial D}{\partial V_{th}} \end{vmatrix} \quad (19)$$

The reduction of matrix dimension comes from the fact that V_{dd} and V_{th} has no effect on area. Therefore from our optimization theory this three dimensional problem is equivalent to a two dimensional one. A direct translation of eq. (19) is: given a circuit that has achieved Pareto optimum in terms of delay and energy by only tuning V_{dd} and V_{th} , increasing or decreasing transistor sizing for all transistors by the same ratio wouldn't change its optimum. Thus this kind of orthogonization in the knobs can decouple the tuning process with multiple knobs and reduce the complexity of optimization. In reality there are always timing and power constraints on designs. For example, lowering V_{dd} results in delay increase which may call for transistor up-sizing, since the range of V_{th} tuning is usually very limited. However based on the Pareto surface property, once the timing requirement is met, no effort is needed on further tuning the knobs since there's no room for overall improvement. A visualization of the above analysis is provided in Fig. 4. The optimization in [9] is a more complicated example of the same type. There a two dimensional optimization is first performed with throughput and energy as objectives and V_{dd} and W as knobs, then the search goes in the third dimension A by tuning the third micro-architectural knob, folding and interleaving. By holding energy and delay constant the matrix is essentially decoupled, which makes this optimization process valid.

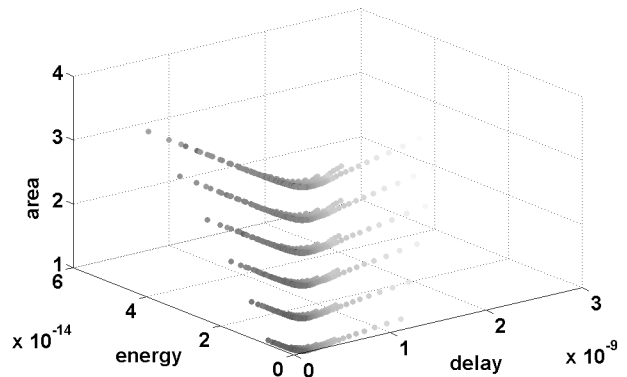


Figure 4. Another three dimensional case: V_{dd} , W and V_{th} are knobs and E , D and A are objectives.

4.2 A PLA Structured Control Logic

The second tested circuit is an industrial control logic with 25 primary inputs, 20 product terms and 1 output, implemented as a dynamic style programmable logic array (PLA) circuit with two PLA planes and a final static gate combining their outputs [15], designed in STMicroelectronics 90nm technology. Again we choose E , D and A as

objectives. The first knob is V_{dd} . The second knob is the NMOS size in the AND plane of the PLA, which directly affects the falling edge of the evaluation signal and is on the critical path. The last knob is the PMOS size of clock-block circuit which generates the OR plane clock by delaying the clock in the AND plane. This sizing is also important in that if the rising edge arrives too early, racing between the two planes may happen and the prechanged dynamic nodes in the OR plane are discharged by mistake. On the other hand a slow rising edge leads to unnecessary delay. The sweeping result is shown in fig. 5, where all the data points are normalized with respect to the smallest data point. Again, the determinants are smaller at better design points.

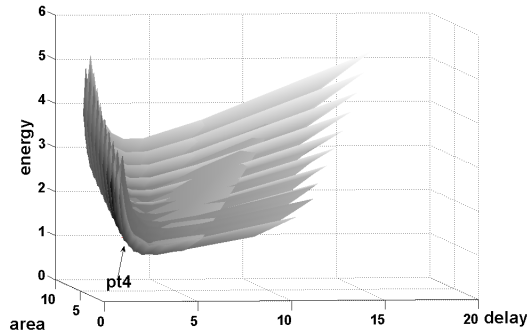


Figure 5. An industrial control logic circuit implemented as multiple dynamic PLA planes with static output stage.

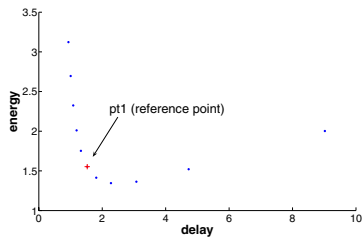


Figure 6. Points having similar area with pt1.

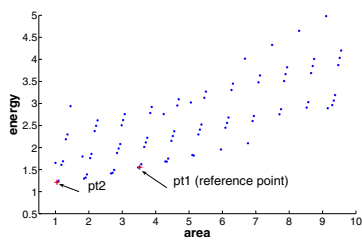


Figure 7. Points having similar delay with pt1.

In particular, we pick a reference design point (*pt1*) at the nominal voltage supply (one volt), and size transistors by 'rules of thumb'. The NMOS in the AND planes are sized about twice as large as the input drivers, and the PMOS for the clock generation circuit is intentionally upsized for a fast rising edge. For better comparison, design points with area, delay and energy within 7% of *pt1* are collected and plotted in Fig. 6, 7 and 8 respectively. The detailed design parameters of those corner points are listed in Table 1. The

reference point happens to fall on the optimum curve in the E-D trade-off in Fig. 6. However, it is suboptimal in the E-A and D-A plots as can be seen from Fig. 7 and 8. Table 1 shows that we can either improve 22% in energy and 70% in area with merely 4% increase on delay by moving to design point *pt2*, or improve 18% in delay and 70% in area with 6% penalty in energy by moving to design point *pt3*. *pt4* is a design point close to the origin picked directly in Fig. 5. It also has a small determinant shown in Table 1, confirming its optimality.

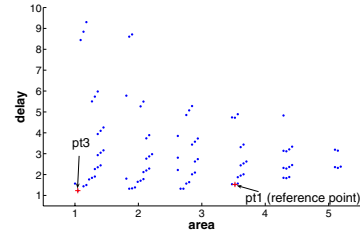


Figure 8. Points having similar energy with pt1.

Notice that while the D-E and D-A plots in Fig. 6 and 8 show trade-off between two conflicting figures of merit, we never have to trade between energy and area. This is reflected in the distribution of data points in Fig. 7. In such cases designers should simply pick the point closest to the origin, which is the absolute optimum. When a trade-off relationship exists, designers are constrained to design specifications. Lower dimensional optimization essentially yields projections from higher dimensional Pareto frontiers. For example, the delay-energy efficiency curves in [10, 17] are essentially the Pareto surface in Fig. 5 projected onto the XY plane. However, any optimization based on the projected curve is blind to the Z axis (in this case area), and is subject to violation of design specifications. In comparison, Fig. 6, 7 and 8 show the intersection of the volume, which are more valuable given a particular design requirement, and can only be obtained through a multi-dimensional optimization framework. With our results in this paper designers are able to infer the optimality in the context of multi-dimensions.

Table 1. Delay/energy/area comparison

| | knob1 | knob2 | knob3 | delay | energy | area | det |
|-----|-------|-------|-------|-------|--------|------|----------|
| pt1 | 1 | 3.5 | 4.3 | 1.5 | 1.55 | 3.52 | 1.41E-24 |
| pt2 | 1 | 1 | 2.7 | 1.56 | 1.21 | 1.05 | 6.87E-25 |
| pt3 | 1.2 | 1 | 2.7 | 1.23 | 1.65 | 1.05 | 6.89E-25 |
| pt4 | 1.1 | 1 | 2.7 | 1.37 | 1.40 | 1.05 | 6.70E-25 |

5 Potential Pitfalls in Optimization

Although it seems straightforward to apply eq. (7) to circuit designs, as works in [16, 8, 9, 5, 11] with sensitivity balance eq. (3), some special care should be taken in this kind of sensitivity based optimization.

First, in all figures 3, 4 and 5 our sweeping were not able to yield points with small determinants on the large-delay-low-energy side. This is due to the smallest transistor sizing limited by the technology. The large determinants indicate that although on the boundary of the design space, those points are not on the theoretical Pareto curve, where

determinants should diminish. This implies the failure of any optimizer whose convergence criterion solely relies on the magnitude of determinants (7).

Next, we project all the data points in Fig. 5 onto the D-E plane, which reverts back to the two dimensional energy-performance optimization, to get Fig. 9. Then we replace energy with power, repeat this step and end up with Fig. 10. An immediate observation is that for these two objectives the boundary design points (e.g., design points *pt5*, *pt6*, *pt7*) remain the same, which is intuitive as the energy at each data point is obtained by multiplying power and delay, but this may not be true in general, and the choice of objectives usually has a direct impact on the optimizing results.

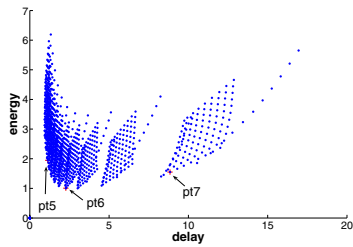


Figure 9. Delay-energy tradeoff for all points.

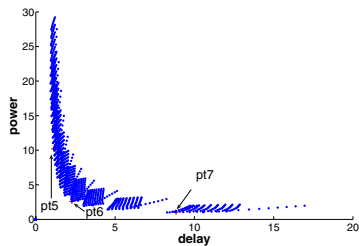


Figure 10. Delay-power tradeoff for all points.

Another observation is that the effective E-D trade-off region is smaller than the P-D trade-off region. For example, although *pt7* in Fig. 10 may be interesting to designers with loose delay constraints but stringent power budgets, it should never be chosen when energy supplants power as the figure of merit, since under these objectives *pt6* is always superior from Fig. 9.

Sometimes with certain knobs and objectives it is possible that the sensitivity values are positive when all sensitivities defined in eq. (4) are balanced. These points should *always be discarded* since they are similar to *pt7* in Fig. 9 in the sense that all objectives increase or decrease at the same time, thus the design can always be further improved. Although it may not be obvious from the view of sensitivity balancing, it's quite clear from our derivation in section 3.1 that this problem arises from the equality constraint in (5). In the two dimensional case (8), sensitivity balance only yields a minimized E under a certain D , without knowing whether a trade-off exists. Furthermore, in the inequality constraint case discussed in section 3.3, the KKT theorem is only necessary but not sufficient. Therefore, designers should always check negativity of sensitivity values or impose it as a constraint in sensitivity balance. Nonetheless,

these points are usually quite far from the effective trade-off region and design points obtained from experience, good rules of thumbs or CAD tools are usually close to the 'knee' area of Pareto curves. Starting from there would guide the optimization in the correct direction.

6 Conclusion

In this work we revealed the underlying principle of sensitivity balance and explained the similarity of research works from different groups on energy-performance optimization. It's shown that the metric of 'sensitivity balance' for optimum is only valid in two dimensions. A general result that applies to arbitrary dimensions is derived with its extension in different scenarios discussed. Applying the proposed optimization theory, designers can introduce any number of objectives that they care about and any number of knobs that are available for tuning. This result not only helps designers to evaluate the optimality of different design points, but also provides a guideline for CAD tools targeting circuit optimization. Finally we demonstrated the optimization theory with two test circuits in various situations. Some potential pitfalls in such sensitivity based optimization frameworks are pointed out. Due to its flexibility, this optimization framework can be adapted for micro-architectural level and analog circuits with appropriate modifications.

References

- [1] R. Brodersen, M. Horowitz, D. Markovic, and B. N. V. Stojanovic. Simulation and optimization of the power distribution network in VLSI circuits. In *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, pages 35–42, 2002.
- [2] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen. Low-power cmos digital design. *IEEE J. Solid-State Circuits*, 27(4):473–484, April 1992.
- [3] E. K. P. Chong and S. H. Zak. *An Introduction to Optimization*. Wiley-Interscience, New York, 2 edition, 2001.
- [4] A. R. Conn, I. M. Elfadel, W. W. Molzen, P. O'Brien, P. Strenski, C. Visweswariah, and C. Whan. Gradient-based optimization of custom circuits using astatic-timing formulation. In *Proc. Design Automation Conf. (DAC)*, pages 452–459, 1999.
- [5] S. Farhana, L. Melinda, and N. Borivoje. Hierarchical power-performance optimization of digital filters for universal radio. In *SRC Student Symposium*, 2006.
- [6] J. P. Fishburn and A. E. Dunlop. Tilos: a posynomial programming approach to transistor sizing. In *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, pages 326–328, 1985.
- [7] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein. Scaling, power, and the future of cmos. In *IEEE International Electron Devices Meeting*, page keynote talk, 2005.
- [8] S. Kao, R. Zlatanovici, and B. Nikolic. A 250ps 64-bit carry-lookahead adder in 90nm cmos. In *Proceedings of the International Solid-State Circuits Conference (ISSCC)*, pages 438–439, 2006.
- [9] D. Markovic, R. Brodersen, and B. Nikolic. A 70gops, 34mw multi-carrier mimo chip in 3.5mm2. In *Symposium on VLSI Circuits*, pages 15–17, 2006.
- [10] D. Markovic, V. Stojanovic, B. Nikolic, M. A. Horowitz, and R. W. Brodersen. Methods for true energy-performance optimization. *IEEE J. Solid-State Circuits*, 39(8):1282–1292, August 2004.
- [11] H. H. Peter. Power-constrained microprocessor design. In *Proc. IEEE Int. Conf. on Computer Design (ICCD)*, pages 14–16, 2002.
- [12] G. Stehr, H. Graeb, and K. Antreich. Performance trade-off analysis of analog circuits by normal-boundary intersection. In *Proc. Design Automation Conf. (DAC)*, pages 958–963, 2003.
- [13] V. Stojanovic, D. Markovic, B. Nikolic, M. Horowitz, and R. Brodersen. Energy-delay tradeoffs in combinational logic using gate sizing and supply voltage optimization. In *Proceedings of the European Solid-State Circuits Conference (ESSCC)*, pages 211–214, 2002.
- [14] I. Sutherland, B. Sproull, and D. Harris. *Logical effort: designing fast CMOS circuits*. San Francisco, CA: Morgan Kaufmann, 1999.
- [15] N. Weste and D. Harris. *CMOS VLSI design: a circuits and systems perspective, 3rd. edition*. Addison Wesley, 2005.
- [16] R. Zlatanovici and B. Nikolic. Power-performance optimization for custom digital circuits. In *Power And Timing Modeling Optimization and Simulation (PATMOS)*, pages 404–414, 2005.
- [17] V. Zyuban and P. N. Strenski. Unified methodology for resolving power-performance tradeoffs at the microarchitectural and circuit levels. In *Proc. Int. Symp. on Low Power Electronics and Design (ISLPED)*, pages 166–171, 2002.
- [18] V. Zyuban and P. N. Strenski. Balancing hardware intensity in microprocessor pipelines. *IBM J. Res. & Dev.*, 47:585–598, 2003.