Statistical Power Minimization for FPGA Dual-Vdd Interconnect with Time Slack Allocation

1. TIME SLACK ALLOCATION WITH WIRE SEGMENTS OF DIFFERENT LENGTHS

we use the net-based formulation which partitions the constraints on path delay into constraints on delay across circuit elements or routing. Let a(v) be the arrival time for vertex v in \mathcal{G} and the timing constraints for PI and PO are as below,

$$a(v) \le T_{spec} \quad \forall v \in \mathcal{PO}$$
 (1)

$$a(v) = 0 \qquad \forall v \in \mathcal{PI} \tag{2}$$

For the edges corresponding to routing in \mathcal{G} , the constraints considering slack can be expressed as

$$a(p_{i0}) + d(p_{i0}, p_{ik}) + S_{ik} \le a(p_{ik})$$

$$0 \le i < N_r \land \forall p_{ik} \in \mathcal{FO}_{p_{i0}}$$
(3)

where vertex p_{i0} is the source of \mathcal{R}_i in \mathcal{G} , vertex p_{ik} is k^{th} sink of \mathcal{R}_i in \mathcal{G} , S_{ik} is the slack allocated to k^{th} sink in \mathcal{R}_i and $d(p_{i0}, p_{ik})$ is the delay from p_{i0} to p_{ik} in \mathcal{R}_i using VddH. For the edges other than routing in \mathcal{G} , the constraints can be expressed as

$$a(u) + d(u, v) \le a(v) \qquad \forall u \in \mathcal{V} \land u \notin SRC \land v \in \mathcal{FO}_u$$
 (4)

where SRC is a subset of V and gives the set of vertices corresponding to routing tree sources.

There is an upper bound for slack, which is the delay increase when VddL is assigned to all the switches in a tree. Clearly, slack more than the upper bound cannot lead to more VddL switches. We define the *useful slack* of each routing tree sink as the slack less than this upper bound. For the rest part of the paper, we use slack to represent the useful slack. The slack upper bound constraints can be expressed as

$$0 \le S_{ik} \le D_{ik} \qquad 0 \le i < N_r \land 1 \le k \le N_k(i) \qquad (5)$$

where $N_k(i)$ is the number of sinks in \mathcal{R}_i and D_{ik} is the delay increase of the path from the source to k^{th} sink in \mathcal{R}_i when VddL is assigned to all the switches in that path.

Given a routing tree with arbitrary topology and allocated slack for each sink, we need to estimate power reduction that can be achieved. We use l_{ik} represent the number of switches in the path from the source to k^{th} sink in \mathcal{R}_i . We first transform slack S_{ik} into s_{ik} , which is expressed in number of switches as follows,

$$s_{ik} = \frac{S_{ik}}{D_{ik}} \cdot l_{ik} \tag{6}$$

We then estimate the VddL switch number that can be achieved using s_{ik} . We use C_{ik} to represent the total load capacitance of the switches in the path from the source to k^{th} sink in \mathcal{R}_i . We use c_{ij} to represent the load capacitance of j^{th} switch in \mathcal{R}_i . We define sink list $S\mathcal{L}_{ij}$ as the set of sinks in the fanout cone of j^{th} switch in \mathcal{R}_i . We then estimate the number of VddL switches that can be achieved given the allocated slack as

$$F_n(i) = \sum_{j=0}^{N_s(i)-1} \min(\frac{s_{ik}}{C_{ik}} \cdot c_{ij} : \forall k \in \mathcal{SL}_{ij})$$
(7)

To estimate the number of VddL switches that can be achieved in tree \mathcal{R}_i , we first deliberately distribute the slack s_{ik} to the switches in the path from source to k^{th} sink in \mathcal{R}_i considering the load capacitance of each switch. For a switch with multiple sinks in its fanout cone, we choose the minimum $s_{ik}c_{ij}/C_{ik}$ as the slack distributed to the switch. We then add the slack distributed to all the switches in \mathcal{R}_i . and get the estimated number of VddL switches. The rationale is that we consider k^{th} sink with minimum $s_{ik}c_{ij}/C_{ik}$ in sink list \mathcal{SL}_{ij} as the most critical sink to j^{th} switch in \mathcal{R}_i .

We then estimate the power reduction of \mathcal{R}_i . The dynamic power reduction of the tree \mathcal{R}_i is estimated as the sum of the dynamic power reduction of each switch in \mathcal{R}_i and can be expressed as,

$$P_{dr}(i) = 0.5 f_{clk} \cdot \Delta V dd^2 \cdot f_s(i) \sum_{j=0}^{N_s(i)-1} \left[min(\frac{s_{ik}}{C_{ik}} \cdot c_{ij} : \forall k \in \mathcal{SL}_{ij}) \cdot c_{ij} \right]$$

$$\tag{8}$$

The leakage power reduction of \mathcal{R}_i is also the sum of the leakage power reduction of each switch in \mathcal{R}_i and can be expressed as,

$$P_{lr}(i) = \sum_{j=0}^{N_s(i)-1} \left[min\left(\frac{s_{ik}}{C_{ik}} \cdot c_{ij} : \forall k \in \mathcal{SL}_{ij} \right) \cdot \Delta P_s(i,j) \right]$$
(9)

where $\Delta P_s(i, j)$ is the leakage power difference of j^{th} switch in \mathcal{R}_i between VddH and VddL. Wire segments with different lengths might be driven by switches with different sizes. The objective is to maximize power reduction given by the sum of (8) and (9). To incorporate (8) and (9) into mathematical programming, we introduce a variable $f_n(i, j)$ for j^{th} switch in \mathcal{R}_i and some additional constraints. The new objective function after transformation plus the *additional constraints* can be expressed as

$$Maximize \sum_{i=0}^{N_r-1} 0.5 f_{clk} \Delta V dd^2 \sum_{j=0}^{N_s(i)-1} f_n(i,j) f_s(i,j) c_{ij} + \sum_{i=0}^{N_r-1} \sum_{j=0}^{N_s(i)-1} f_n(i,j) \Delta P_s(i,j)$$
(10)

s.t.

$$f_n(i,j) \le \frac{s_{ik}}{C_{ik}} c_{ij} \quad 0 \le i < N_r \land 0 \le j < N_s(i) \land \forall k \in \mathcal{SL}(1)$$

We formulate the *time slack allocation problem* using objective function (10), additional constraints (11), slack upper bound constraints (5), plus timing constraints (1), (2), (3) and (4). It is

easy to verify that $(1) \sim (5)$ and (11) are linear, and the objective function (10) is linear too. Hence we have the following theorem.

THEOREM 1. The time slack allocation problem is a linear programming (LP) problem.

2. STATISTICAL DELAY AND POWER MODELS

We consider the variation in threshold voltage (V_{th}) , effective channel length (L_{eff}) , and gate oxide thickness (T_{ox}) . Similar to [1] where ASIC is assumed, each variation (ΔP) is decomposed into global (die-to-die) variation (ΔP_g) and local (within-die) variation (ΔP_l) . Both global and local variations are modeled as normal random variables. To make presentation simple, we denote ΔL_{eff} , ΔV_{th} and ΔT_{ox} as L, V and T, respectively. L, V and T can be decomposed into local (L_l, V_l, T_l) and global (L_g, V_g, T_g) components.

Power reduction of an interconnect switch by changing the Vddlevel from VddH to VddL considering process variation is presented as below,

$$\Delta P = \Delta P_0 e^{-c_1 L} e^{-c_2 V} e^{-c_3 T} \tag{12}$$

$$\Delta E = \Delta E_0 + \frac{\partial \Delta E}{\partial V} V = \Delta E_0 + c_4 V \tag{13}$$

 ΔP is leakage power reduction, ΔE is dynamic energy per signal switch reduction when driving unit load capacitance. c_1, c_2, c_3 and c_4 are curve fit parameters. ΔP_0 and ΔE_0 are the nominal values. ΔP exponentially depends on L, V and T. Under this model, ΔP is a lognormal (LN) random variable as follows,

$$\Delta P \sim LN(ln(\Delta P_0), (c_1^2 \sigma_L^2 + c_2^2 \sigma_V^2 + c_3^2 \sigma_T^2))$$

We assume a first-order Taylor expansion of dynamic energy reduction ΔE is adequate, which therefore can be modeled as a normal random variable as follows,

$$\Delta E \sim N(\Delta E_0, c_4^2 \sigma_V^2)$$

A first-order Taylor expansion of circuit element delay is also assumed as adequate,

$$d = d_0 + \frac{\partial d}{\partial L}\Delta L + \frac{\partial d}{\partial V}\Delta V = d_0 + c_5 L + c_6 V$$
(14)

where d_0 is the nominal value of circuit element delay, c_4 and c_5 are curve fit parameters. Delay d is a normal random variable as follows under this model,

$$\Delta d \sim N(\Delta d_0, c_5^2 \sigma_L^2 + c_6^2 \sigma_V^2)$$

3. STATISTICAL POWER OPTIMIZATION

3.1 Statistical Chip-level Time Slack Allocation

To handle the variability of process parameters, we reformulate the time slack allocation problem as a robust linear program problem which can be solved efficiently using interior-point convex methods [2]. We first the objective function (10) into an equivalent formulation as below, which places the power reduction into the constraint set

$$Miminize \sum_{i=0}^{N_r - 1} \sum_{k=0}^{N_k(i) - 1} S_{ik}$$
(15)

s.t.

$$\sum_{i=0}^{N_r-1} \sum_{j=0}^{N_s(i)-1} \Delta P_d(i,j) f_n(i,j) \ge \Delta \hat{P}_{dyn}$$
(16)
$$\Delta P_d(i,j) = c_{ij} f_{clk} f_s(i,j) \Delta E(i,j)$$

$$\sum_{i=0}^{N_r-1} \sum_{j=0}^{N_s(i)-1} \Delta P_s(i,j) f_n(i,j) \ge \Delta \hat{P}_{leak}$$
(17)

where $\Delta E(i, j)$ is the dynamic energy per signal switch reduction of switch j in \mathcal{R}_i when driving unit capacitance, $\Delta \hat{P}_{dyn}$ and $\Delta \hat{P}_{leak}$ are the optimal dynamic and leakage power reduction achieved by the time slack allocation problem without considering process variation. The statistical equivalents of the power reduction constraints (16) and (17) can be expressed as,

$$P(\sum_{i=0}^{N_r-1}\sum_{j=0}^{N_s(i)-1}\Delta P_d(i,j)f_n(i,j) \ge \Delta \hat{P}_{dyn}) \ge \eta$$
(18)

$$P(\sum_{i=0}^{N_r-1}\sum_{j=0}^{N_s(i)-1}\Delta P_s(i,j)f_n(i,j) \ge \Delta \hat{P}_{leak}) \ge \eta$$
(19)

We first handle the probabilistic dynamic power reduction constraint (18). Let $x = \sum_{i,j} \Delta P_d(i,j) f_n(i,j)$, where

$$\Delta P_d(i,j) = c_{ij} f_{clk} f_s(i,j) [\Delta E_0(i,j) + c_4(i,j)V]$$

x is the sum of random normal variables and can be modeled as another normal random variable. Considering the fact that all the switches on chip share the same V_g , the mean μ_x and variance σ_x^2 of x can be expressed as,

$$\mu_x = f_{clk} \sum_{i,j} [c_{ij} f_s(i,j) \Delta E_0(i,j) f_n(i,j)]$$
(20)

$$\sigma_x^2 = f_{clk}^2 [\sum_{i,j} c_{ij} f_s(i,j) f_n(i,j)]^2 \sigma_{V_g}^2 + f_{clk}^2 \sum_{i,j} [c_{ij} f_s(i,j) f_n(i,j)]^2 \sigma_{V_l}^2$$
(21)

Using the translation-invariance property of a normal distribution variable, We can express (18) as

$$P(\frac{-x+\mu_x}{\sigma_x} \le \frac{-\Delta \vec{P}_{dyn} + \mu_x}{\sigma_x}) \ge \eta \tag{22}$$

Since $(-x+\mu_x)/\sigma_x \sim N(0,1)$, $P(x \geq \Delta \hat{P}_{dyn}) \geq \eta$ is $\phi(-\Delta \hat{P}_{dyn}+\mu_x)/\sigma_x) \geq \eta$ where $\phi(\cdot)$ is the *cdf* of N(0,1). We can finally express the dynamic power constraint (18) as follows,

$$\mu_x - \phi^{-1}(\eta)\sigma_x \ge \Delta \hat{P}_{dyn} \tag{23}$$

Similarly, for leakage power reduction constraint, we let $y = \sum_{i,j} \Delta P_s(i,j) f_n(i,j)$, where

$$\Delta P_s(i,j) = \Delta P_0 e^{-c_1(i,j)L - c_2(i,j)V - c_3(i,j)T}$$

y is the sum of lognormal variables and can be modeled as another lognormal random variable [3]. Considering the fact that all the switches share the same global variances, we can calculate the mean μ_y and variance σ_y^2 of y. We first define:

$$g_0(c_1, c_2, c_3) = c_1^2(\sigma_{L_g}^2 + \sigma_{L_l}^2) + c_2^2(\sigma_{V_g}^2 + \sigma_{V_l}^2) + c_3^2(\sigma_{T_g}^2 + \sigma_{T_l}^2)$$

$$A_{ij} = \Delta P_s(i, j) f_n(i, j)$$

 μ_y and σ_y^2 can then be expressed as follows,

$$\mu_{y} = \sum_{i,j} \Delta P_{0}(i,j) f_{n}(i,j) e^{\frac{g_{0}(c_{1}(i,j),c_{2}(i,j),c_{3}(i,j))}{2}} (24)$$

$$\sigma_{y}^{2} = \sum_{i,j} \{ [(\Delta P_{0}(i,j) f_{n}(i,j))^{2} e^{g_{0}(c_{1}(i,j),c_{2}(i,j),c_{3}(i,j))}]$$

$$\cdot [e^{g_{0}(c_{1}(i,j),c_{2}(i,j),c_{3}(i,j)} - 1] \}$$

$$+ \sum_{i,j} \sum_{i',j'} 2COV(A_{ij}, A_{i'j'}) (25)$$

where μ_y is the sum of means of A_{ij} and σ_y^2 is the sum of variance of A_{ij} and the covariance of each pair of A_{ij} . The covariance is calculated as follows,

$$COV(A_{ij}, A_{i'j'}) = E[A_{ij}A_{i'j'}] - E[A_{ij}]E[A_{i'j'}] (26)$$

$$E[A_{ij}A_{i'j'}] = (\Delta P_0(i, j)f_n(i, j)) \cdot (\Delta P_0(i', j')f_n(i', j'))$$

$$\cdot e^{\frac{g_0(c_1(i, j) + c_1(i', j'), c_2(i, j) + c_2(i', j'), c_3(i, j) + c_3(i', j'))}{2}}$$

$$E[A_{ij}] = \Delta P_0(i, j)f_n(i, j)e^{\frac{g_0(c_1(i, j), c_2(i, j), c_3(i, j))}{2}}$$

We then use the method from [1] to obtain the mean and variance $(\mu_{N,y}, \sigma_{N,y}^2)$ of the normal random variable corresponding to the lognormal variable y.

$$\mu_{N,y} = \ln(\mu_y^2 / \sqrt{\mu_y^2 + \sigma_y^2}), \sigma_{N,y}^2 = \ln(1 + \sigma_y^2 / \mu_y^2)$$
(27)

Similar to the dynamic power constraint (23), we can express the leakage power constraint (19) as

$$\mu_{N,y} - \phi^{-1}(\eta)\sigma_{N,y} \ge \ln(\Delta P_{leak})$$

Based on (27), we can finally express (19) as

$$\ln(\mu_y^2/\sqrt{\mu_y^2 + \sigma_y^2}) - \phi^{-1}(\eta)\sqrt{\ln(1 + \sigma_y^2/\mu_y^2)} \ge \ln(\Delta \hat{P_{leak}})$$
(28)

Similar to [4], we perform a least square error fitting for (28) and get a linear function of μ_y and σ_y as follows,

$$(a_1 - a_2 \phi^{-1}(\eta))\mu_y + (a_3 - a_4 \phi^{-1}(\eta))\sigma_y \ge \ln(\Delta \hat{P}_{leak})$$
(29)

Considering the uncertainty of timing constraints, the deterministic timing constraints (1) can be transformed into the probabilistic constraints as follows,

$$P(a(v) \le T_{spec}) \ge \zeta \quad \forall v \in \mathcal{PO}$$

$$(30)$$

where ζ is the timing yield. Similar to [5, 4], we use a heuristic method to model the circuit element delays as $d(u,v)^0 + \phi^{-1}(\zeta)\sigma_{d(u,v)^0}$, where $d(u,v)^0$ is the nominal delay value of edge e(u, v) and $\sigma_{d(u,v)^0}$ is the standard deviation of the circuit element delay. We then can tranform (30) into closed form analytically. For PI and PO, the constraints stay the same as (2) and (1), respectively. For the edges corresponding to routing in \mathcal{G} , the probabilistic constraints considering slack can be expressed as

$$a(p_{i0}) + d(p_{i0}, p_{ik})^{0} + \sigma_{d(p_{i0}, p_{ik})^{0}} + S_{ik} \leq a(p_{ik})$$

$$0 \leq i < N_{r} \land \forall p_{ik} \in \mathcal{FO}_{p_{i0}} (31)$$

$$\sigma_{d(p_{i0}, p_{ik})^{0}}^{2} = \left[\sum_{\forall b_{ij} \in \mathcal{L}_{ik}} c_{5}(i, j)\right]^{2} \sigma_{L_{g}}^{2} + \sum_{\forall b_{ij} \in \mathcal{L}_{ik}} c_{5}(i, j)^{2} \sigma_{L_{l}}^{2}$$

$$+ \left[\sum_{\forall b_{ij} \in \mathcal{L}_{ik}} c_{6}(i, j)\right]^{2} \sigma_{V_{g}}^{2} + \sum_{\forall b_{ij} \in \mathcal{L}_{ik}} c_{6}(i, j)^{2} \sigma_{V_{l}}^{2}$$

where b_{ij} is j^{th} switch in \mathcal{R}_i , \mathcal{L}_{ik} is the path from p_{i0} to p_{ik} and $\sigma_{d(p_{i0}, p_{ik})^0}$ is the standard deviation of routing path delay $d(p_{i0}, p_{ik})^0$ considering local and global variations. For the edges other than routing in \mathcal{G} , the probabilistic timing constraints can be expressed as

$$a(u) + d(u,v)^{0} + \sigma_{d(u,v)^{0}} \leq a(v)$$

$$\forall u \in \mathcal{V} \land u \notin S\mathcal{RC} \land v \in \mathcal{FO}_{u} \quad (32)$$

$$\sigma_{d(u,v)^{0}}^{2} = c_{5}(u,v)^{2}(\sigma_{L_{g}}^{2} + \sigma_{L_{l}}^{2}) + c_{6}(u,v)^{2}(\sigma_{V_{g}}^{2} + \sigma_{V_{l}}^{2}) \quad (33)$$

We formulate the *statistical time slack allocation problem* using objective function (15), additional constraints (11), slack upper bound constraints (5), statistical power constraints (23) and (29), plus statistical timing constraints (1), (2), (31) and (32). It is easy to verify that the statistical power constraints (23) and (29) are second-order conic functions, and the objective function (15) and all other constraints are linear. Hence we have the following theorem.

THEOREM 2. The statistical time slack allocation problem is a second-order conic programming (SOCP) problem.

An SOCP problem is convex and can be solved efficiently using the interior point methods with a globally optimal solution [2]. The complexity of solving a SOCP problem is close to $O(N^{1.3})$ in the size of the benchmark circuit. By solving the statistical time slack allocation problem, we allocate time slack to each routing tree considering probabilistic timing and power constraints.

3.2 Statistical Net-level Bottom-up Assignment and Refinement

After statistically allocate time slack to each routing tree, we perform a net-level bottom-up assignment similar to that as in [6] to leverage the allocated slack (see Figure 1). For each tree \mathcal{R}_i , VddH is first assigned to all the switches in \mathcal{R}_i . We then iteratively perform the following steps in a bottom-up fashion. A switch is defined as a *candidate switch* if it follows into two categories. In the first category, a candidate switch is 'untried', and it does not drive any switch. In the second category, VddL has been assigned to all the fanout switches of a candidate switch. We assign VddL to a candidate switch and mark the switch as 'tried'. After updating the circuit timing, we reject the assignment and restore the Vdd-level of the switch to VddH if the delay increase at any sink exceeds the allocated slack. Different from [6] that uses the nominal delay value, we use $d_{ij}^0 + \phi^{-1}(\zeta)\sigma_{d_{ij}^0}$ as the delay of j^{th} switch in \mathcal{R}_i . VddL can be assigned to j^{th} candidate switch in \mathcal{R}_i only if all the

sinks in
$$SL_{ij}$$
 have available slack greater than Δd_{ij} , where

$$\Delta d_{ij} = d_{ij}^0 (V ddL) - d_{ij}^0 (V ddH) + \phi^{-1}(\zeta) [\sigma_{d_{ij}^0 (V ddL)} - \sigma_{d_{ij}^0 (V ddH)}]$$

The iteration terminates when there is no candidate switch in \mathcal{R}_i .

Bottom-up assignment within \mathcal{R}_i :
Assign VddH to all switches in \mathcal{R}_i and mark them as 'untried';
While $(\exists \text{ candidate switch})$
Assign VddL to a candidate switch j ;
If (timing constraints violated)
{
Find all the upstream switches of j in \mathcal{R}_i ;
Assign VddH to j and those upstream switches, and
mark them as 'tried';
}
Else mark j as 'tried';
Update slack for sinks in $\mathcal{S}L_{ij}$;
}

Figure 1: Statistical net-level bottom-up assignment.

After net-level assignment, we may further reduce power by leveraging surplus slack. To leverage surplus slack, we mark all the VddH switches as 'untried' but keep the VddL switches as 'tried', and then perform the sensitivity based algorithm dTLC-S in [6] but use $d^0 + \phi^{-1}(\zeta)\sigma_{d^0}$ as the switch delay to achieve more VddL switches and further reduce power.

4. REFERENCES

- R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Parametric yield estimation considering leakage variability," in *Proc. Design Automation Conf.*, June 2004.
- [2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, 2004.
- [3] Ho-Yan Wong and Lerong Cheng and Yan Lin and Lei He, "FPGA device and architecture evaluation considering process variations," in *Proc. Intl. Conf. Computer-Aided Design*, November 2005.
- [4] M. Mani, A. Devgan, and M. Orshansky, "an efficient algorithm for statistical minimization of total power under timing yield constraints," in *Proc. Design Automation Conf.*, June 2005.
- [5] Kim and et al, A Heuristic for optimizing stochastic activity networks with applications to statistical circuit sizing. preprint.
- [6] Y. Lin and L. He, "Leakage efficient chip-level dual-vdd assignment with time slack allocation for FPGA power reduction," in *Proc. Design Automation Conf.*, June 2005.