J Comb Optim (2012) 23:292–300
DOI 10.1007/s10878-010-9355-1

The complexity of VLSI power-delay optimization by interconnect resizing

Konstantin Moiseev · Avinoam Kolodny · Shmuel Wimer

Published online: 21 September 2010

Abstract The lithography used for 32 nanometers and smaller VLSI process technologies restricts the interconnect widths and spaces to a very small set of admissible values. Until recently the sizes of interconnects were allowed to change continuously and the implied power-delay optimal tradeoff could be formulated as a convex programming problem, for which classical search algorithms are applicable. Once the admissible geometries become discrete, continuous search techniques are inappropriate and new combinatorial optimization solutions are in order. A first step towards such solutions is to study the complexity of the problem, which this paper is aiming at. Though dynamic programming has been shown lately to solve the problem, we show that it is NP-complete. Two typical VLSI design scenarios are considered. The first trades off power and sum of delays (L_1), and is shown to be NP-complete by reduction of PARTITION. The second considers power and max delays (L_∞), and is shown to be NP-complete by reduction of SUBSET_SUM.

Keywords Power-delay optimization · VLSI interconnects · NP-completeness

1 Introduction

The interconnecting wires in VLSI chips are routed in several metal layers stacked one above the other, where the wires are typically running in alternating orthogonal directions as shown in Fig. 1 (Weste and Harris 2010). Figure 2 illustrates the connection of two circuits of the chip, one is called driver and the other is called receiver.

K. Moiseev · A. Kolodny
EE Dept., Technion, Israel Institute of Technology, Haifa, Israel

S. Wimer (✉)
Eng. School, Bar-Ilan University, Ramat-Gan, Israel
e-mail: wimers@macs.biu.ac.il

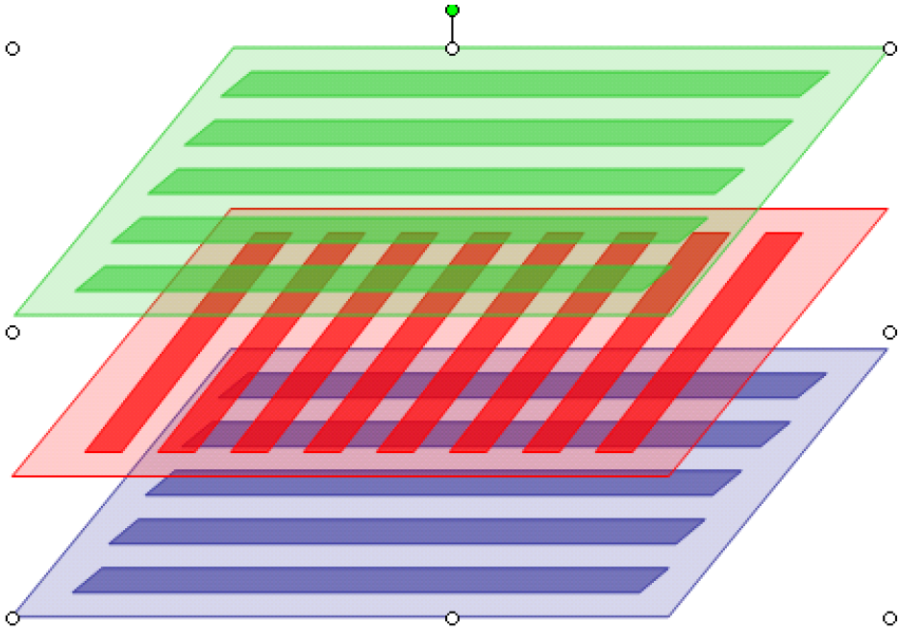

Fig. 1 The interconnecting metal layer regime in VLSI chips. Metal layers are stacked one above the other and directions of interconnects are alternating between adjacent layers

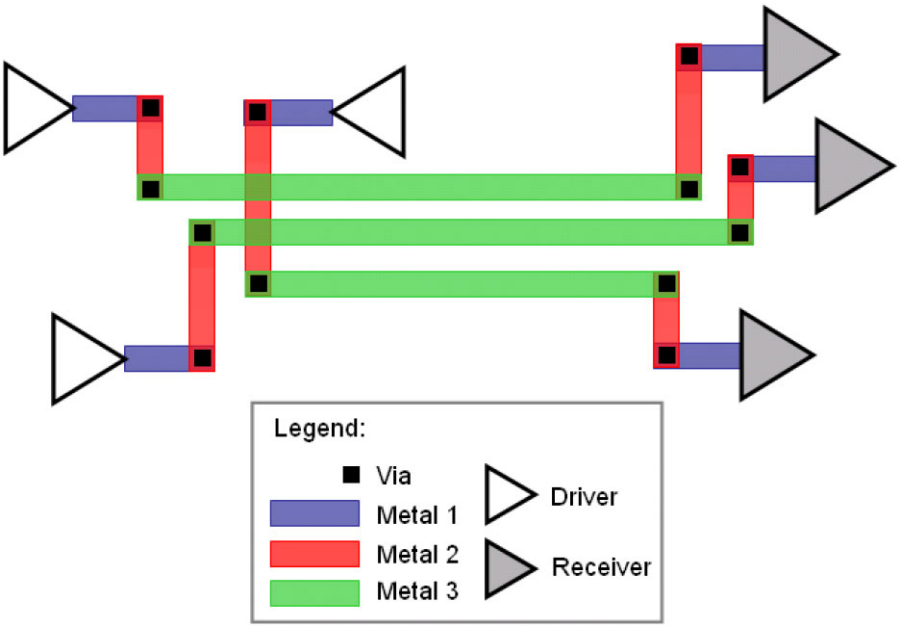

Fig. 2 Interconnecting circuits in VLSI chips. A driving circuit is connected at the near end of a network. The signal is propagating along metal wires to the receiving circuit connected at the far end of the network

The interconnecting wires at the near and far ends reside on a lower metal layer, but switch to upper layers along their way in order to achieve high electrical performance.

Power consumption and speed of VLSI systems and their tradeoff, aka power-delay tradeoff, are important design considerations in state-of-the art manufacturing technologies. The scale down of VLSI manufacturing technologies is lasting for more than four decades, obeying the well-known Moore's Law (Moore 1965), and this trend will continue for the next decade at least (ITRS 2009). Though technology progression enables the integration of complex systems on silicon die, it makes the design effort for high performance chips more and more difficult. The lasting trend towards higher speed is increasing the power consumption, while recent demand for mobile products is driving reduction of power dissipation (ITRS 2009). Unfortunately, power and speed are often in conflict with each other and their tradeoff is delicate and challenging. As a part of the VLSI design optimization techniques, interconnects are subject to small adjustments for setting their widths and spaces (Cong et al. 2001; Wimer et al. 2006).

Physical connectivity must be maintained under any horizontal shift of vertical wires or vertical shift of horizontal wires. Shifting wires in one layer doesn't affect the spacing and width of the orthogonal wires in the layers above it and below it. The length changes of wires in layers above and below of optimized layer is negligible for all practical cases (Moiseev et al. 2009). Until recently the sizes of interconnects were allowed to change continuously and the implied power-delay optimal tradeoff could be formulated as a convex programming problem, for which classical search algorithms are applicable (Zuber et al. 2009).

A new degree of optimization difficulty was introduced with the appearance of 32 nanometer and smaller process technologies (ITRS 2009), where the lithography restricts the admissible sizes and spaces of interconnects to very few values. Once the admissible geometries and their distances of each other become discrete, continuous search techniques are inappropriate and new combinatorial optimization solutions are in order. The complexity of delay-area optimization has been discussed in Li et al. (1993) with regard of sizing the drive strength of logic cells. Though dynamic programming has been shown lately to solve our interconnect problem (Moiseev et al. 2010), studying its complexity is important and discussed in the rest of the paper. Section 2 sets interconnect physical modeling and its related power and delay, where Sect. 3 proves the NP-completeness of their optimization.

2 Delay and power modeling of interconnects in a bundle

Let $\sigma_1, \ldots, \sigma_n$ be n signals of a wire bundle, and let $I_1, \ldots, I_n$ be their corresponding wires positioned between two shielding wires I_0 and I_{n+1} connected to ground, as shown in Fig. 3. As shown in the figure, R_i represents the power drive of a driver where a signal starts, while C_i represents the capacitive load of the receiver at the terminating end of the signal. Let $w_1, \ldots, w_n$ be wire widths and $s_0, \ldots, s_n$ be the spaces between them. It is assumed that admissible wire widths and spaces are taken from finite, very small sets, representing gridded (discrete) design rules.

$$w_i \in \boldsymbol{W} = \{W_1, \ldots, W_q\}, \qquad s_i \in \boldsymbol{S} = \{S_1, \ldots, S_p\} \tag{2.1}$$

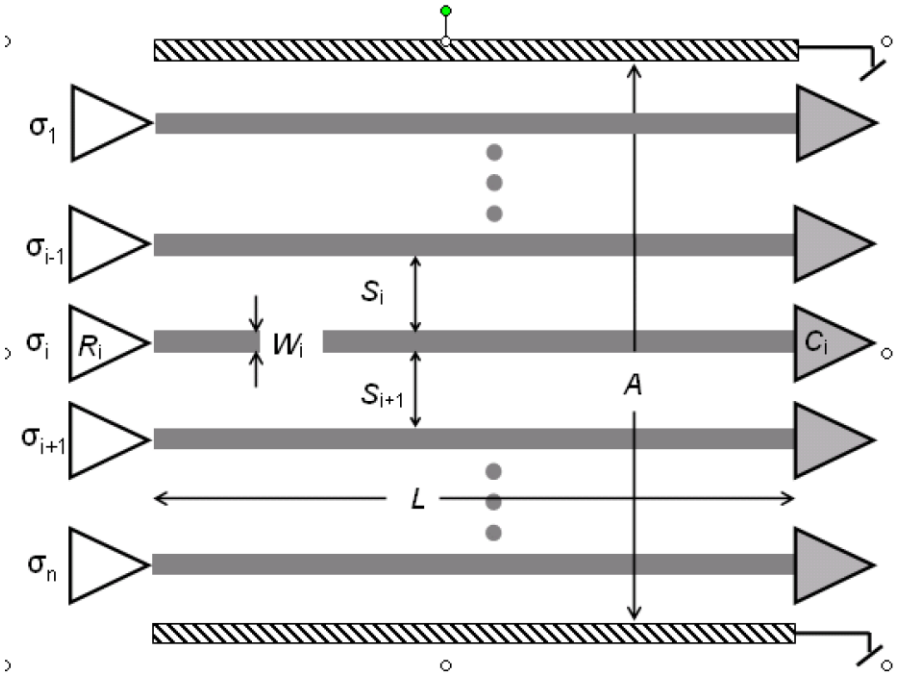

Fig. 3 A fundamental model of interconnecting bus comprising parallel wires laid on the same layer between two shielding wires

Sometimes, a mix of discrete values with continuous ranges is allowed, but design practice usually employs only a limited set of values, turning the problem into pure discrete. Lithography may sometimes prohibit certain width and space combinations by imposing interdependencies between the values in (2.1). We'll ignore such restrictions as those don't affect the complexity of the problems. The area allocated for the wire bundle dictates a total width limit A, satisfying:

$$\sum_{i=1}^{n} w_i + \sum_{i=0}^{n} s_i \leq A \tag{2.2}$$

The delay of signal σ_i can be approximated by Elmore model (Boese et al. 1993) as follows:

$$D_i(s_{i-1}, w_i, s_i) = \alpha_i + \beta_i w_i + \gamma_i / w_i + (\delta_i + \varepsilon_i / w_i)(1/s_{i-1} + 1/s_i), \quad 1 \leq i \leq n \tag{2.3}$$

The coefficients $\alpha_i, \beta_i, \gamma_i, \delta_i$ and ε_i capture process parameters, driver's resistance and capacitive load, and interconnect length, which is fixed in this setting. The dynamic switching power P_i consumed by σ_i is given by:

$$P_i(s_{i-1}, w_i, s_i) = \kappa_i w_i + \eta_i (1/s_{i-1} + 1/s_i), \quad 1 \leq i \leq n \tag{2.4}$$

The coefficients κ_i and η_i capture process parameters, signal activity factors and interconnect length. Signal activity factor is the amount of switching relative to the clock signal. It can range from zero if the signal never switches (e.g., shields or power delivery wires) to one if it toggles twice at every cycle (e.g., clocks). Signal activity factors are derived from functional simulations which check the signal activity in representative scenarios, and then averaging those over all cases (Magen et al. 2004).

Delay and power models in (2.3) and (2.4) are commonly used in Moiseev et al. (2009), and the parameters in their expressions are not subject to optimization. The total sum of delays, maximal delay and total interconnect power consumption are given respectively by:

$$D^{\text{sum}}(\bar{s}, \bar{w}) = \sum_{i=1}^{n} D_i(s_{i-1}, w_i, s_i) \tag{2.5}$$

$$D^{\max}(\bar{s}, \bar{w}) = \max_{1 \le i \le n} D_i(s_{i-1}, w_i, s_i), \quad \text{and} \tag{2.6}$$

$$P(\bar{s}, \bar{w}) = \sum_{i=1}^{n} P_i(s_{i-1}, w_i, s_i) \tag{2.7}$$

The total delay in (2.5) is in fact L_1 metric, while the max delay in (2.6) is L_∞ metric. Let T_i be the required time of σ_i and $\Delta_i = T_i - D_i$ be its slack. It was shown in Wimer et al. (2006) that maximizing $\sum_{i=1}^{n} \Delta_i$ is equivalent to minimizing $\sum_{i=1}^{n} D_i$, and has the same solution. It was also shown that maximizing minimal Δ_i is a similar problem to minimizing the maximal D_i, since both are convex and same algorithm will solve both. Hence, without loss of generality we'll consider just delays in the discussion.

We show below that finding the minimum delays in (2.5) and (2.6) (or the power in (2.7)) such that the power in (2.7) (or delay in (2.5) and (2.6)) doesn't exceed certain limit, is an NP-complete problem. In the proof we ignore the area constraint, since an area constrained solution implies unconstrained solution, but not vice versa. This follows from the number of distinct possible areas, which is linearly bounded by $n|\boldsymbol{W}||\boldsymbol{S}|$. We could then invoke the algorithm of the area constrained problem for each possible area and obtain the solution for the unconstrained one. Hence the latter problem is generally easier than the former one.

3 NP completeness of power-delay optimization

Once all parameters of the bundle are set, namely, drivers, capacitive loads and activity factors, the optimal sizing problem is equivalent to the following. Let "base" power and delay be calculated for the setting in which all wire widths and spaces are at minimum, namely, w_1 and s_1. We then seek an assignment of extra widths and spaces such that the total power (delay) is maximally reduced while total delay (power) change doesn't exceed certain limit. In the sequel we show that a simpler decision sub-problem, called MIN_DLYPWR, is NP-complete.

MIN_DLYPWR:

Instance: A n-wire bundle with given drivers, capacitive loads and activity factors, whose wire widths and spaces are given in (2.1).
Question: Is there a setting of the widths and spaces of wires in the bundle such that delay reduction from the base delay is δD at least, while power increase from the base power is δP at most?

It follows from the delay and power equations given in (2.3) and (2.4), respectively, that both are monotonic decreasing in spacing. Wider wires always increase power, but may increase or decrease delay, depending on driver's resistance. We prove that the MIN_DLYPWR problem is NP-complete by showing that any instance of the NP-complete PARTITION problem (Garey and Johnson 1979) can be transformed in polynomial time into a special instance of MIN_DLYPWR, such that the answer to PARTITION is YES if and only if it is so for the special MIN_DLYPWR instance. The proof follows some ideas used in Li et al. (1993) which proves that the problem of trading off area and delay by cell resizing is NP-complete.

Theorem 1 *MIN_DLYPWR in NP-complete.*

Proof MIN_DLYPWR clearly belongs to NP, as given a guess of widths and spaces, one needs only to substitute those in the appropriate equations, which requires polynomial time. (Notice that in the presence of an area constraint, the problem remains NP as a summation of wire widths and spaces determines whether the area constraint is met.) An instance I of a PARTITION problem attempts to answer whether for a given set B whose elements have size $s(b) \in \mathbb{Z}^+$ for any $b \in B$, there is a subset $B' \subseteq B$ satisfying $\sum_{b\in B'} s(b) = \sum_{b\in(B-B')} s(b)$.
MIN_DLYPWR instance $f(I)$ is built as follows:

1. For every element $b \in B$ of PARTITION we allocate a wire in the bundle.
2. Drivers of wires have zero resistance (infinite current drive) and zero internal delay, hence they don't affect signal delays (via interconnect capacitances). The coefficients α_b, β_b, δ_b, ε_b, η_b are set to 0. The coefficients γ_b and κ_b are set so that $\gamma_b = C_b$ (capacitive load of wire b) and $\kappa_b = F_b$ (activity factor of wire b), yielding $D(s_{b-1}, w_b, s_b) = \gamma_b/w_b$ and $P(s_{b-1}, w_b, s_b) = \kappa_b w_b$.
3. We fix the allowable spaces to minimum value only, namely, $s_b \in \boldsymbol{S} = \{S_1\}$. It means that cross coupling capacitance does not affect this MIN_DLYPWR instance.
4. Wire width has only two admissible values $w_b \in \boldsymbol{W} = \{W_1, W_2\}$, $W_1 < W_2$. All wire widths are initially set to $w_b = W_1$.
5. The area limit A of the bundle is sufficiently large to accommodate any width sizing, so it doesn't affect this MIN_DLYPWR instance.
6. Every signal corresponding to $b \in B$ is assigned with an activity factor $F_b = s(b)/(W_2 - W_1)$ and a capacitive load $C_b = s(b)W_1W_2/(W_2 - W_1)$. Under these assignments (2.5) and (2.7) turn into:

$$D^{\text{sum}} = \sum_{b\in B}(1/w_b)s(b)W_1W_2/(W_2 - W_1) \tag{3.1}$$

$$P = \sum_{b \in B} w_b s(b)/(W_2 - W_1) \tag{3.2}$$

7. We finally set the power increase upper bound and delay reduction lower bound to be equal to each other such that $\delta P = \delta D = \sum_{b \in B} s(b)/2$.

It is obvious that $f(I)$ can be constructed in polynomial time. Assume that the answer to MIN_DLYPWR $f(I)$ problem is YES. Notice that because drivers' resistance was set to zero, delay is monotonic decreasing in wire width. Power is always monotonic increasing in wire widths. Hence there exists only a single value where they are equal to each other, the only value for which a YES answer holds for the MIN_DLYPWR problem. This value must be $\sum_{b \in B'} \delta P_b = \sum_{b \in B'} \delta D_b$. By definition, a YES answer to MIN_DLYPWR implies a subset $B' \subseteq B$ of wires which have been upsized from W_1 to W_2, decreasing delay and increasing power such total delay decrease and total power increase satisfy $\sum_{b \in B'} \delta D_b \geq \delta D$ and $\sum_{b \in B'} \delta P_b \leq \delta P$, respectively. It follows from (3.1) and (3.2), and the setting (7), that $\sum_{b \in B'} \delta P_b = \sum_{b \in B'} \delta D_b = \sum_{b \in B} s(b)/2$. Calculation of delay reduction (power increase is similar) yields $(\sum_{b \in B} s(b))/2 = \sum_{b \in B'} \delta D_b = \sum_{b \in B'} (1/W_1 - 1/W_2) s(b) W_1 W_2/(W_2 - W_1) = \sum_{b \in B'} s(b)$, which implies that $(B', B - B')$ is a YES answer to PARTITION.

Conversely, if $B' \subseteq B$ is a YES answer to PARTITION, we widen the wires corresponding to $b \in B'$ form W_1 to W_2. The delay given in (2.3) is thus reduced for each wire of B' by $\delta D_b = C_b(1/W_1 - 1/W_2) = s(b)$, while the power given in (2.5) is increased by $\delta P_b = F_b(W_2 - W_1) = s(b)$. Summing over all wires of B' obtains a YES answer to MIN_DLYPWR $f(I)$ problem. □

Consider now the problem of minimizing the power in (2.7) such that the maximal wire delay in (2.6) doesn't exceed a certain value, a problem we call MIN_MAX_DLYPWR. In this case we'll tradeoff power decrease for delay increase as follows.

MIN_ MAX_ DLYPWR:

Instance: same as in MIN_DLYPWR.
Question: Is there a setting of the widths and spaces of wires in the bundle such that the power decrease from the base power is δP at least while delay increase δD_i, $1 \leq i \leq n$, from the base delay is δD at most?

Theorem 2 *MIN_MAX_DLYPWR is NP-complete.*

Proof MIN_MAX_DLYPWR clearly belongs to NP. We'll reduce a well known NP-complete problem called SUBSET_SUM (Garey and Johnson 1979) into MIN_MAX_DLYPWR. An instance I of a SUBSET_SUM problem attempts to answer whether for a given set B whose elements have size $s(b) \in \mathbb{Z}^+$ for any $b \in B$, and a given number $M \in \mathbb{Z}^+$, there is a subset $B' \subseteq B$ satisfying $\sum_{b \in B'} s(b) = M$.

The base delay and power in this case are obtained by initially setting all wire widths to W_2, which results in maximum base power. Settings 1 to 5 of MIN_MAX_DLYPWR instance $f(I)$ are similar to those in MIN_DLYPWR proof.

It follows from $W_2 > W_1$ that the base delays are minimal and they increase whenever a wire is narrowed. Setting 6 of Theorem 1 is modified such that the capacitive load is set to $C_b = MW_1W_2/(W_2 - W_1)$ for all wires $b \in B$. Under this assignment (2.6) turns into:

$$D^{\max} = \max_{b \in B}\{M(1/w_b)W_1W_2/(W_2 - W_1)\} \tag{3.3}$$

Consequently wire narrowing from W_2 to W_1 results in delay increase $\delta D_b = M$ for every narrowed wire. Finally, setting 7 of Theorem 1 is modified to $\delta P = \delta D = M$.

The theorem follows by similar arguments as in Theorem 1. Narrowing a wire $b \in B$ from W_2 to W_1 adds $s(b)$ to total power reduction at the expense of increasing signal's delay by M. It follows immediately that $\delta P = \delta D = M$ iff $\sum_{b \in B'} \delta P_b = \max_{b \in B'}\{\delta D_b\} = M$, and this holds iff the answer to SUBSET_SUM is YES. □

It has been shown that the decision problems MIN_DLYPWR and MIN_MAX_DLYPWR are NP-complete, where the area constraint has been dropped. In VLSI practice we are interested in the function describing the power-delay dependency (tradeoff function), where area is usually constrained. This is a convex function describing the minimum power (delay) that can be achieved for a delay (power) not exceeding a certain value. A dynamic programming algorithm finding the power-delay tradeoff function for real industrial problems has been reported in Moiseev et al. (2010). This algorithm approximates the function to any desired accuracy $\varepsilon > 0$, while its complexity is a polynomial in $1/\varepsilon$, the number n of wires, the number of admissible wire widths $|\boldsymbol{W}|$ and spaces $|\boldsymbol{S}|$.

4 Conclusions

In this paper we have shown that several typical problems of power-delay optimization by interconnect resizing in VLSI design turn to be NP-complete once the design rules of process technology are discrete rather than continuous. The transition into discrete design rules is a must in nanometer-scale manufacturing process technologies, and in the near future more and more design optimization problems may face similar situations.

Acknowledgement The authors are grateful for the anonymous reviewers for their useful comments which helped in improving the manuscript.

References

Boese KD, Kahng AB, McCoy BA, Robins G (1993) Fidelity and near optimality of Elmore-based routing constructions. Digest of technical papers, ICCAD, pp 81–84

Cong J, He L, Koh CK, Pan Z (2001) Interconnect sizing and spacing with consideration of coupling capacitance. IEEE Trans Comput Aided Des Integr Circuits Syst 20(9):1164–1169

Garey MR, Johnson DS (1979) Computers and intractability. Freeman, New York

ITRS—International Technology Roadmap for Semiconductors, 2009 edn. http://www.itrs.net/Links/2009ITRS/Home2009.htm

Li W-N, Lim A, Agrawal P, Sahani S (1993) On the circuit implementation problem. IEEE Trans Comput Aided Des Integr Circuits Syst 12(8):1147–1156

Magen N, Kolodny A, Weiser U, Shamir N (2004) Interconnect power dissipation in a microprocessor. In: International workshop on system-level interconnect prediction, pp 7–13

Moiseev K, Wimer S, Kolodny A (2009) Power-delay optimization in vlsi microprocessors by wire spacing. ACM Trans Des Automat Electron Syst 14(4):55

Moiseev K, Kolodny A, Wimer S (2010) Interconnect bundle sizing under discrete design rules. IEEE Trans Comput Aided Des Integr Circuits Syst 29(10) (to appear)

Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8)

Weste N, Harris D (2010) CMOS VLSI design: circuit and system perspective. Addison Wesley/Longman, Reading/Harlow

Wimer S, Michaely S, Moiseev K, Kolodny A (2006) Optimal bus sizing in migration of processor design. IEEE Trans Circuits Syst-I 53(5):1089–1100

Zuber P, Bahlous O, Ilnesher T, Ritter M, Stechele W (2009) Wire topology optimization for low power CMOS. IEEE Trans VLSI Syst 17(1):1–11