Wiring Layer Assignments with Consistent Stage Delays*

Andrew B. Kahng
University of California at Los Angeles, CS Dept.
3731 Boelter Hall
Los Angeles, CA 90095-1596
abk@cs.ucla.edu

Dirk Stroobandt†
Ghent University, ELIS Dept.
Sint-Pietersnieuwstraat 41
B-9000 Gent, Belgium
dstr@elis.rug.ac.be

ABSTRACT

Wire sizing, repeater insertion and repeater sizing are necessary to limit delay in on-chip interconnections. When these techniques are applied to nets that are already routed, the results heavily depend on the routing layer chosen for the wire. In this paper, we present a layer assignment method that assigns wires to the layer that is best fit. The method is based on a consistent target delay constraint and uses wire sizing and repeater insertion and sizing. It also considers a repeater area constraint and takes the impact of vias into account. A greedy optimization approach is used with the number of layers needed for the wiring as its cost function. Our layer assignment method can be used in conjunction with a priori wirelength estimation models so that it applies both as a guide for the router as well as for placement tools. Our model suggests that vias can severely impact the solution when tight delay constraints are applied, and that this actually sets an upper bound to the number of wires that can be accommodated in any layer stack. Empirical results provide some answers as to the best form of the layer stack. Layer stacks with monotonically increasing wire height on the layers are optimal for tight delay constraints. If longer delays are allowed, the addition of a low-level layer on top of the layer stack might be beneficial.

Categories and Subject Descriptors

B.7.2 [**Hardware**]: Integrated Circuits—*Placement and routing* ; J.6 [**Computer Applications**]: Computer-Aided Engineering—*CAD* ; I.6.5 [**Computing Methodologies**]: Simulation and Modeling—*Model Development*

General Terms

Layer assignment, routing, delay, via impact, wire length.

*This research was supported by Cadence Design Systems, Inc. and the MARCO Gigascale Silicon Research Center.

†Dirk Stroobandt is a Postdoctoral Fellow of the Fund for Scientific Research (F.W.O.) – Flanders. This research was performed during his stay at UCLA as a visiting researcher.

SLIP 2000, San Diego, CA.

1. INTRODUCTION

Deep submicron (DSM) routing tools are not only concerned with connectivity, but must also take into account delay constraints, yield, power, etc. Current routing schemes are based on wire tapering, repeater insertion and repeater sizing [1; 2; 3].

Simply finding a routing layout that results in the minimum overall wiring area is already intractable. Additionally, delay constraints require wire sizing and repeater insertion, on a wire by wire basis, i.e., given the wire's length (or even location) and given the wiring layer(s) where it is routed. If the routing solution to which such techniques are applied is obtained by conventional routing, we may end up with inferior overall results. In particular, the decision to assign a wire to a certain layer (or layers) is becoming increasingly important.

In this paper, we present a layer assignment method that (i) uses the optimal wire size and optimal number and size of repeaters for each wire, (ii) ensures that delay constraints are met, (iii) takes into account a total repeater area constraint, and (iv) accounts for the impact of vias. The model uses a priori estimation techniques. It can therefore be applied to obtain a layer assignment prediction before any layout step (including placement) is performed. With more information (e.g., the exact wire length distribution after placement), the model can afford more accurate layer assignment predictions. In general, our layer assignment model has such potential applications as: (i) improving CAD layout tools, (ii) studying the effects of technological parameters on the routing solution, and (iii) optimizing the fabrication process (e.g., to define the best wire width and spacing parameters for each layer). Indeed, the model and studies that we describe are easily integrated within such a framework as is provided by the recent GTX technology extrapolation system [4].

Section 2 introduces the layer assignment problem and the various models used. Our layer assignment method is outlined in Section 3 and Section 4 explains it in more detail. In Section 5, we derive an upper bound (due to the via impact) for the number of wires that can be accommodated on a finite layer stack. Empirical verification shows that the conventional technique of providing fat wires at the highest layers outperforms other layer schemes but that for some examples another scheme might be preferable.

2. PROBLEM DEFINITION AND MODELS

The problem we address is the following: find the optimal assignment of wires to wiring layers subject to delay con-

straints and constraints on the total area used for the repeaters. The optimization objective is the total number of layers needed to route all wires.
Our degrees of freedom in this problem are (i) choice of layer parameters, (ii) wire width, (iii) number of repeaters, and (iv) size of repeaters. In the next subsections, we explain the models that define how these parameters influence layer selection.

2.1 Layer assignment model

We assume a layer assignment model in which wires are assigned to *tiers* (pairs of layers) with one layer for the horizontal and one layer for the vertical wire segments [5; 6]. Tiers are grouped in *tier types*. All layers (tiers) of the same tier type have the same technological parameters, i.e., a fixed wire height, a fixed wire spacing and a minimum and maximum value for the wire widths. Also the type of routing material (Al, Cu, ...), uniform dielectric permittivity, and the dielectric thickness above and below the wires are fixed. The total number of layers in a group of tiers of the same type is to be decided by the layer assignment method but it has to be an even number. Indeed, we need to ensure that there is an equal amount of space reserved for horizontal and vertical connections. In an environment where, e.g., the design has not yet been placed, we can omit this requirement and use the layer assignment result to guide the placement (and the routing) to make efficient use of the directional differences.
The layer assignment method takes as input an arbitrary number of tier types, with technological parameters per type. The order of the tier types (bottom to top) is user-defined. It is thus possible to apply schemes other than the traditional fat-wires-on-top approach. The method searches for the optimal number of layers for each tier type (this can include zero layers) and returns the wires that have to be routed on each tier (we assume each wire uses only one tier). We use the following terminology: a tier i is *lower* (*higher*) than a tier j if the layers of tier i are below (above) the ones of tier j. A tier i is *fatter* than a tier j if its wire height is larger.

2.2 Delay equation and repeater model

The delay model is important since the method is driven by the delay constraint. We use Sakurai's delay equation [7]

$$T_d = 0.377 R_w C_w + 0.693 \left(R_o \left(C_j + C_i\right) + R_o C_w + R_w C_i\right), \quad (1)$$

with R_w and C_w the wire resistance and capacitance, R_o the (effective) output resistance of the gate that drives the wire and C_j and C_i the junction and input capacitances of the gate that is driven by the wire.
The delay can be minimized by tuning several parameters. It depends on the wire length ℓ and wire width W through the wire resistance R_w and wire capacitance C_w. The BACPAC model [8; 9] provides the following relations (the capacitance model includes fringing capacitances)

$$R_w \sim \frac{\ell}{W}, \qquad C_w = \ell \left(a\, W + b - c\, e^{-d\, W}\right), \quad (2)$$

with a, b, c and d technology-derived constants for a given tier type. With the length of the wire and the layer choice fixed, the delay can be lowered by increasing the wire width. We assume wires are sized uniformly over their entire length. We do not consider wire tapering, not only because it makes the solution of our problem much more difficult to obtain but also because it is less compatible with modern-day routers and because it is not so valuable when repeaters are already used [10].
Gate sizing can also be used to minimize delay, using the following dependencies on the gate width

$$R_o \sim \frac{1}{W_g} \qquad C_j \sim W_g \qquad C_i \sim W_g. \quad (3)$$

If repeaters are inserted to reduce the delay, Equation (1) changes. For the sake of simplicity, we assume that the repeaters are inserted at equal distances and that a repeater is also inserted at the beginning of the line to drive the line. With these assumptions, the time delay is broken up into N_r parts of equal delay[1] (with N_r the number of repeaters)

$$T_d = N_r \left(0.377 \frac{R_w}{N_r} \frac{C_w}{N_r} + 0.693 \left(R_o \left(C_j + C_i\right) + R_o \frac{C_w}{N_r} + \frac{R_w}{N_r} C_i\right)\right) \quad (4)$$

where R_o, C_j and C_i are now parameters of the repeaters driving the wire segments. They have the same dependencies on the repeater width as in Equations (3).
Tuning wire widths and number and size of repeaters not only affects delay, but also impacts area. The layer assignment solution in fact trades off between delay and area. Apart from the direct effect of increasing the wire and repeater parameters, the number of repeaters also indirectly affects area usage on other layers due to the via impact (repeaters break up wires into multiple segments, and each must be connected to the wiring layer). We therefore also need a via impact model.

2.3 Via impact model

The impact of vias on routing, while increasingly important, has always been crudely estimated, e.g., each layer reduces the effective area that can be used for wiring on all layers below it by 15% [11]. More recent models [6; 12] try to assess via impact by observing that there are two types of vias [12]. *Turn* vias are used to switch wires between layers with different routing directions and do not take up additional space beyond that used for the routing of the wires. *Terminal* vias, on the other hand, are used to guide a wire to its tier (in our layer assignment model a wire uses only one tier). These vias take up space on all underlying tiers and block wires on those tiers. The model of Chong and Brayton [6] takes only the actual via area into account. The total wiring area reduction on a layer is then estimated to equal $N_v\, A_v$ where N_v is the number of terminal vias on that layer and A_v is the area a single via takes. The *via impact factor* (wiring area reduction factor due to via blockage) is then

$$f = \frac{N_v\, A_v}{A}, \quad (5)$$

with A the originally available area for wiring. Chen et al [12] acknowledge the via blockage effect more realistically, with a larger via blockage factor

$$f = \sqrt{\frac{N_v\, A_v}{A}}. \quad (6)$$

[1] We do not take into account possible differences between the input capacitance of a repeater and that of the gate driven by the wire.

Both models have been extensively studied in [13] and compared to actual routing results. While [13] shows that there is room for improvement to these models, we will use the best model available to date, Equation (6). Our layer assignment model relies only on the fact that f is an increasing function of N_v (which should be true in all via impact models) and better alternatives can be easily substituted once they become available.

2.4 Interconnection length distribution

In our layer assignment model, all wires are classified according to their lengths. We thus need a wire length distribution that describes the number of wires for each length.[2] Our model does not require any knowledge on how this wire length distribution is obtained. It could be the result of a measurement of distances between placed gates if the model is used after placement. It could also be predicted by an a priori wire length estimation method such as described in [14; 15; 16; 17], based on the *Rent exponent* [18] that describes the topological complexity of designs. An overview of such methods is presented in [19].

3. LAYER ASSIGNMENT WITH CONSISTENT STAGE DELAYS

Our layer assignment method takes the following inputs:
1. Technological parameters for capacitances/resistances.
2. For each layer i: the wire height H_i, spacing S_i and minimum W_{min} and maximum W_{max} wire width.
3. Die area A_{die} (fixed die) including the maximal area to be used for repeater insertion (fraction f_A of the area) as well as the routing efficiency factor η_r (see below).
4. Target delay T_{target} (maximal allowed delay for a wire).
5. The complete wirelength distribution (or necessary inputs to estimators, such as Rent exponent and number of gates).
6. Implementation-related parameters (error bound for iteration, maximum number of "best" moves to be performed in one pass, etc.).

3.1 Cost function

The objective function (cost function) to be minimized by the layer assignment method is the total number of layers, defined as

$$C = \sum_{i=1}^{N_t} L_i, \tag{7}$$

with N_t the total number of tier types and L_i the number of layers at tier type i. The number of layers L_i on tier type i should be an (even) integer. However, this would mean that the cost function only changes in discrete steps whenever a layer is (nearly) fully occupied. Such a cost function is not very conducive to finding a good solution using iterative improvement. Therefore, in a first phase, we let the number of layers be a real number that reflects the amount of space effectively used on a layer. The cost function then becomes

$$C = \sum_{i=1}^{N_t} L_i = \sum_{i=1}^{N_t} \frac{A_i}{A}, \tag{8}$$

with A_i the actual area needed for wiring on tier type i and A the wiring area available on a single layer (equal for all layers). The available area is not necessarily equal to the total layer area because several effects make it impossible to fully pack all layers with wires [13]. We combine all those effects into a single factor η_r (often called the *routing efficiency factor* because it is largely dependent on the quality of the router). If A_{die} represents the actual die area, then the effective wiring area is $A = \eta_r A_{die}$.
The actual area used on tier type i can be divided into a part $A_{w,i}$ used for the actual wiring and a part $A_{via,i}$ 'lost' due to vias needed for wires on higher layers:

$$A_i = A_{w,i} + A_{via,i}. \tag{9}$$

The area taken up for the wiring of a wire k on tier type i is given by $A_{w,i}(k) = \ell(k)\,(W(k) + S_i)$ where $\ell(k)$ is the wirelength of wire k, $W(k)$ is its width and S_i is the wire spacing on tier type i. With $\mathbf{I}_i$ the set of all wires on tier type i we can write

$$A_{w,i} = \sum_{k \in \mathbf{I}_i} \ell(k)\,(W(k) + S_i). \tag{10}$$

The part of the cost function that takes care of the via impact is given by[3]

$$A_{via,i} = \sum_{j=1}^{L_i} A\, f_i(j), \tag{11}$$

with $f_i(j)$ the via impact factor on layer j of tier type i. This factor depends on all wires on layers higher than j (both in tier type i and all tier types that are above i). Using the via impact model of Equation (6) and denoting by $\mathbf{J}_j$ the set of all wires on layers above layer j, we have

$$f_i(j) = \sqrt{\sum_{k \in \mathbf{J}_j} \frac{N_v(k)\, A_{v,i}(k)}{A}}, \tag{12}$$

with $N_v(k)$ being the number of vias due to wire k and $A_{v,i}(k)$ being the area each via occupies on layer i. For calculating the via area $A_{v,i}(k)$ on a tier of type i for a wire k, we assume that (i) vias occupy a square area with side equal to the minimum wire width of their corresponding layer (below), (ii) a line (1-dimensional array) of such vias is used for wires that are wider than the minimum wire width, (iii) the number of vias in the line covers the entire wire width, and (iv) for the part of the via stack on the lower layers, the via sizes scale with the minimal wire width on the corresponding layers but the number of vias in the line remains the same. These assumptions lead to (the line of vias is considered a single via in Equation (12))

$$A_{v,i}(k) = (W_{m,i} + S_i)^2\, \frac{W(k)}{W_{m,k}}, \tag{13}$$

where $W_{m,i}$ ($W_{m,k}$) is the minimal wire width on tier type i (on tier type of wire k) and $W(k)$ is the wire width of a wire k. The term S_i was introduced to include the via spacing.[4] The number of terminal vias on layer j due to a wire k on a higher tier can be estimated using the number of repeaters

[2] Multi-terminal nets are broken up into separate source-sink paths.

[3] Note that the summation only makes sense for integer values of L_i. However, we will average out all vias on tier i over its layers and, afterwards, again allow real values for L_i.

[4] Note that we used spacing around all vias in the line. For large differences of $W_{m,i}$ and $W_{m,k}$, the vias might be far enough apart to let another wire pass in between them.

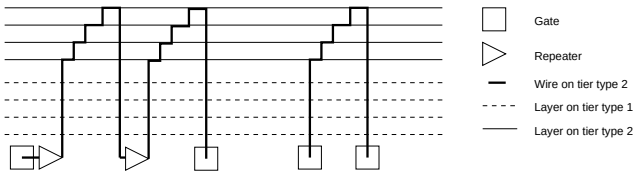

Figure 1: The number of terminal vias on all layers of tier types 1 and 2 for a wire on tier type 2: with two repeaters and without repeaters. Because wires are not assigned to a specific layer within a tier type, we average them out over all layers and assume turn vias to connect the pieces and a single terminal via stack to connect the other line end.

$N_r(k)$ since each repeater (except for the driver) adds two vias per layer that is lower than its own layer and the two endpoints of the connection add another two (left wire in Figure 1)

$$N_v(k) = 2\,N_r(k). \tag{14}$$

If there are no repeaters, then $N_r(k)$ in Equation (14) should be replaced by 1 (not 0) because the endpoints of the wire still need two terminal vias per layer (right wire in Figure 1). Because we do not (yet) distinguish between the layers of a tier type, we assume that equal portions of a wire on tier type i are routed on all of the tier's layers.[5] Each layer on tier type i (except for the top one – but we disregard this anomaly for simplicity) then contains one terminal via and the average number of terminal vias on any layer of tier type i is

$$N_v(i,k) = N_r(k). \tag{15}$$

Again, if no repeaters are used, $N_r(k)$ has to be substituted by 1. Substituting Equations (14), (15) and (13) into Equation (12) yields (on a single layer j of tier type i; the dependency on j is averaged out)

$$f_i = \sqrt{\frac{(W_{m,i}+S_i)^2}{A}\left(2\sum_{k\in\mathbf{J}\setminus\mathbf{I}_i} N_r(k)\frac{W(k)}{W_{m,k}} + \sum_{k\in\mathbf{I}_i} N_r(k)\frac{W(k)}{W_{m,k}}\right)}. \tag{16}$$

The number of layers on tier type i can then be found by combining Equations (8) through (11):

$$L_i = \frac{1}{A}\sum_{k\in\mathbf{I}_i} \ell(k)\,(W(k)+S_i) + L_i\,f_i, \tag{17}$$

$$L_i = \frac{1}{A\,(1-f_i)}\sum_{k\in\mathbf{I}_i} \ell(k)\,(W(k)+S_i). \tag{18}$$

In [20], we single out the change in the cost function (Equation (8)) if only a single wire k changes.

3.2 Layer assignment method

To be able to optimize one wire at a time without having to worry about the influence on other wires, the cost function (and thus the number of layers per tier type) has to be a real number instead of an integer. Because the final solution has to be integer, we propose a method in two phases: (i) optimize the layer assignment with the cost function of Equation (8), and (ii) find the best way of rounding the real numbers to integers so as to minimize the total number of layers and re-assign wires accordingly. In this paper, we focus on Phase 1; any actual layer/tier assignment requires some heuristic for Phase 2 (see [20]).

[5] One can check that this assumption leads to the same result as assuming that wires are distributed randomly on all layers and averaging out all resulting vias over all the layers.

Phase 1

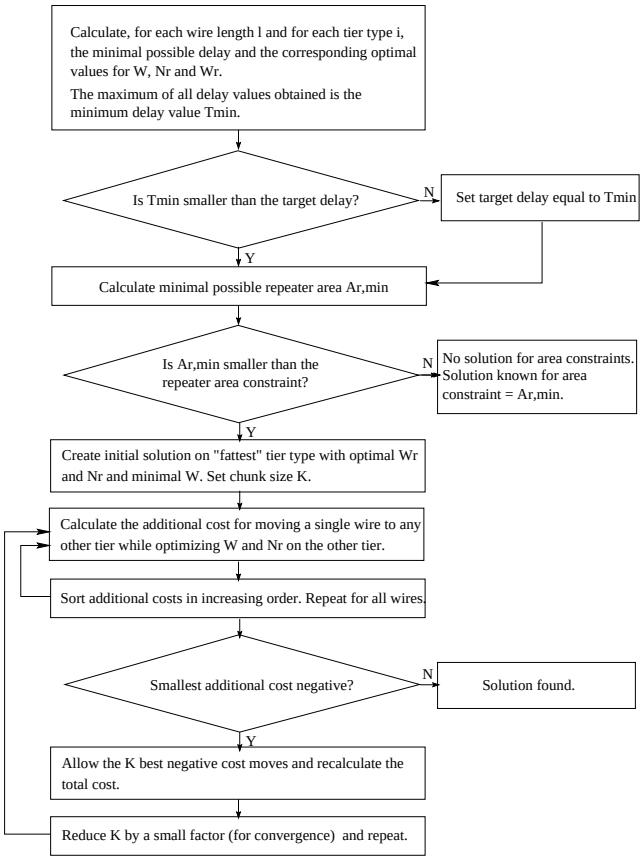

Figure 2: Phase 1 of the layer assignment method.

Phase 1 of the layer assignment method is outlined in Figure 2. It optimizes the wire width W, the number N_r and size W_r of the repeaters, and the tier type to which the wire is assigned. The method is based on a greedy approach in which the K moves that result in the largest reduction of the cost function are performed first. K is called the *chunk size*. At any time, the delay constraint is met by all wires and the remaining degrees of freedom are used to minimize the number of layers (cost function). The solution that is produced by the algorithm of Figure 2 does not yet guarantee that it lies within the constraints for the repeater area. In fact, the optimal repeater width (for minimal delay) is used throughout this algorithm. However, we check that a solution exists that obeys the area constraint. A post-processing step checks the outcome of the algorithm against the area constraint and, if the constraint is not met, gradually reduces the repeater area (see [20] for details).

4. DETAILED MODEL DESCRIPTION

In this section, we clarify selected parts of the algorithm that is outlined in Figure 2. More details are in [20].

4.1 Calculation of the minimal delay

Since the delay equation monotonically increases with the wire length ℓ, the longest wires define the value for the best (minimum) delay we can obtain. The delay equation is a quadratic function of N_r and W_r, hence optimal values for these parameters to minimize delay are easily found [20]. The equation is much harder to solve for W because of the exponential terms (Equation (2)). However, using the results for N_r^{opt} and W_r^{opt}, the delay equation is proportional to the square root of $R_w\,C_w$, which is given by

$$R_w\,C_w = A + \frac{B}{W} - \frac{C}{W}\,e^{-D\,W}, \qquad (19)$$

with A, B, C and D positive constants.

If $B > C$, which always holds due to the physics of the problem, Equation (19) is a decreasing function of W.[6] The minimal delay T_{min} is thus found for $W = W_{max}$. The target delay T_{target}, provided by the user, must be larger than or equal to T_{min} in order to have a solution to the layer assignment problem. The further the target delay is from T_{min}, the more freedom we have to optimize the cost function. (Notice that a tight delay constraint forces an area-inefficient solution).

4.2 Aiming at a target delay

With a target delay larger than T_{min}, more choices of W, N_r, and W_r meet the delay constraint; we use this flexibility to improve the cost function. Given a target delay, the delay equation can again be solved easily to obtain values for the number and size of the repeaters. The quadratic function has either two positive solutions (of which we choose the smaller) or none. For the wire width W, the delay equation can be written (using Equation (19)) as

$$T_d = U + \frac{V}{W} + X\,W - \frac{Y}{W}\,e^{-D\,W} - Z\,e^{-D\,W} \le T_{target}, \qquad (20)$$

with all constants positive and independent of W. Written as

$$e^{-D\,W}\,\left(\frac{Y}{W} + Z\right) \ge U - T_{target} + \frac{V}{W} + X\,W \qquad (21)$$

this equation has a left hand side (LHS) that is a decreasing function of W, and a right hand side (RHS) that has a minimum at

$$W' = \sqrt{\frac{V}{X}}. \qquad (22)$$

Since $0 \le \exp(-D\,W) \le 1$ (for $0 \le W$), a necessary condition for a solution is

$$\left(\frac{Y}{W} + Z\right) \ge U - T_{target} + \frac{V}{W} + X\,W \Leftrightarrow W_1 \le W \le W_2$$

with W_1 and W_2 the solutions of the quadratic equation.

If the discriminant of the quadratic equation is negative or the intervals $W_1 \le W \le W_2$ and $W_{min} \le W \le W_{max}$ do not overlap, there is no solution that meets the target delay constraint (this means the choices for N_r and W_r were insufficient). Otherwise, we have the following possibilities (see Figure 3):

[6] This might no longer be true when inductance is included into the delay model.

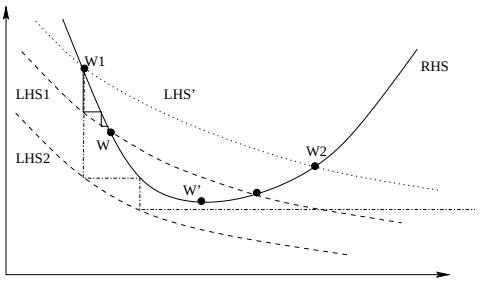

Figure 3: Iteration for obtaining W_{min} with target delay.

1. $W_1 < W' \Rightarrow$ both LHS and RHS of Equation (21) are decreasing functions of W around W_1, and LHS < RHS (because the exponential function is smaller than 1, and we solved W_1 from RHS = LHS' without the exponential factor). Calculating the LHS for $W = W_1$, solving the RHS for W with the calculated value and iteratively improving the value in the LHS with the value for W obtained in the RHS will converge to the solution if one exists. This is shown in Figure 3 if the LHS is given by the curve LHS1. If the RHS does not have a solution for a certain W obtained through iteration, then there is no solution to Equation (21), as is shown in the figure for curve LHS2.

2. $W_1 \ge W' \Rightarrow$ the RHS of Equation (21) is an increasing function of W for all $W > W_1$, and RHS > LHS: there is no solution.

If a solution is found but it is larger than W_{max}, then there is no solution within our restrictions on the wire width. If the value for W is lower than W_{min} then there are again two possibilities: if LHS(W_{min}) $\ge$ RHS(W_{min}) then W_{min} is the best solution (and the delay is lower than the target delay), otherwise there is no solution (W_{min} is larger than the second solution of LHS=RHS).

4.3 Initial solution

To ensure the largest degree of freedom (with regard to the delay constraint), we assume all wires are initially located at the "fattest" tier. We choose the optimal values for number of repeaters and repeater size that achieve minimal delay. The lowest W for each wire is then calculated according to the previous subsection. Because our initial solution uses the "fattest" (i.e., best) tier, a solution will always be found. After this step, the initial total cost function is calculated.

4.4 Move wires to optimal tier

In the next step, we compute the cost change ΔC for a change of each wire (length) separately and for all movements from a tier s to a tier t (s and t might be equal; this allows internal optimization on a tier). The cost is calculated for the best new situation on the new tier, i.e., the optimal W-N_r combination on that tier.

All possible best moves are ordered by increasing cost and the K best ones are actually performed. The parameter K, the *chunk size*, is user-defined. K is automatically reduced during the subsequent optimization steps to promote convergence (when wires change subsequently, the cost function changes with every wire change so that all costs need to be recomputed frequently to obtain the correct ΔC). The optimization algorithm ends when no more wires can be moved to reduce the cost.

After this optimization step, we obtain an optimal solution that adheres to the delay constraint. However, the repeater area has been kept at the optimal value (for minimal delay). If the total repeater area exceeds the area we allowed, we have to reduce it [20].

4.5 Repeater area constraint

The details of the calculations for checking the repeater area constraint can be found in [20]. If the constraint is easily met (which is the assumption we adopt in the result section of this paper), we will have no problem in optimizing the number of layers. If, however, the area constraint is very tight, this will not leave much room for cost improvement and we will end up with many layers. (Imagine if all wires were forced to be twice as wide – a very modest assumption, – increasing the number of layers from 6 to 12!) More importantly, the wire area increase will also result in a significant increase of the via impact, as will be described next. Hence, repeater area constraints should be carefully considered and our model enables the user to explore the implications of such limits.

5. DISCUSSION AND RESULTS

5.1 Via limit

Equation (18) has a solution only if $f_i < 1$. Hence, the via impact must be bounded. Let us look at the implications of this, in the simplified context of all wires on a single tier. From Equation (16), the constraint $f_i < 1$ implies

$$\frac{(W_{m,i} + S_i)^2}{A} \sum_k N_r(k) \frac{W(k)}{W_{m,k}} < 1. \qquad (23)$$

Even if the target delay is chosen such that we do not need repeaters and such that all wires can have the minimal wire width, there is some via impact and the total number of wires N_w that we can allow is bounded by

$$N_w < \frac{A}{(W_{m,i} + S_i)^2}. \qquad (24)$$

Equation (24) quite sensibly indicates that the total number of connections in the design must be less than the effective routing area divided by the area occupied by a single via. For the example 250nm design in [8] (with 10,000,000 transistors and a logic area of 54.08mm^2) on a tier with 4μm wire pitch, the bound would be around 7 million wires. This might seem a large number at first sight but one has to keep in mind that this means a very large number of routing layers. Also, several factors will lower the bound significantly.
In order to obtain Equation (24), we assumed that all wires have minimum wire width and that no repeaters would be used. This would only be possible for very relaxed delay constraints. Wiring limits are much worse if we bound the total number of layers by L_{max}; writing Equation (18) as

$$L_i = L_w \frac{1}{1 - f_i} < L_{max}, \qquad (25)$$

with L_w the number of layers used for wiring alone, the bound on f_i suddenly reduces to

$$f_i < 1 - \frac{L_w}{L_{max}}. \qquad (26)$$

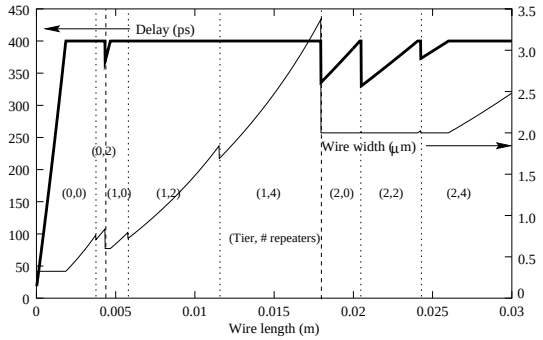

Figure 4: Tier assignment with wire width, number of repeaters and delay obtained for each wire length.

With L_{max} a conservative 10 layers and L_w taking eight of them, f_i would already be reduced to 0.2, reducing the number of wires we can allow to less than 300,000.
The actual bound on the number of wires that follows from via impact should be computed from Equation (23), taking into account the delay constraints and the best solution that obeys them. The above reasoning, however, already shows that we cannot rely on an increasing number of wiring layers to route more interconnections on the same-area die.

5.2 Empirical observations

5.2.1 *Target delay influence*

In Figure 4, the result of our layer assignment method is plotted for a design consisting of three tiers with parameters as in Table 1 (layer stack A). These parameters, as well as the parameters defining resistances and capacitances, are the same as the default parameters for a 250nm process technology in the recent estimation tool BACPAC [8; 9]. The length distribution we used is a theoretical distribution, decaying with ℓ^{2p-3} [14] with a Rent exponent $p = 0.6$. Wire lengths between $20\mu m$ and $30mm$ are considered in intervals of $20\mu m$. The number of wires of length $20\mu m$ was chosen to be 100,000 (total number of wires approx. 187,000). The total available wiring area per layer is chosen to be $20mm^2$ and the routing efficiency $\eta_r = 0.5$.
From Figure 4, we can draw the following conclusions:
1. Longer wires are optimally assigned to higher tiers.
2. The shortest wires obey the target delay easily with minimal wire width. From the moment the target delay is met exactly, the wire width increases to keep meeting the delay. From the moment a threshold wire width is passed, repeaters are inserted and the wire width can be lowered. It then rises again until the next threshold. A second threshold occurs when the wire is moved to a higher tier. Because the wire height then increases, both the wire width and the number of repeaters can decrease. It is beneficial to move wires to a higher tier even if there is some degree of freedom left for the delay constraint. This general scheme can be observed for all examples; only the threshold values change.
Although we have no formal proof that our method finds the optimal result in terms of the number of layers, we have empirically verified that the result is optimal. We disabled the choice of the best cost for every wire tier change (i.e., the choice is made between all improvements at random) and obtained exactly the same results as in the original experiment. We even allowed a gradually decreasing number of moves that increase the cost without any effect on the

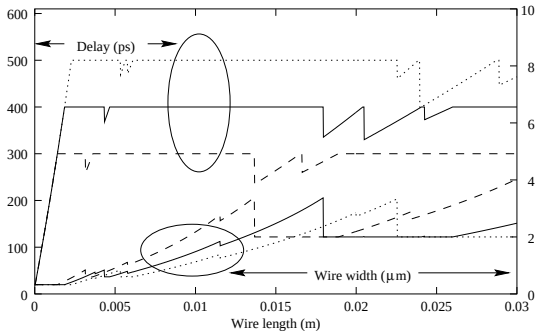

Figure 5: Tier assignment for different delay constraints.

result (except for the fact that it took much longer to find the optimal solution). This clearly shows that our cost function is "smooth" and that the probability of being trapped in a local optimum is nonexistent or at least very small. This was also verified by increasing the "chunk size," i.e., the number of moves allowed without recomputing the cost function; with no effect on the solution.[7]

Of course, the delay constraint has a large effect on the result. In Figure 5, the results for three different delay constraints are plotted. For lower (i.e., tighter) delay constraints, the upper tiers are used for more wires since wire widths must increase starting from shorter lengths (lower set of curves in Figure 5). From the figure, it seems the threshold for moving to a higher tier (big dips in the curves) is only dependent on the wire width value.

5.2.2 Optimal layer stack

Our layer assignment method allows an interesting study of the optimal layer stack. In a first study, we investigate the difference between a uniform layer stack (all tiers have the same type) and two non-uniform ones. Parameters for three different layer stacks `B`, `C` and `D` are given in Table 1. The resulting numbers of layers per tier type are presented in Table 2. It is interesting to note that the uniform layer stack `B` does not result in all wires being at the bottom tier type. This is due to the fact that the via impact on a tier type i is found to be $L_i f_i$ (Equation (17)). While it is true that the via impact on the lower tiers is larger when more wires are moved up, the number of layers that "feel" this impact decreases. At the same time, the upper tiers have less via impact so an increase in the number of layers does not hurt that much. This results in many wires residing on the higher tier type.

Adding non-uniformity to the tier stack allows shorter wires to move down (to profit from the decrease in wire widths). As can be seen from Table 2, increasing the non-uniformity decreases the total number of layers needed. Of course, at some point, smaller wire widths on lower tiers will no longer help because the delay constraint prevents us from using these widths. At the same time, larger wire widths at the upper tiers will reduce the number of wires that are optimally routed there until none will be routed. Our layer assignment method thus enables a search for the optimal layer stack architecture, for a given length distribution and per-stage target delay. The method can also easily handle different delay constraints on different wires, i.e., a 2-D distribution of wire lengths and required wire performances. (Possibly, there are scaling properties of both lengths and required delay performances of wires in well-optimized – delay slack-budgeted, gate-sized, buffer-clustered, remapped, etc. – placed circuits. One might even speculate as to the nature of applicable "temporal Rent parameters" that capture such properties and provide compact descriptions of the 2-D length-delay distribution for placed circuits.)

An interesting question to ask is whether the traditional approach of putting tiers with increasingly fat wires higher in the stack, is in fact optimal. We investigated three other layer stacks (`E`, `F` and `G`) with parameters given in Table 1. Layer stack `E` is the inverse of layer stack `A`, i.e., with "fat" wires on the bottom. Clearly, this layer stack is less optimal in terms of number of layers needed than the original one (see Table 2), mainly because of the huge via impact on the lower tiers. We can also investigate non-monotonic layer stacks such as `F`, which has a "fat" tier type on the bottom and a "conventional" layer stack on top of that. As could

[7]Our experiments use chunk size $K = 100$ (with no reduction in consecutive iteration steps). We have tried various values of $K = 1, 5, 10$ etc. and obtained exactly identical solutions from the model. Hence, K is nothing more or less than a means to expedite model evaluation.

Stack	Tier type	H	S	T	W_{min}	W_{max}
`A`	0	0.5	0.325	0.65	0.325	10
	1	0.9	0.6	0.9	0.6	10
	2	2.5	2	1.4	2	10
`B`	0	1.3	0.6	1.3	0.6	20
	1	1.3	0.6	1.3	0.6	20
	2	1.3	0.6	1.3	0.6	20
`C`	0	1.1	0.5	1.1	0.5	20
	1	1.3	0.6	1.3	0.6	20
	2	1.5	0.7	1.5	0.7	20
`D`	0	0.9	0.4	0.9	0.4	20
	1	1.3	0.6	1.3	0.6	20
	2	1.7	0.8	1.7	0.8	20
`E`	0	2.5	2	1.4	2	10
	1	0.9	0.6	0.9	0.6	10
	2	0.5	0.325	0.65	0.325	10
`F`	0	2.5	2	1.4	2	10
	1	0.9	0.6	0.9	0.6	10
	2	2.5	2	1.4	2	10
`G`	0	0.9	0.6	0.9	0.6	10
	1	2.5	2	1.4	2	10
	2	0.9	0.6	0.9	0.6	10

Table 1: Technological parameters (wire height H and spacing S, dielectric thickness T, and minimal W_{min} and maximal W_{max} wire width, all in μm) for different layer stacks with three tier types each.

Stack	L_0	L_1	L_2	L'_0	L'_1	L'_2	L_{tot}
`A`	1.44	2.03	1.72	1.58	2.10	1.76	5.4436
`B`	0.56	1.27	5.67	0.69	1.40	5.94	8.0251
`C`	1.93	0.44	4.23	2.25	0.46	4.39	7.1053
`D`	1.57	0.45	3.93	1.77	0.46	4.08	6.3134
`E`	5.59	2.40	0.88	15.67	2.77	0.93	19.3665
`F`	0	4.34	1.86	0	5.21	1.92	7.1285
`G`	1.77	1.64	2.79	2.13	1.97	2.89	7.0002

Table 2: Layer assignments for the three different layer stacks of Table 1 (L_x is the number of layers used for wiring on tier type x, L'_x the number of layers needed with inclusion of the via impact).

be expected, the bottom tier type is not used at all. This indicates that the conventional approach is indeed better. However, a conventional layer stack with a "non-fat" tier type at the top (G) does not result in a vacant top tier. Indeed, the results show it is beneficial to move some of the wires to the top tier type! The reason is again that the number of layers needed because of via impact is given by $L_i f_i$. Although the bottom and top tier types have identical parameters, and the via impact is lower if wires are moved down, the increase in the number of layers on the bottom tiers annihilates the cost gain. Thus, in some cases, the conventional layer stack configuration might not be the best one. Note, however, that the difference in total cost is small (compare L_{tot} for stacks F and G in Table 2) and that this non-conventional layer stack is a particular situation. When we decrease the target delay from 400 ps to 300 ps, the top tier is no longer used because the tighter delay constraint requires more repeaters and increases the via impact. For a higher delay constraint however, the top tier was used to a large extent (in the limiting case of no delay constraint at all, the "fat" tier is no longer used and the solution degenerates to the uniform case with only two tier types).

6. CONCLUSION

The assignment of wires to wiring layers will soon become one of the most critical design (and process) optimizations. In this paper, we have presented a method for assigning wires to layers that uses a uniform stage delay constraint and optimizes the number of layers needed for the wires subject to this constraint. The model includes uniform wire sizing, repeater insertion and repeater sizing and also takes the via impact into account, as well as a constraint on the total repeater area. It is shown that the via impact can severely increase the required routing area, especially when delay constraints are very tight. The via impact limits the number of wires that one can accommodate on any arbitrarily large number of layers.

Our layer assignment method can also handle different delay constraints on different wires, i.e., a 2-D distribution of wire lengths and required wire performances, possibly based on what we call "temporal Rent parameters" that capture scaling properties of both lengths and required delay performances of wires in well-optimized placed circuits.

Our layer assignment method provides, apart from an effective algorithmic solution, some interesting conclusions:

1. The maximum wire width used on a certain tier type does not depend on the delay constraint but only on the interaction between the layer stack parameters..

2. A monotonic non-uniform layer stack is better than a uniform one, provided that the "fattest" layers are on top of "less fat" layers. An optimal layer stack can be found.

3. Non-monotonic layer stacks are worse than monotonic layer stacks (with "fattest" layers on top) for tight delay constraints. For a higher delay constraint, a non-monotonic layer stack with a "non-fat" layer on top might be better.

Our layer assignment model can easily be used to investigate different layer stack solutions and to search for the optimal layer stack parameters. Further investigation might reveal the threshold values for input parameters that allow us to make such conclusions beforehand and ensure the optimality of a monotonic layer assignment so as to make the layer assignment task more "trivial".

7. REFERENCES

[1] J. Lillis, C.-K. Cheng, and T.-T.Y. Lin. "Optimal wire sizing and buffer insertion for low power and a generalized delay model." In *IEEE J. Solid-State Circuits*, 31 (3): pp. 437–447, 1995.

[2] C. C. N. Chu and D. F. Wong. "A new approach to simultaneous buffer insertion and wire sizing." In *IEEE/ACM Intl. Conf. on Comp.-Aid. Design*, pp. 614–621, 1997.

[3] C.J. Alpert, A. Devgan, and S.T. Quay. "Buffer insertion for noise and delay optimization." In *35th IEEE/ ACM Design Automation Conf.*, pp. 362–367, 1998.

[4] A. Caldwell, A. B. Kahng, F. Koushanfar, H. Lu, I. Markov, M. Oliver and D. Stroobandt, "GTX: The MARCO GSRC Technology Extrapolation System." To appear in *Proc. ACM/IEEE Design Automation Conf.*, 2000, See: `http://vlsicad.cs.ucla.edu/GSRC/GTX/`.

[5] R.H. Otten and R.K. Brayton, "Planning for Performance." In *Proc. Design Automation Conf.*, pp. 122–127, 1998.

[6] P. Chong and R. K. Brayton. "Estimating and optimizing routing utilization in DSM design." In *Workshop notes 1st Intl. Workshop on System-Level Interconnect Prediction*, pp. 97–102, 1999.

[7] T. Sakurai. "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's." *IEEE Trans. Electron Devices*, 40: pp. 118–124, 1993.

[8] D. Sylvester and K. Keutzer. "Getting to the bottom of deep submicron." In *Proc. ICCAD*, pp. 203–211, 1998. `http://www.eecs.berkeley.edu/~dennis/bacpac/`.

[9] D. Sylvester and K. Keutzer. "System-level performance modeling with BACPAC – Berkeley advanced chip performance calculator." In *Workshop notes 1st Intl. Workshop on System-Level Interconnect Prediction*, pp. 109–114, 1999.

[10] C.J. Alpert, A. Devgan, and S.T. Quay. "Is wire tapering worthwhile?" In *1999 IEEE/ACM Intl. Conf. on Computer-Aided Design*, pp. 430-435, 1999.

[11] G. A. Sai-Halasz. "Performance trends in high-performance processors." In *Proc. IEEE*, pp. 20–36, 1995.

[12] Q. Chen, J.A. Davis, P. Zarkesh-Ha, and J.D. Meindl. "Via impact and via-limited chip size," 1999. Georgia Inst. of Techn. Private communication.

[13] A.B. Kahng, S. Mantik, and D. Stroobandt. "Requirements for models of achievable routing." In *Proc. Intl. Symp. on Physical Design (ISPD)*, 2000.

[14] W. E. Donath. "Placement and average interconnection lengths of computer logic." *IEEE Trans. Circuits & Syst.*, CAS–26: pp. 272–277, 1979.

[15] W. E. Donath. "Wire length distribution for placements of computer logic." *IBM J. of Research and Development*, 25: pp. 152–155, 1981.

[16] J. A. Davis, V. K. De, and J. D. Meindl. "A stochastic wire-length distribution for gigascale integration (GSI) – PART I: Derivation and validation." *IEEE Trans. on Electron Devices*, 45 (3): pp. 580–589, 1998.

[17] D. Stroobandt and J. Van Campenhout. "Accurate interconnection length estimations for predictions early in the design cycle." *VLSI Design, Special Issue on Physical Design in Deep Submicron*, 10, 1999.

[18] B. S. Landman and R. L. Russo. "On a pin versus block relationship for partitions of logic graphs." *IEEE Trans. on Comput.*, C–20: pp. 1469–1479, 1971.

[19] P. Christie and D. Stroobandt. "The interpretation and application of Rent's rule." *IEEE Trans. on VLSI Systems, Special Issue on System-Level Interconnect Prediction*, 1999. Submitted.

[20] A. B. Kahng and D. Stroobandt. "Wiring layer assignments with consistent stage delay." Technical Report CSD-200005, UCLA CS Dept., March 2000.