Toward Accurate Models of Achievable Routing

Andrew B. Kahng, Stefanus Mantik and Dirk Stroobandt

Abstract— **Models of achievable routing, i.e., chip wireability, rely on estimates of *available* and *required* routing resources. Required routing resources are estimated from placement, or (a priori) using wire length estimation models. Available routing resources are estimated by calculating a nominal "supply," then taking into account such factors as the efficiency of the router and the impact of vias.**

Models of achievable routing can be used to optimize interconnect process parameters for future designs or to supply objectives that guide layout tools to promising solutions. Such models must be accurate in order to be useful, and must support empirical verification and calibration by actual routing results.

In this paper, we discuss the *validation* of such models and we apply our validation process to three existing models. We find notable inaccuracies in the existing models when matched against real data. We then present a thorough analysis of the assumptions underlying these models; based on this analysis, we discuss requirements for predictors of routing resources and make suggestions for a new model of achievable routing.

Keywords— **VLSI Routing Estimation, Via Impact Model, Interconnect Process Optimization.**

I. Introduction

IN predictive system implementation methodologies, it is increasingly critical to have accurate models of the routing resources needed to implement interconnect structures. Such models have many applications. Donath's pioneering wire length estimation model [9] based on Rent's rule [16] has been used by Bakoglu [1] as the basis of a system-level performance model. Recent models have addressed either the estimation of the total wire length *required* by the design ("demand") [7], [8], [20], or the estimation of effectively *available* total track length ("supply") on chip [5], [6], [18]. The latter is much smaller than the nominal supply of signal wiring tracks, for reasons that notably include router efficiency and the impact of vias.

Predictions of required and available routing resources together comprise a "model of achievable routing." Given such a model, one can predict the number of wiring layers needed to route a given design in a given technology. Or, if the number of wiring layers and their technology parameters (e.g., wire pitch) are fixed, one can obtain an "oracle" that predicts whether the design is routable in the given resource. These particular applications, along with extrapolations to future designs and process technologies, have been extremely popular and influential [1], [4], [11], [12], [18], [21].

Predictions of achievable routing can be made post-placement based on actual pin locations, or else pre-placement based on a model of the wire length distribution. The model can then provide a priori knowledge about the routing, before any layout step has been performed. One application of such models is to optimize the interconnect process [6], [15] (number of layers, wire pitch on each layer) for a certain class of target designs. Future models should therefore include wire sizing, buffer insertion, tapering, etc.

Optimization of the layout flow also becomes possible. Early predictions are needed for, e.g., wireplanning methodologies [17] where a global wire plan is instantiated beginning at the conceptual stage of physical implementation. At the placement stage, better estimates of routing feasibility can guide placers and reduce incremental placement/routing iterations. Finally, routers could benefit from knowledge of their "routing efficiency" and effectively available routing resources on each layer to improve convergence.

All techniques described in this paper are a priori (i.e., before layout) non-constructive estimation models. Global routing may also be used as a constructive estimator. Indeed, proponents of global routing – notably Scheffer and Nequist [19] – have argued that interconnect estimation can only be performed constructively. This is in some sense a religious issue. We believe there is still a need for the development of strong *non-constructive, a priori* interconnect estimation methods. One of the main reasons for this is that the a priori techniques can also be applied when only limited information on the placement, or even the netlist, is present. Especially for research on future designs (of which almost nothing is known yet) or optimization studies of, e.g., wiring layer parameters, the global routing approach is no longer an alternative to a priori estimation methods. These methods are also ideal for any application where only the average behavior of nets is important, as is the case in the study of the required routing resources.

Contributions of This Work

To effectively guide physical chip implementation, models of achievable routing must be accurate: they must permit empirical verification and calibration by actual routing results. Although accuracy at the level of individual nets is unlikely, models should at least provide an accurate understanding of global parameters of the final route (total wire length, distribution of wires onto various layer pairs, amount of detours or vias, etc.). With this in mind, it is noteworthy that *no existing model of achievable routing presents validation results using real place-and-route data.*

This work was supported by Cadence Design Systems, Inc. and by the MARCO Gigascale Silicon Research Center project on Calibrating Achievable Design.

Andrew B. Kahng is with the UCSD CSE Department, La Jolla, CA, USA. E-mail: abk@cs.ucsd.edu .

Stefanus Mantik is with the UCLA CS Department, Los Angeles, CA, USA. E-mail: stefanus@cs.ucla.edu .

Dirk Stroobandt is Postdoctoral Fellow of the Fund for Scientific Research (F.W.O.) – Flanders and affiliated with Ghent University, ELIS Department, Gent, Belgium. E-mail: dstr@elis.rug.ac.be . This research was performed during his stay at UCLA as a visiting researcher.

Our work centers on (i) understanding the reasons for this validation gap, (ii) processes for model validation, and (iii) necessary improvements in models of achievable routing. Section II reviews three recent models. One has been very influential in technology extrapolation systems; the other two are very recent and attempt to explicitly model the impact of vias on achievable routing. In Section III, we make the case for a thorough validation of (current and future) models of achievable routing through the use of real placement and routing tools. We find that the three recent models predict the available routing resources very differently; indeed, our experimental validation process reveals that none of them is very accurate. In Section IV, we try to assess the reasons behind the failure of existing models. In particular, we experimentally verify their assumptions to expose those assumptions that do not hold. Based on this empirical verification and analysis, we propose some improvements to models of achievable routing in Section V. Our main focus is on the routing efficiency factor and a new via impact model.

II. Models of Achievable Routing

As noted above, all models of achievable routing distinguish between *required* and *available routing resources*. Required routing resources are defined to be the total length of the interconnections that the chip must accommodate. For the a priori context, this total is estimated by wire length distribution models [7] such as those of Donath [10], Davis et al. [8], or Stroobandt et al. [20]. However, for post-placement applications, actual terminal locations of signal nets can be used; this is the approach used in our work, and hence we do not consider any effects of inaccuracies in the estimation of required wiring resources. Rather, our focus will be on models of available routing resources. Note, though, that the techniques presented here are still applicable together with the use of estimated wire length distribution models, whereas a global routing approach would require the actual instantiation of the nets from the distribution.

Available routing resources are significantly less than the nominal total track length on all layers.[1] The first reason for this is that net terminal locations limit the solution space for the routing, so that even an optimal routing solution will not use all tracks completely. Second, routers are not 100% efficient because heuristics are used to solve the NP-hard routing problem (i.e., the optimal solution is out of reach). Third, often a wire must make a detour because vias that connect other wires to higher layers block its path (see Figure 1). There is in fact a *cascade effect* of via blockage, since detours form additional blockages for other wires.

A. A Common Model Framework

Although the first reason given above depends on the netlist topology and on the placement, it is generally combined with the second reason, which depends on the router, into a single *routing efficiency* factor η_r. The impact of the vias on the available routing resources is represented by the *via impact factor* v_i (also called the "via blockage factor"), which represents the fraction of the total available space that is not available due to the via blockage effect on a specific layer i. Finally, the ratio of the total available track length within layer i to the supplied (nominal) track length on the layer, which we call the *utilization factor* U_i, can be written as

$$U_i = \eta_r(1 - v_i). \tag{1}$$

Of course, since all models focus on the resources used for signal nets, the resource used for power and ground wires and for clock distribution must be left out of the estimated available resources. This leads to another factor, the fraction of routing resources used for signal nets only, which we define here as the *signal net fraction* s_i (on layer i). This changes Equation 1 to

$$U_i = \eta_r(1 - v_i)s_i. \tag{2}$$

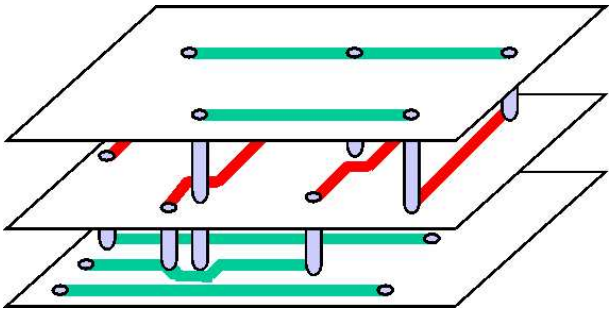

Fig. 1. Wires have to make detours due to via blockage.

B. Review of Existing Models

Sai-Halasz The first model to account for the effects of router efficiency and via impact was used by Sai-Halasz [18] to predict performance trends in microprocessors. The model assumes that power and ground wires take up 20% of each level ($s_i = 0.8$), that the routing efficiency is 40% and that each layer blocks 12% to 15% of the wire capacity of all layers underneath it if the wire pitches are equal. If the pitches differ, this factor has to be increased by taking the ratio of the pitches into account. For N_l layers (numbered from 1 to N_l, going bottom to top), the via impact factor on layer i in Sai-Halasz' model is defined to be

$$1 - v_i = \sum_{k=i+1}^{N_l} 0.85^{\frac{p_i}{p_k}}, \tag{3}$$

where p_x is the wire pitch on layer x.

The Sai-Halasz model has been used by a number of other researchers in "technology extrapolation" to predict future achievable design [11], [12], [21]. However, since it is based only on factors for "good design practice" and attempts to ensure a routable design, it tends to be rather pessimistic about the available resources.

Chong In a paper specifically on estimating routing utilization [6], Chong and Brayton devised a model that

[1] We follow existing practice in the literature by considering the effects listed here within "supply" analysis.

takes as inputs the number of gates, the average area per gate, the average gate pitch, the average fanout of a gate, and the number of layers in the design. It then optimizes the wire width on the layers and predicts the total number of interconnects routed on each layer, the length of the longest interconnect on each layer, and the total available track length.

The model consists of two main parts: the *layer assignment model* and the *available resources model*. The layer assignment model takes a wire length distribution as input (the a priori wire length distribution model of Davis et al. [8] is used, but any other model could be applied). It then assigns interconnects (defined as source-sink pairs) to the layers under the assumptions that (i) layer pairs form *tiers* (one layer provides the horizontal, the other the vertical routing direction), (ii) interconnects can only reside on a single tier, and (iii) shorter interconnects are routed on lower tiers. (The layer assignment model is enhanced with an optimization for wire sizes and addition of delay constraints, but this is not of interest for the present discussion.)

The available resources model reduces the supplied resources on tier m by a constant routing efficiency factor (equal to 0.65 on all layers in their examples) and by the via impact factor according to Equation 1. The latter equates the area "lost" due to via blockage with the total area of all vias that either pass through tier m or connect signals to tier m. Each interconnect on a layer on or above tier m is assumed to contribute two via stacks (one for each terminal) and hence four vias on tier m. The total number of such interconnects (and hence the number of vias) on tier m is defined by the layer assignment model.

In summary, the key points of Chong's model are that the via impact is estimated solely by the total area of the vias, and that the number of vias is estimated from the layer assignment model. It seems likely that at least the first point can lead to underestimation of via impact, since no detour or cascade effect is modeled.

Chen The model of Chen et al. [5] is specifically targeted at the via impact. It classifies vias as either *terminal vias* (those vias that serve the terminals of interconnects) or *turn vias* (those that arise from routing necessity, connecting "doglegs" of interconnects). Turn vias do not add to the via blockage because they are an internal part of the interconnect and can be left out. Only terminal vias are taken into account (this is the case for Chong's model as well). The number of terminal vias on each layer is estimated by a model very similar to Chong's layer assignment model. The authors then distinguish between two cases: (i) *sparse vias*, where the average distance between vias is larger than the average length of an interconnect on that layer, and (ii) *dense vias* otherwise. In the sparse via case, the authors acknowledge that the via impact is indeed limited to the footprint area of the vias (as in the Chong model). However, Chen et al. make the case that realistic situations correspond to the dense via regime.

The via impact model for dense vias presented in [5] assumes that, for every X potential tracks, one track is congested by dense vias and must be given up. The value of X is calculated from the average number of vias per layer side length (and hence $X\,p_i = \sqrt{A_i}/\sqrt{N_i}$, where p_i is the wire pitch on layer i, A_i is the layer area and N_i the number of terminal vias on that layer).[2] If a via is assumed to take p_i^2 area, the via impact factor $(1/X)$ equals the square root of Chong's impact factor, which is based on the via area only.

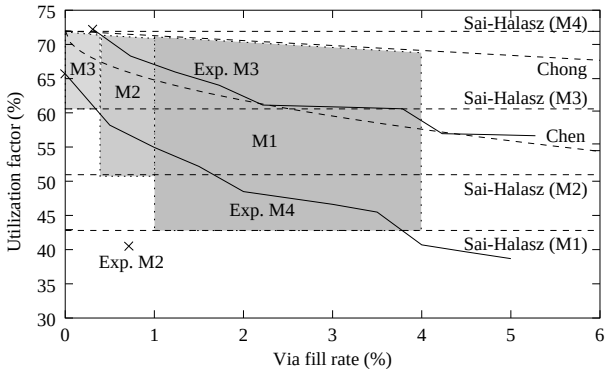

Fig. 2. Utilization factor as a function of the via fill rate for the reviewed models (dashed lines) and experimental results (solid lines). The shaded regions represent the range of predictions across the three models for the different layers in a typical design.

As in the other models, power/ground and clock nets are subtracted from the supplied track length, a routing efficiency factor is used (the authors of [5] use values between 40% and 66% depending on the router and the type of the circuit), and then the via impact factor is included to obtain the final estimate of available resources.

III. Model Validation

We claim that correctness of assumptions and models can be validated *only* by testing the models against comparable experimental results, where "comparable" indicates that the main input parameters to both the model and the experiment (e.g., the number of gates in the system, the number of wiring layers, the wiring pitches, etc.) are identical. In this light, we have to realize that none of the reviewed models has been validated with results of real placement and routing tools. While there are probably reasonable explanations for this, the result is that even a simple comparison between those models reveals huge differences [14].

A. A Simple Comparison of Previous Models

The three models differ only in the way they estimate the via impact. Figure 2 plots for each model the utilization factor U_i as a function of the *via fill rate* f, which we define to be the ratio of the number of terminal vias over the total number of track intersections on a layer. The combined effect of routing efficiency η_r and signal net fraction s_i is set to 72% for all models. We make the following observations.

[2] The expression in [5] is slightly more complicated because the authors also include possibly different wire-to-wire and wire-to-via spacings.

1. Chong estimates the via impact as the total via footprint area, and hence the utilization factor decreases *linearly* with the via fill rate. The same behavior is predicted for all layers (but with a higher f for lower layers).

2. Chen predicts that the utilization factor decreases with the *square root* of f. Again, the same behavior is predicted for all layers (with a higher f for lower layers).

3. Sai-Halasz' model is *independent of the number of vias*, and simply reduces the utilization factor by 15% for each subsequent layer (for simplicity, we assume wire pitches to be constant across all layers).

Experimental measurements (both our own and those of [5]) show that the via fill rate is between 1% and 4% for Metal 1 (`M1`), and much lower than 1% for all higher layers. Given such values of f, Sai-Halasz always has the most pessimistic prediction, Chong always has the most optimistic one, and Chen predicts somewhere in between. The shaded regions in Figure 2 represent the *range of predictions*, across all three models, for the different layers. For `M1`, the predictions of the utilization factor vary by more than 25% in absolute terms, and for `M2` the variance is still 20%. Furthermore, the Sai-Halasz model, with a routing efficiency of 40% and a 20% loss of space for power and ground routing, predicts a `M1` utilization factor of 20%, *a factor of three to four* less than the value predicted by Chong.

B. Experimental Tests of Previous Models

It is tempting to conclude from Figure 2 that Sai-Halasz overestimates the via impact (and underestimates the utilization factor), that Chong underestimates the via impact, and that Chen's model is probably the most accurate. However, such a conclusion is valueless if not backed up by experimental data. In the interest of having comparable inputs, we focus on congested designs.[3] To assess the via impact for congested designs, our experimental setup is as follows.

1. We use a "typical" industry standard-cell block design (approximately 42,000 cells, dating from early 1999) that is routable in a five-layer technology (we use Cadence placement and gridded routing tools with the same 1 μm pitch for all routing layers; via size is .62μm; all pins for cells are on `M1`; the die size is 1.89 mm $\times$ 1.89 mm and the minimal spacing 0.38 μm).

2. We ensure a (globally) congested design by removing the top layer, then gradually removing randomly chosen nets and rerouting the design (each time) until we find that the partial netlist is just routable again. (This procedure creates a maximally congested design in the sense that no net can be added back in without making the design unroutable; the total wire length for the congested design in four layers is 3.96 m.) A maximum routing efficiency value of 72% was found and applied in the Chong and Chen models.

[3]We acknowledge that the three existing models are in some sense meant only to predict the edge of routability, i.e., for congested designs.

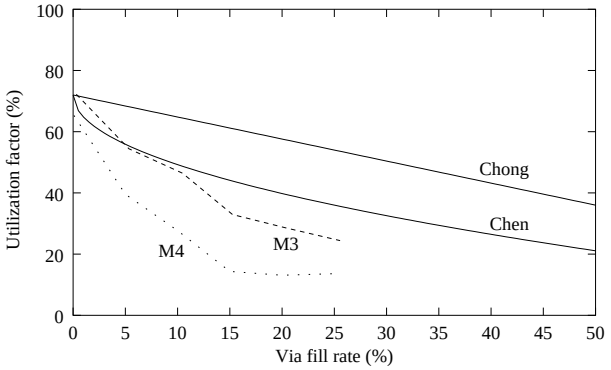

Fig. 3. Utilization factor for large via fill rate values: experimental results for `M3` (dashed line) and `M4` (dotted line) versus reviewed models (solid lines).

For the congested design, the utilization factor on each layer is represented by an x in Figure 2.[4] All x points (except for the one for `M3`, to which value the routing efficiency of 72% was tuned) are far from the model predictions.

3. To see how the utilization rate varies with the number of vias, we extend the experiment by adding *virtual vias* on track intersections.[5] The virtual vias mimic the effect of additional wires that are routed on virtual upper layers. Since blocking track intersections on `M1` and `M2` can potentially cause a net terminal (i.e., pin) to be blocked (thus preventing the router from finding any solution), we do not add virtual vias on `M1` and `M2`. For each number of virtual vias, we apply the same approach of gradually removing randomly chosen nets and rerouting the design until the partial netlist is just routable again.

Results for the extended experiment are plotted as solid lines in Figure 2 for `M3` and `M4`. While the addition of virtual vias mimics the behavior of the router for higher numbers of layers, the actual number of such higher layers is unknown. Thus, the model of Sai-Halasz can be checked only against the original congested result (without virtual vias). This comparison does not show a close match in Figure 2. The Chen model follows experiment data well for `M3`, but not for any other layer. Thus, Figure 2 shows that (i) no model accurately predicts the utilization factor on all layers, even though we tuned the routing efficiency to fit the experiments and (ii) no model correctly predicts the relationship between via impact and the number of vias. Section IV investigates the reasons for this.

The differences between model predictions and experimental results are especially worrisome if we recall that the primary purpose of these models is to predict the number of routing layers required by (future) designs. In the literature, the models that we have reviewed have been used to make claims on the limits on layer number or chip size in future VLSI systems. Increasing the number of layers dramatically and shrinking the die size, while acknowledging that wire sizes cannot shrink as much, results in a dramatic increase of the via fill rate on all layers. Figure 3 compares

[4]The value for `M1` was too low (2.66%) to be plotted.

[5]This is achieved by defining a LEF macro with via-shaped obstructions, and superposing the macro onto the original core region.

TABLE I

NUMBER OF TERMINAL VIAS PREDICTED BY CHONG AND BY CHEN, COMPARED TO THE EXPERIMENTALLY MEASURED NUMBER. THE LAST LINE IN THE TABLE SHOWS THE ESTIMATED AND REAL NUMBER OF LAYERS NEEDED TO ROUTE THE DESIGN.

Layer	Chong	Chen	Experiment
M1	71266	30973	113452
M2	27562	0	23585
M3	0	0	9894
M4	0	0	0
Total	98828	30973	146931
Layers needed	2	2	4

the experiment data with a very high number of virtual vias on M3 and M4 to the predictions by Chong and Chen.[6] We see that the via impact is significantly underestimated by both models. The real limits on number of layers and chip size will therefore be much more stringent than the models currently predict.

Finally, our experimental validation of the models not only adjusts the routing efficiency factor to better fit the experimental values but, more importantly, applies the via impact models of Chong and Chen to the *actual* number of terminal vias instead of the estimated number. In Table I, the number of terminal vias predicted by the Chong and Chen layer assignment models is compared to the actual number for the original experiment (no virtual vias). The difference between the (otherwise similar) layer assignment models of Chong and Chen is that Chong includes the terminal vias on the layer the wire is connected to (although they do not really add to the blockage) whereas Chen only counts vias that go through the layer. The number of terminal vias clearly is also underestimated by both models. Applying the models to estimate the number of layers needed, the combined effects of underestimating the number of terminal vias and underestimating the impact of each of these vias have large consequences. Both Chong and Chen predict that the design will be routable in two layers, while it is barely routable in four (since we have assured a congested design)!

IV. EXPERIMENTAL ANALYSIS OF THE ASSUMPTIONS OF EXISTING MODELS

If the experimental validation of the result of a model reveals that the model is not correct (as is the case for all of the reviewed models), one can try to experimentally verify the assumptions that lead to the result. Let us recall the main assumptions made by the various models:

1. The routing efficiency is constant over all layers (its value is assumed to be 40% by Sai-Halasz, 65% is used in examples by Chong, and Chen reports values between 40% and 66%).

2. The via impact is a constant factor (12% to 15%) of the available space on the upper layer (for Sai-Halasz' model and equal wire pitches).

3. The via impact models of Chong and Chen depend on the number of terminal vias and the assumptions that (a) interconnects are routed on a single tier (layer pair) and (b) shorter interconnects are routed on lower tiers.

4. The via impact is linear (Chong) in or increases with the square root of (Chen) the number of terminal vias.

In this section, we review these assumptions in detail.

A. Routing Efficiency

A.1 Routing Efficiency is Constant?

If the routing efficiency and signal net fraction are constant over all layers, then the utilization factor should monotonically increase with the layer number. Indeed, in our experiments where the wire pitches are the same on all four layers, the number of terminal vias is always larger on lower layers (see Table I). The via impact thus decreases with the layer number and applying Equation 2 results in an increasing utilization factor. However, Figure 2 supports this reasoning only for M1 through M3. The top layer (M4) actually accommodates less wire length than M3 although there is no via impact on M4.

A naive explanation is that the design is not fully congested, and hence not all available space on the top layer was used. However, our experiments force the design to be fully congested. The real explanation is that the congestion differs for different layers. Two effects cause this: (i) M1 can only be used for signal routing for a small amount of its total track length because of pin blockage and M1 features in cell layouts, and (ii) the via impact is higher for the layers on the bottom of the layer stack. Figure 4 shows the actual length that can be used on the layers because of these two effects, for a hypothetical four-layer design (a) and five-layer design (b). The letters H and V indicate the routing direction (horizontal or vertical) for each layer. In the four-layer design of Figure 4(a), every V-layer has a higher utilization rate than the H-layer beneath it. If we assume that wrong-way routing is prohibited and that the length needed in each direction is equal, this implies that the H-layers will be fully congested, but the V-layers will still have space left. Adding a fifth layer has the opposite effect. In Figure 4(b), the V-layers will dominate the congestion. This analysis shows that the routing direction of the second topmost layer always dominates the congestion, whatever the number of layers is (under the assumption of alternate routing directions).

Such results seem to indicate that any accurate model should introduce a different routing efficiency for each direction. Another option is to take advantage of the difference in available routing space. Since the minimum required length in each direction is fixed by the placement, one could guide the placement such that the required length is balanced over the directions as predicted by the model. Or, one could force the router to make most of the unavoidable detours in the less congested direction.[7] Again,

[6] Again, no comparison to Sai-Halasz' model is possible because we do not know the number of virtual layers introduced.

[7] Certainly, modern place-and-route tools are aware of such considerations. Our point is that models of achievable routing need similar

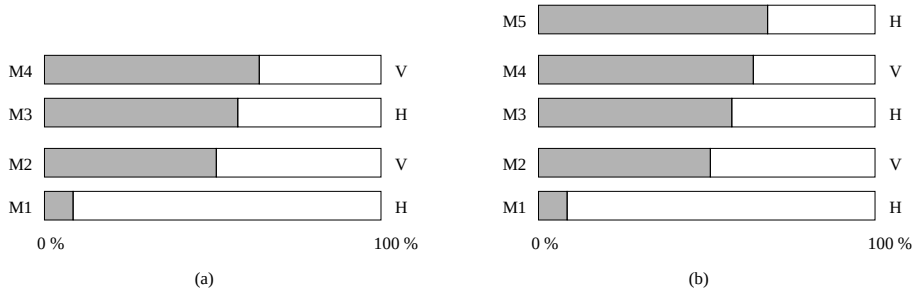

Fig. 4. Difference between the available track length in the horizontal and vertical directions. The second topmost layer defines the direction which is most congested.

guiding the layout tools to a "desired solution" is only feasible if the desired solution is obtained through an accurate model of via impact.

A.2 There is More Than Routing Efficiency Alone

In Section II, we noted that the routing efficiency represents various effects that reduce the total available routing space. Some effects are dependent only on the netlist, some depend on the placement, and some are related to the efficiency of the router. Since designs, placement tools and routing tools can be freely combined, it is important to distinguish between those effects. We therefore propose to decompose the current routing efficiency factor into three separate factors

$$\eta_r = \eta_n \, \eta_p \, \eta_r', \tag{4}$$

where η_n covers the routing space reduction due to the netlist (for an optimal placement and routing), η_p the reduction because of the quality of the placement tool, and η_r' the real routing efficiency.

A.3 Routing Efficiency or Routing Inefficiency?

Even more fundamental questions are raised by the counterintuitive definition of routing efficiency. Indeed, consider the following thought experiment:

1. Consider a given placement of a given netlist.
2. First route this design with a very good router.
3. Then route the same design with a very bad router.
4. Measure the resulting utilization factor for both routers.

Clearly, the routing efficiency of the bad router should be much lower than that for the good router. However, actual wire lengths will of course be longer for the bad router since it will make more detours. According to previous models, the "routing efficiency" is higher for the bad router! The problem is that the routing efficiency factor as defined in previous works does not really model the efficiency of the router, but rather its ability to fill whatever space it has, even if that is done by making unnecessary detours.

We therefore propose to define the routing efficiency factor based on the routing space *that is used efficiently*. This can be easily done by measuring (and modeling) the utilization rate based not on the actual length, but on the shortest possible length (the minimum Steiner tree length defined by the terminal locations). Hence, the utilization factor should be defined as

$$U_i = \frac{SL_i}{TL_i}, \tag{5}$$

instead of

$$U_i = \frac{AL_i}{TL_i}, \tag{6}$$

with SL_i the minimum Steiner tree length of all successfully routed nets on layer i, AL_i the actual routed length on layer i, and TL_i the supplied track length on layer i. With this definition, the routing efficiency of a bad router is lower than that of a good router because in a congested design a bad router is not able to route as many nets as a good one.

B. Congestion

The models reviewed in this paper (implicitly) assume a fully congested design. The discussion in the previous subsection invalidates this assumption unless the placement and routing tools can be tuned to obtain a fully congested design on all layers at the same time. Even if we could tune the layout tools, a model for the amount of congestion is still needed. Although the fully congested prediction is necessary to find the minimum number of layers needed to route a given design, the routing space that those layers provide will almost never be fully used (and if the prediction is such that we are balancing between N_l and $N_l + 1$ layers, we should probably opt for $N_l + 1$ layers to make sure the design is routable). Hence, the real routing will not be as congested as the predicted routing.

Predictions for designs that are not fully congested could leave unused space at the topmost layer, which lowers the number of terminal vias required to connect wires to that layer and hence creates more space on lower layers too. A model that accounts for the amount of congestion would probably also guide the layout tools to a solution that divides congestion problems equally among layers.

To assess the effects of congestion, we consider a fully congested design, routed on four layers, and gradually remove wires. Since a congested design requires more detours, the actual length decreases much more rapidly than the Steiner length when the wires are removed and the congestion is lowered. This is illustrated in Figure 5. When the design is not at all congested anymore, the actual length follows the Steiner length very closely.[8] Since the actual

awareness.

[8]The fact that the actual length gets even lower than the Steiner

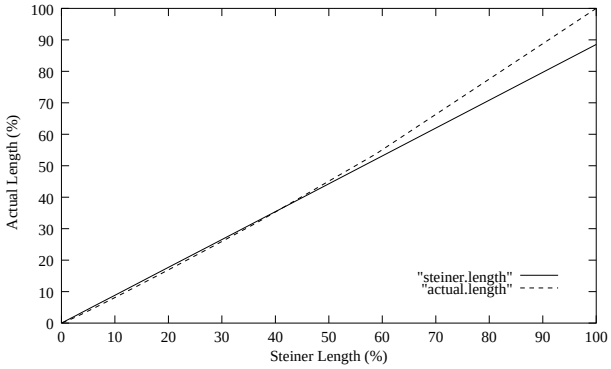

Fig. 5. The actual wire length versus the Steiner length for fully congested and less congested designs

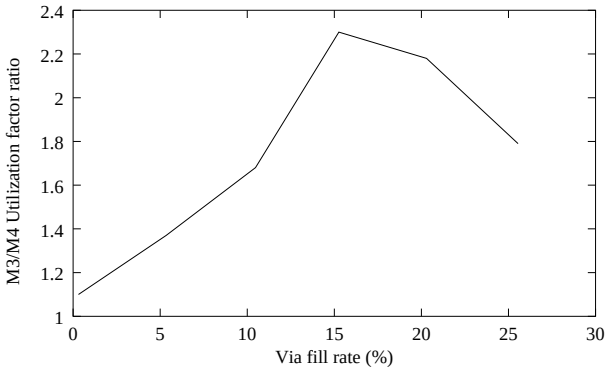

Fig. 6. The utilization factor at M3 relative to that at M4 for the experiments with virtual vias. A higher number of virtual vias (higher via fill rate) corresponds to a higher layer stack. Utilization ratios are certainly not constant.

TABLE II

THE UTILIZATION FACTOR U_i ON LAYER i, RELATIVE TO U_{i+1}: SAI-HALASZ' MODEL OF CONSTANT RELATIVE FACTORS VERSUS EXPERIMENTAL VALUES THAT ARE NOT CONSTANT.

Layer	Sai-Halasz	$\frac{U_i}{U_{i+1}}$	Only M3	$\frac{U_3}{U_4}$
M1/M2	0.85	0.07	min	1.10
M2/M3	0.85	0.56	avg	1.74
M3/M4	0.85	1.10	max	2.30

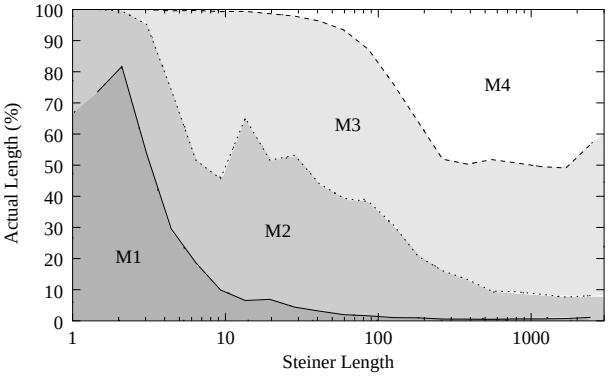

Fig. 7. The distribution of lengths over the layers for point-to-point connections of various lengths.

length starts to rise much faster than the Steiner length when the design becomes congested, it is difficult to model congestion accurately. However, such a congestion model is necessary if we want to use models of achievable routing as guides for layout tools.

C. A Constant Via Impact Factor

If Sai-Halasz' assumption of a constant via impact factor on all layers is true, then the utilization factor of layer i relative to that of layer $i+1$ should be a constant. Table II shows this relative utilization factor for the experiments presented in Figure 2. The result of the experiment without virtual vias is compared to Sai-Halasz' model in the left part of the table. The ratio of utilization factors is obviously not the same for all layers. The ratio is so low for M1/M2 because M1 is largely blocked by the cell pins. The factor for M3/M4 is larger than 1 because the top layer is probably not fully utilized, as discussed earlier. A similar reason (M2 is also underutilized) causes a low value for M2/M3. Such effects are not included in Sai-Halasz' model. The right part of Table II presents the results for all experiments (with virtual vias), only for M3/M4 (since no virtual vias were added on the other layers), and shows the minimum, average and maximum value of the ratio. Even if we observe the results for the same layers but for different via fill rates, the ratio is certainly not a constant, invalidating Sai-Halasz' basic assumption.

length is due to (i) our Steiner length approximation (we used the Batched Iterated 1-Steiner implementation for Steiner tree estimation from the University of Virginia [13]) and (ii) to the fact that Steiner lengths are measured from the center of the bounding box for all gate pins that are connected to the same net, whereas the actual net only connects to the closest one ("group Steiner" problem [2]).

The relation between the relative utilization factors at layers M3 and M4 and the number of terminal vias on M4 (including virtual vias) is shown in Figure 6. (The figure also represents the relative utilization factors of M3 and M4, for an increasing number of virtual layers.) The ratio of utilization factors increases with the via fill rate, which means that the utilization factor for M4 decreases more rapidly than that for M3 (until it saturates). This seems to indicate that with high via fill rates, the router is no longer able to connect wires to the top layer. None of the existing models is able to predict this (note that the via fill rates on M3 and M4 are almost the same in our experiment because the number of virtual vias is much larger than the original number of vias, hence both Chong and Chen predict the relative utilization factor to be very close to 1).

D. Interconnects on a Single Tier and Shorter Interconnects on Lower Tiers?

In Figure 7, we experimentally test the two assumptions of the layer assignment model of both Chong and Chen. The figure shows the percentage of the length of point-to-point connections that is routed on each layer as a function of the total length. The assumption that shorter wires are generally routed on lower layers seems to be (roughly) validated. However, the figure also shows that more than two layers are used for routing the nets of a single length.

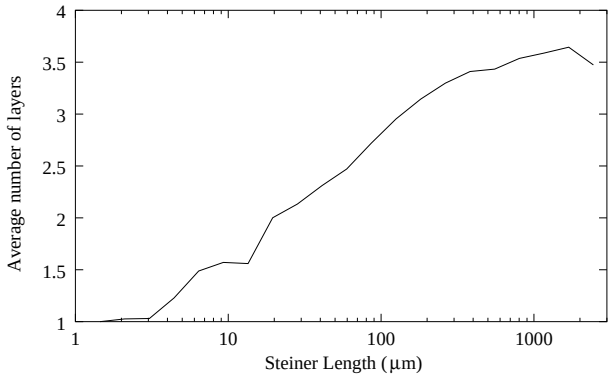

Fig. 8. The average number of layers used for single point-to-point connections as a function of their length. Long wires use more than two layers.

This is not only due to the fact that different nets of the same length are routed on different layers. Indeed, Figure 8 shows the average number of layers used for routing the nets as a function of their total length. Quite naturally, the very short wires are routed on a single tier (or even a single layer) but this no longer holds for the longer wires.[9]

The impact of routing interconnects on several layers is mainly that terminal vias are exchanged for turn vias. Indeed, wires are not connected straight to higher layers (with a stack of terminal vias) but with stops along all intermediate layers (using turn vias). This, of course, can have a tremendous effect on the via impact models that are based on the number of terminal vias. Moreover, the assumption that turn vias do not harm the routing solution becomes questionable if a single net turns too often.

Let us define a track *segment* on layer i such that (i) its length is equal to the wire pitch at layer $i-1$ and (ii) its midpoint is an intersection of tracks at layers i and $i-1$. Note that the number of segments in one track at layer i equals the number of tracks at layer $i-1$. With this definition, every turn in a wire uses one track segment on both layers, increasing the number of used track segments from $T = \ell + 1$ for a straight line of length ℓ segments to $T = \ell + v + 1$ if it uses v turn vias. This effect should also be taken into account.

E. Relation Between Via Impact and Number of Vias

The results of Figure 2 show that even if the via fill rates are the same, the curves for different layers do not coincide. One reason is the different routing efficiency factors. However, from Figure 6 we can deduce that even an adjustment in the routing efficiency factor could only cause the curves to overlap in a very small region of f. Since the (virtual) via fill rate corresponds to different layers, this leads us to the conclusion that the via impact factor is also layer dependent and that this dependency cannot be explained by the difference in the *number* of terminal vias alone, as the models of Chong and Chen assume. We believe that the major problem in their models is the fact that they do not capture the real wiring effects that are caused by via blockages.

Whenever a via blocks the path of a wire, it either has to be rerouted (probably with a detour) to a totally different location, or it can just be routed around the via. The latter solution creates a (larger) blockage for wires in the adjacent track and this leads to the "cascade" (or "ripple") effect. Chong's model does not consider this effect, and Chen's assumes that the blockage caused by this effect can be modeled by assuming a track blocked by dense vias simply cannot be used over its entire length. A better understanding of the impact of the ripple effect (although this is certainly not straightforward) is necessary to model it more accurately.

V. Toward a New Model of Achievable Routing

Based on the observations of Section IV we suggest new additions to the models of achievable routing that take care of many of the deficiencies found. Like the models described in Section II, our new model is aimed at fully congested designs (it does not contain a general congestion model as proposed in Section IV-B). Although our new model improves on existing models, we do not wish to claim this is the "end all solution." Further research is definitely needed and the suggestions made hereafter intend to put this research on the right tracks.

A. Routing Efficiency Model

We keep the definition of the routing efficiency as in the previous models, i.e., it contains effects of netlist, placement and routing. Since we investigate the models of achievable routing, we do not really need to split the routing efficiency factor into separate factors as proposed in Section IV-A. However, we make sure that the *efficiency* of the router, rather than the inefficiency, is measured by using Steiner lengths instead of actual lengths. Measuring this routing efficiency from experiments and using it in the models ensures (by definition) that router dependent effects are not observed (because they are all taken into account by the routing efficiency value, specific for the layout tool used).

The difference in congestion between `H` and `V` layers is taken into account by (i) introducing a *fill factor* for the non-congested routing direction that models the fact that not all available space in this direction can be used and (ii) changing the tier (layer pair) model in such a way that `H` and `V` layers of the same tier can have a different utilization. Therefore, we define tiers as (overlapping) layer pairs of neighboring layers such that tier 1 combines M1 and M2, tier 2 combines M2 and M3, and so on. With this, M1 and M2 (e.g.) can have very different utilization rates because M2 is also part of a tier that contains M3. We assume that wires routed on a tier have, in principle, the same length on both layers of the tier (thus obeying the requirement – or assumption – that the total utilization factor for all `H` layers together is equal to that for all `V` layers together). However, we acknowledge that real routers allow wrong-way routing and that the space available in the non-congested layers

[9]We did not investigate how extra vias or layer usage resulted from antenna routing rules. This may be necessary in future models.

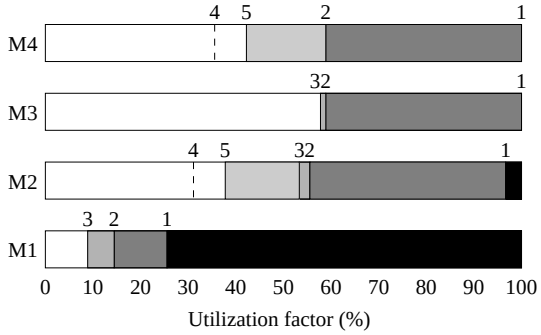

Fig. 9. Factors that reduce the available routing space: (1) signal net ratio and space unavailable due to cell terminals (on M1), (2) routing efficiency, (3) via impact, (4) fill factor and (5) H-V factor (reduces the fill factor and hence increases the utilization factor for non-congested layers). Layers M1 and M3 are fully congested and therefore have no fill factor.

therefore can be used more efficiently than if wrong-way routing was not allowed. The fill factor is thus adjusted to allow a larger part of the wires on the non-congested layers by taking into account a factor we call the *H-V factor*.[10] The different factors in our model are shown in Figure 9.

B. Layer Assignment Model

We retain the assumption that shorter interconnects are routed on lower layers but we consider the fact that (longer) wires occupy more than a single tier by splitting wires into segments and assigning the segments (rather than the wires) to a specific tier based on the segment length (rather than the total wire length). The segments are defined as Steiner segments, i.e., segments that connect a pin to a Steiner point or to another pin. The Steiner tree is constructed by building an MST, then instantiating connections shortest-first. This emulates the order in which connections are typically implemented in batch gridded routers (shortest connections are made between subtrees using, e.g., bidirectional A* routing, until there is a single tree for the net). We use the derivation of Borah et al. [3] for constructing a Steiner tree from a MST. The order for Steiner tree construction has been changed into a shortest-net first order.

Splitting wires into segments has the additional advantage that the influence of turn vias on the via impact no longer has to be taken into account since all turns (except for those between adjacent H and V layers – which do not impact other wires) are converted into segment terminals.

C. New Via Impact Model

Another observation is that the effects of blockages on wires should differ for different wire lengths. Indeed, the probability that a wire is blocked should be monotone in the number of track intersections it must cross as well as in the probability that a via blocks one of those intersections. If the vias are uniformly distributed over the area (which we assume), the via fill rate f equals the probability that an intersection (or wire segment) is blocked by a via. Since any wire segment of length ℓ track segments occupies $\ell+1$ track intersections, the probability P_1 that a specific route for the wire segment is **not** blocked can be estimated as

$$P_1 = (1-f)^{\ell+1} \tag{7}$$

because none of the intersections may contain a via. Of course, there are several possible routes between any two points and the probabilities that they are blocked are not independent. In general, if $N_r(\ell)$ routes are possible then the probability that at least one of them is not blocked is given by

$$\begin{aligned} P_{nb} = &\sum_{n_1=1}^{N_r(\ell)} P_1(n_1) - \sum_{n_1=1}^{N_r(\ell)} \sum_{n_2=n_1+1}^{N_r(\ell)} P_2(n_1,n_2) \\ &+ \sum_{n_1=1}^{N_r(\ell)} \sum_{n_2=n_1+1}^{N_r(\ell)} \sum_{n_3=n_2+1}^{N_r(\ell)} P_3(n_1,n_2,n_3) - \ldots \\ &(-1)^{N_r(\ell)-1} \sum_{n_1=1}^{N_r(\ell)} \sum_{n_2=n_1+1}^{N_r(\ell)} \cdots \sum_{n_{N_r}=N_r(\ell)}^{N_r(\ell)} P_{N_r}(n_1,\ldots,n_{N_r}) \end{aligned} \tag{8}$$

where $P_i(n_1,\ldots,n_i)$ is the probability that no routes with indices n_1 through n_i are blocked. Because wires can overlap, the number of intersections that have to be free of vias is different for different combinations of routes, and so is the probability P_i. The first term in Equation 8 is the probability that each of the possible routes is free of vias (separately) and the other terms are needed to prevent double counting the cases where two (or more) routes are free of vias.[11]

The enumeration in Equation 8 proves to be quite hard and we have not yet managed to find a closed form expression as a function of f and ℓ. A huge simplification lies in assuming that only two possible routes are allowed: the upper and lower L-shaped route. In this case, Equation 8 can be written as

$$P_{nb} = (P_l + P_u - P_b) \tag{9}$$

where P_l (P_u) is the probability that the lower (upper) L route is not blocked and P_b is the probability that the entire bounding box is not blocked by vias. Because we assume that the probabilities that different track intersections are blocked are independent and equal to the via fill rate f, we have

$$P_l = (1-f)^{\ell+1} \tag{10}$$

$$P_u = (1-f)^{\ell+1} \tag{11}$$

$$P_b = (1-f)^{2\ell} \tag{12}$$

and hence[12]

$$P_{nb} = 2(1-f)^{\ell+1} - (1-f)^{2\ell}. \tag{13}$$

[10] In our model this factor is measured from the experiments.

[11] Suppose there are 4 possible routes and assign a 0 to routes that are not blocked by a certain via combination and a 1 to all routes that are blocked. Assume that all combinations of 0's and 1's for the 4 routes occur exactly once (16 combinations). We need to know the number of cases with one or more 0's (15). The first term in Equation 8 adds up the cases with a 0 for at least one route (4x8=32 cases) and hence greatly overestimates the number of possibilities that are free of vias (underestimates the via impact).

[12] One can show that the probabilities that the wire segment is blocked and that it is not blocked add up to 1.

```
 1. Assume first that there is no via impact
 2. Repeat
 3. {
 4.    Calculate, for each layer, the utilization factor
          as in Equation 2.
 5.    Calculate the fill factor for the non-congested
          direction
 6.    Multiply the utilization factor for non-congested
          layers by the fill factor
 7.    For all layers starting from M1
 8.    {
 9.       Repeat
10.       {
11.          Fill the layer with the remaining wire segments
                from the previous tier
12.          Fill the layer with its share of the shortest
                wire segments until it is full
13.          Calculate the via blockage for all wire
                segments on the layer
14.          Recalculate the utilization factor using the
                new via impact factor
15.       } until the via impact factor does not change more
             than 0.1%
16.    }
17. } until the utilization factors for all layers do not
       change more than 0.1%
```

Fig. 10. Pseudo code of our new model for achievable routing.

If we retain the assumption of other models that a wire (segment) cannot be routed if its shortest path is blocked (i.e., we do not allow ripple effects), the via impact factor for a segment of length ℓ could be estimated to be

$$v_i(l) = 1 - P_{nb} = 1 - 2(1-f)^{\ell+1} + (1-f)^{2\ell}. \qquad (14)$$

It should be noted that the restriction of the number of possible routes to only 2 is very restrictive[13] and that it will undoubtedly lead to an overestimation of the via impact, especially for longer segments.

D. Model for Achievable Routing

Our new model for achievable routing is presented (in pseudo code) in Figure 10. It takes as input a list of all segments of Steiner trees (i.e., their individual lengths) or, alternatively, a list of segments from an a priori wire length estimation model, the supplied track length on each layer, the amount of track segments unavailable for signals due to power and ground routing (the signal fraction s_i per layer), the routing efficiencies per layer,[14] the number of cell pins on M1, the fraction of all wires that is routed on H-layers (H-V fraction)[15] and the number of terminal vias measured on each tier (between the layers of the tier).

In line 11 of the pseudo code each layer inherits the wire segments that were left over from the assignment to the tier for which this layer is the upper layer (none for M1). In line 12 the layer receives its share of wire segments assigned to the tier for which this layer is the lower layer, until the layer is full. For each segment length, the fraction of wire segments to be assigned to this layer is defined by the H-V factor. The remaining part will then, in the next iteration of the for-loop, be assigned to the upper layer of this tier. In the for-loop, the shortest wire segments are assigned to the lowest layers (they are chosen first).

[13]The "natural" extension to also allow Z-shaped routes already is much more complicated.

[14]We measured these values for the experiment without virtual vias and used them to estimate the via impact for cases with virtual vias. Except for the top layer, the utilization factor of all layers is also influenced by the via impact factor, which is unknown. Therefore we used our model to estimate the via impact for the case without virtual vias and adjusted the utilization factors with these estimates to obtain routing efficiency values.

[15]Without loss of generality, we assume here that the routing direction of M1 is horizontal.

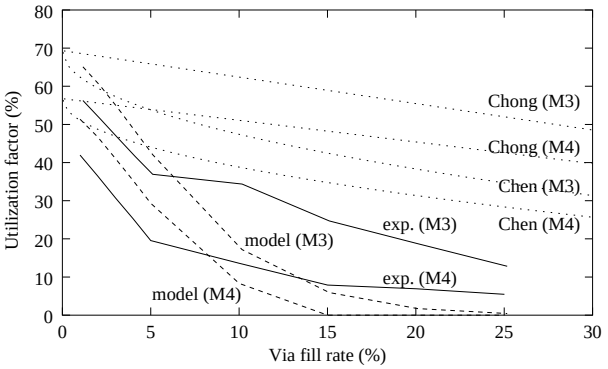

Fig. 11. Utilization factor for large via fill rate values: experimental results (solid lines) versus reviewed models (dotted lines) and our new model (dashed lines) for M3 and M4.

The model still needs to be extended to include a lot more possible routes to calculate the via impact.[16] Figure 11 shows the results of the simplified model (only L-shaped routing) for the same design as we used in Section III, with the same experiments (adding virtual vias and measuring the utilization factor). As expected, our model underestimates the utilization factor for longer wires (higher layers) because of the restrictive assumption that only L-shaped routes are allowed. A few experiments with other designs showed very similar results. Our future work therefore focuses on the inclusion of more possible routes so that the via impact reduces and the results should be much closer to the actual measured results. However, the effect of the different wire lengths on M3 and M4 (for the same via fill rates) is, for the first time, acknowledged by our new model.

In our model, we use the actual (measured) amount of terminal vias on each layer to compute the via fill rate f (and thus, indirectly, the via impact). Although estimates of the number of terminal vias might be available from previous routing results, we should be able to estimate the number of terminal vias in the model itself. However, not all wire segments lead to two stacks of terminal vias all the way down to M1 because segments can be connected to other segments on the same or neighboring layers. In order to model the number of terminal vias, we should keep track of the connections between the segments inside the model and use the layer assignments of connected wire segments to estimate the number of terminal vias that are needed. We plan to include this estimation in future extensions to our model. Also, we used actual placement information to

[16]Because of the large discrepancies for longer wires we limited the segment length to 20 track segments to calculate the via impact.

find the distribution of wire segment lengths. This distribution can also be found a priori, i.e., before placement, by using wire length estimation techniques [7], [8], [10], [20] and therefore our model of achievable routing is more generally applicable than constructive approaches based on global routing. As noted before, we did not use these a priori estimates here because we intended to investigate only models of available routing resources and not models of required routing resources.

VI. Conclusion

Experimental verification of achievable routing models is needed in order to accept such models. Highly accurate models are especially needed for interconnect process optimization, matching interconnect resources to individual designs at early design stages, or guiding layout tools to solutions predicted by the models. In this paper, we have presented a way to experimentally analyze models of achievable routing, and applied the analysis to three existing models. None of the models seems to be accurate. We have investigated the reasons for the deviations between experimental results and identified issues that need to be addressed before an accurate model can be expected.

Based on our study of previous models, we have suggested a new model of achievable routing. Whereas previous models largely underestimate the via impact and only take the number of terminal vias into account, our new model tries to capture the effect of the wire length of the wire segments routed on a particular layer. The results show that this model is able to account for the difference in layers, even if the number of terminal vias is equal. However, our model overestimates the via impact. We have argued that this is due to restricting the number of possible routes for wire segments to L-shapes only. The inclusion of many more possible routes undoubtedly will improve the results but the enumeration that goes with it remains unsolved to date. Future work will focus on this problem.

References

[1] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley Publishing Co., 1990.

[2] C. D. Bateman, C. H. Helvig, G. Robins and A. Zelikovsky, "Provably-Good Routing Tree Construction with Multi-Port Terminals," *Proc. ACM/SIGDA Intl. Symp. on Physical Design*, 1997, pp. 96–102.

[3] M. Borah, R. M. Owens and M. J. Irwin, "An Edge-Based Heuristic for Steiner Routing," *IEEE Trans. on Computer Aided Design*, vol. 13 (12), Dec 1994, pp. 1563–1568.

[4] A. E. Caldwell, Y. Cao, A. B. Kahng, F. Koushanfar, H. Lu, I. L. Markov, M. R. Oliver, D. Stroobandt and D. Sylvester, "GTX: The MARCO GSRC Technology Extrapolation System," *Proc. ACM/IEEE Design Automation Conf.*, 2000, pp. 693-698. See: `http://vlsicad.cs.ucla.edu/GSRC/GTX/`.

[5] Q. Chen, J. Davis, P. Zarkesh-Ha and J. Meindl, "A Compact Physical Via Blockage Model," to be published in *IEEE Trans. on VLSI Systems, Special Issue on System-Level Interconnect Prediction*, December 2000.

[6] P. Chong and R. K. Brayton, "Estimating and Optimizing Routing Utilization in DSM Design," *Workshop notes Intl. Workshop on System-Level Interconnect Prediction*, D. Stroobandt and A. B. Kahng, editors, 1999, pp. 97–102.

[7] P. Christie and D. Stroobandt, "The Interpretation and Application of Rent's Rule," to be published in *IEEE Trans. on VLSI Systems, Special Issue on System-Level Interconnect Prediction*, December 2000.

[8] 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*, vol. 45(3), 1998, pp. 580–589.

[9] W. E. Donath, "Placement and Average Interconnection Lengths of Computer Logic," *IEEE Trans. on Circuits & Syst.*, vol. CAS–26, 1979, pp. 272–277.

[10] W. E. Donath, "Wire Length Distribution for Placements of Computer Logic," *IBM J. of Research and Development*, vol. 25, 1981, pp. 152–155.

[11] J. C. Eble, V. K. De, D. S. Wills and J. D. Meindl, "A Generic System Simulator (GENESYS) for ASIC Technology and Architecture Beyond 2001," *Proc. 9th Annual IEEE Intl. ASIC Conf.*, 1996, pp. 193–196.

[12] B. M. Geuskens, "Modeling the Influence of Multilevel Interconnect on Chip Performance," *Ph.D. thesis*, Rensselaer Polytechnic Institute, 1997. See: `http://latte.cie.rpi.edu/ripe.html`.

[13] J. Griffith, G. Robins, J. S. Salowe and T. Zhang, "Closing the Gap: Near-Optimal Steiner Trees in Polynomial Time," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 13(11), 1994, pp. 1351–1365.

[14] A. B. Kahng, S. Mantik and D. Stroobandt, "Requirements for Models of Achievable Routing," *Proc. ACM Intl. Symp. on Physical Design*, 2000, pp. 4–11.

[15] A. B. Kahng and D. Stroobandt, "Wiring Layer Assignments with Consistent Stage Delays," *Proc. ACM Workshop on System-Level Interconnect Prediction*, 2000, pp. 115–122.

[16] B. S. Landman and R. L. Russo, "On a Pin Versus Block Relationship for Partitions of Logic Graphs," *IEEE Trans. on Computer*, vol. C–20, 1971, pp. 1469–1479.

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

[18] G. A. Sai-Halasz, "Performance Trends in High-Performance Processors," *Proc. of IEEE*, 1995, pp. 20–36.

[19] L. Scheffer and E. Nequist, "Why Interconnect Prediction Doesn't Work," *Proc. ACM Workshop on System-Level Interconnect Prediction*, 2000, pp. 139–144.

[20] 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*, vol. 10, 1999, pp. 1–20.

[21] D. Sylvester and K. Keutzer, "Getting to the Bottom of Deep Submicron," *Proc. Intl. Conf. on CAD*, 1998, pp. 203–211. See: `http://www.eecs.berkeley.edu/~dennis/bacpac/`.

Andrew B. Kahng ADD BIOGRAPHY, ADD BIOGRAPHY.

Stefanus Mantik ADD BIOGRAPHY, ADD BIOGRAPHY.

Dirk Stroobandt was born in Gent, Belgium, in 1972. He graduated in 1994 and obtained the PhD degree in electrotechnical engineering in 1998 from Ghent University. From 1994 to 1998, Dirk Stroobandt was research assistant and currently he is Post-doctoral Fellow with the Fund for Scientific Research - Flanders (Belgium) (F.W.O.). He is affiliated with the Department of Electronics and Information Systems (ELIS) of Ghent University. His research is oriented towards a priori estimations of interconnection lengths in electronic systems and its applications to CAD, computer architecture evaluation (especially optoelectronic threedimensional architectures) and design optimization.

Dirk Stroobandt visited the group of Andrew B. Kahng at the CS Department of UCLA as a post-doctoral researcher from July 1999 to June 2000. He is the inaugural winner of the ACM Outstanding Doctoral Thesis Award in Design Automation and co-organized (in 1999) and chaired (in 2000) the International Workshop on System-Level Interconnect Prediction (SLIP). He is guest editor of two special issues of the IEEE Transactions on VLSI Systems on System-Level Interconnect Prediction.