Requirements for Models of Achievable Routing [*]

Andrew B. Kahng
UCLA CS Dept.
3731 Boelter Hall
LA, CA 90095-1596 USA
abk@cs.ucla.edu

Stefanus Mantik
UCLA CS Dept.
3731 Boelter Hall
LA, CA 90095-1596 USA
stefanus@cs.ucla.edu

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

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 wirelength 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 within models of achievable routing.

Categories and Subject Descriptors

B.7.2 [**Hardware**]: Integrated Circuits— *Placement and routing*; C.4 [**Computer Systems Organization**]: Performance of Systems—*Modeling techniques*; J.6 [**Computer Applications**]: Computer-aided Engineering—*CAD*; F.2.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity—*Routing and layout*

General Terms

VLSI Routing Estimation, Via Impact Model, Interconnect Process Optimization.

1. 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 wirelength estimation model [8] based on Rent's rule [13] 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 wirelength *required* by the design ("demand") [7; 16], or the estimation of effectively *available* total track length ("supply") on chip [4; 5; 15]. 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. (Such predictions can be made post-placement based on actual pin locations, or else pre-placement based on models of wirelength distribution.) 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; 3; 10; 11; 15; 17].

Models of achievable routing can also 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 [5] (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 [14] 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.

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

†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.

Contributions of This Work

To effectively guide physical chip implementation, models for 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 wirelength, 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.*

Our work centers on (i) understanding the reasons for this validation gap, (ii) processes for model validation, and (iii) necessary improvements in future models of achievable routing. Section 2 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 3, 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 4, 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 conclude in Section 5 by proposing requirements for new models of achievable routing.

2. MODELS FOR ACHIEVABLE ROUTING

As noted above, all models for 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 wirelength distribution models [6] such as those of Donath [9], Davis et al. [7], or Stroobandt et al. [16]. 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.

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.

2.1 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/ground and 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

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

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

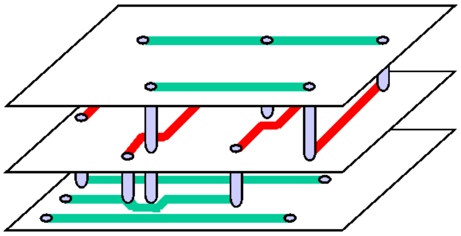

Figure 1: Wires have to make detours due to via blockage.

2.2 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 [15] 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 reduced 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 [10; 11; 17]. 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 [5], Chong and Brayton devised a model that 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 wirelength distribution as input (the wirelength distribution model of Davis et al. [7] 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. [4] 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 [4] 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.

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 [4] 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.

[2] The expression in [4] is slightly more complicated because the authors also include possibly different wire-to-wire and wire-to-via spacings.

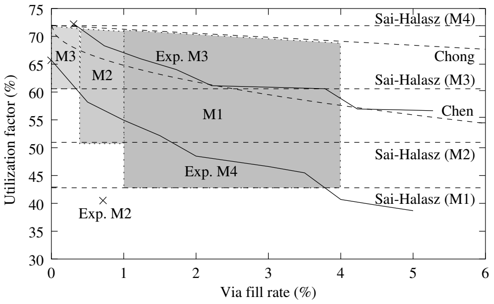

Figure 2: Utilization factor as a function of the via fill rate for the reviewed models (dashed lines) and experimental results (solid lines). The shaded regions represent the range of predictions across the three models for the different layers in a typical design.

3. MODEL VALIDATION

We emphatically 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, it is startling 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.

3.1 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.

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 [4]) 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, 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 an `M1` utilization factor of 20%, *a factor of three to four* less than the value predicted by Chong (!).

3.2 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`).

2. We ensure a congested design by removing the top layer, then gradually removing randomly chosen nets and rerouting the design 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.) A maximum routing efficiency value of 72% was found and applied in the Chong and Chen models.

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 did not add virtual vias on `M1` and `M2`.)

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 4 investigates the reasons for this.

The differences between model predictions and experimental results are especially appalling 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 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 severely 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 1, 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). Both Chong and Chen predict that the design will be routable in two layers, while it is barely routable in four (!) 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. Clearly, both layer assignment models and via impact models produce wrong results.

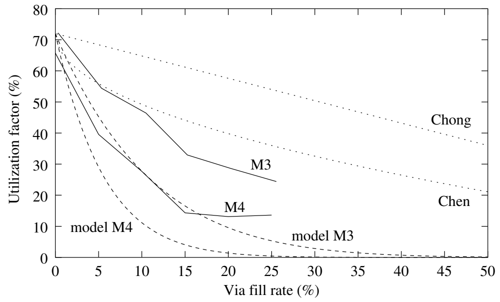

Figure 3: Utilization factor for large via fill rate values: experimental results (solid lines) versus reviewed models (dotted lines) and a simple model based on the probability that an average length wire is blocked by vias (dashed lines).

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

Table 1: Number of terminal vias predicted by Chong and by Chen, compared to the experimentally measured number.

4. 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

[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.

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

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

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

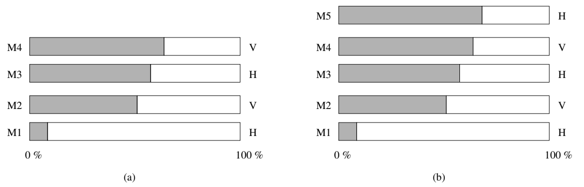

Figure 4: Difference between the available track length in the horizontal and vertical directions. The second topmost layer defines the direction which is most congested.

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.

4.1 Routing Efficiency

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 1). 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 wirelength 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 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.

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

There is More Than Routing Efficiency Alone

In Section 2, 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.

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 wirelengths 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.

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

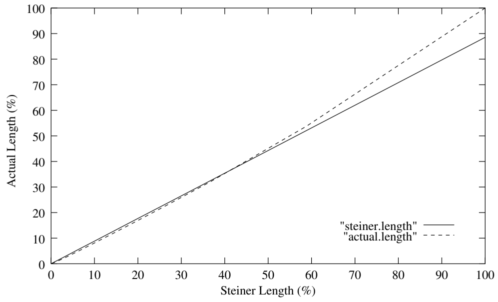

Figure 5: The actual wirelength versus the Steiner length for fully congested and less congested designs

4.2 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 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 for achievable routing as guides for layout tools.

4.3 A Constant Via Impact Factor

If Sai-Halasz' assumption is true, then the utilization factor of layer i relative to that of layer $i+1$ should be a constant. Table 2 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

[8]The fact that the actual length gets even lower than the Steiner 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 [12]) 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]).

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

Table 2: 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.

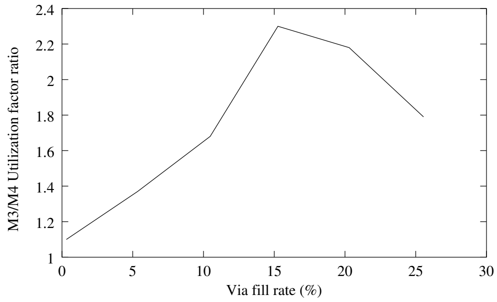

Figure 6: The utilization factor at `M3` relative to that at `M4` for the experiments with virtual vias. A higher number of virtual vias (higher via fill rate) corresponds to a higher layer stack. Utilization ratios are certainly not constant.

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 2 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.

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 the two highest virtual layers, 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 the 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, hence both Chong and Chen predict the relative utilization factor to be very close to 1).

4.4 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

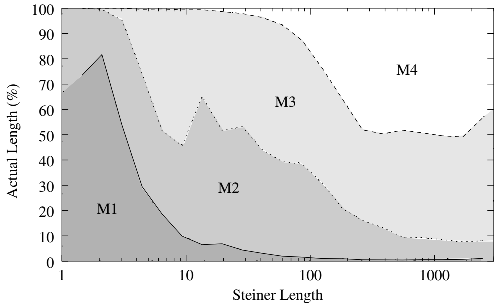

Figure 7: The distribution of lengths over the layers for point-to-point connections of various lengths.

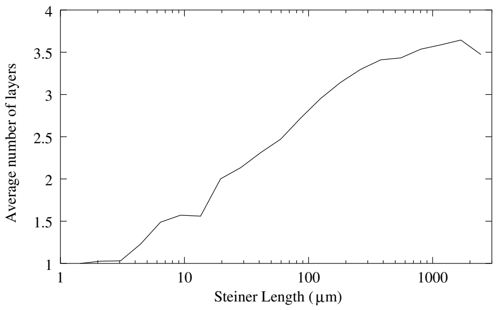

Figure 8: The average number of layers used for single point-to-point connections as a function of their length. Long wires use more than two layers.

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. 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 $T = l+1$ for a straight line of length l segments to $T = l + v + 1$ if it uses v turn vias. This effect should also be taken into account.

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

4.5 Relation Between the Via Impact and the 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.

Proposal of a Simple Model

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.

A second 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, the via fill rate f equals the probability that an intersection (or wire segment) is blocked by a via. Since any wire of length l segments occupies $l+1$ track intersections,[10] the probability P_{nb} that the wire is **not** blocked can be estimated as

$$P_{nb} = (1-f)^{l+1} \tag{7}$$

because none of the intersections may contain a via. If we retain the assumption of other models that a wire cannot be routed if its shortest path is blocked (i.e., we do not allow ripple effects), the via impact factor for a wire of length l could be estimated to be

$$v_i(l) = 1 - P_{nb} = 1 - (1-f)^{l+1}. \tag{8}$$

Of course, this new via impact model is far too simple. It does not even acknowledge that longer wires have more possibilities for finding a shorter route (which decreases $v_i(l)$). A comparison with the experimental results for the average wirelengths on each layer (Figure 3) indeed reveals that this simple model overestimates the via impact. However, the more interesting result is that this simple model seems to introduce the effects of the wirelength in a manner consistent with the general trend of the experimental results. Therefore, we believe that it is a good basis for future models, provided that many other effects can be accommodated as well.

[10]For simplicity, we leave out the previously described effect of turn vias.

5. FUTURE MODEL REQUIREMENTS

Based on the results of our analysis in the previous section, we can set the following basic requirements for future models for achievable routing, without claiming we have answers to all those issues yet.

1. A more thorough understanding of the routing efficiency factor. The following should be considered.

- A different efficiency factor for each layer that takes into account the congestion differences between horizontal and vertical layers. This might be done by introducing a *fill factor* for each layer that accounts for the amount of space that remains unused.
- The factorization of the routing efficiency factor as $\eta_r = \eta_n \, \eta_p \, \eta'_r$, 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. An accurate model of all those efficiency factors requires a thorough understanding of topological properties and influences (η_n), as well as of the behavior of placement (η_p) and routing (η'_r) tools.
- The use of the minimum Steiner tree length (based on terminal locations) for successfully routed nets, instead of the actual routed length, to model and measure the routing efficiency.

2. A model that introduces a factor that relates the increase in Steiner length to an increase in actual length and accounts for the amount of congestion.

3. A layer assignment model that takes into account more than one tier for long interconnects.

4. A study to assess the number of turn vias and their impact on the overall via impact factor.

5. A via impact model that really takes the effects on the wiring into account and considers

- interconnection lengths on the layer (our proposed simple model should be a good starting point for this);
- the impact of the ripple effect; and
- the effect of detours on the other wires.

Some of the requirements are, of course, more important than others. An ordering of the requirements is difficult. Based on our results, we believe that the most vital issues are those related to the routing efficiency factor and the fact that the via impact factor should better account for real wiring phenomena. The distribution of interconnects over more than 2 layers and the impact of turn vias are also worth investigating (note that the two problems are related!). A detailed congestion model is mainly needed for the calibration of achievable routing models to guide layout tools, not so much for predictions of the minimal number of layers. In the latter case, the congestion model could be another (constant) factor that relates the Steiner length to the actual length for a fully congested design.

Clearly, a great deal of research remains to be done. Our point is that any future models must be driven, validated and calibrated by real-world data.

6. CONCLUSION

Experimental verification of models for achievable routing 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. This leads us to a set of requirements for future models of achievable routing, as well as requirements for experimental validation of such models.

7. 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] 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/`.

[4] Q. Chen, J. Davis, P. Zarkesh-Ha and J. Meindl, "Via Impact and Via-Limited Chip Size", *Private communication*, Georgia Institute of Technology, 1999.

[5] 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.

[6] 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*, 2000.

[7] 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.

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

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

[10] 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.

[11] 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`.

[12] 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.

[13] 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.

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

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

[16] 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.

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