Interconnect Optimization Strategies for High-Performance VLSI Designs*

Andrew B. Kahng, Sudhakar Muddu, Egino Sarto

Silicon Graphics, Inc., Mountain View, CA 94039

{muddu,sarto}@mti.sgi.com, abk@cs.ucla.edu

Abstract

Interconnect tuning and repeater insertion are necessary to optimize interconnect delay, signal performance and integrity, and interconnect manufacturability and reliability. Repeater insertion in interconnects is an increasingly important element in the physical design of high-performance VLSI systems. By interconnect tuning, we refer to the selection of line thicknesses, widths and spacings in multi-layer interconnect to simultaneously optimize signal distribution, signal performance, signal integrity, and interconnect manufacturability and reliability. This is a key activity in most leading-edge design projects, but has received little attention in the literature. Our work provides the first technology-specific studies of interconnect tuning in the literature. We center on global wiring layers and interconnect tuning issues related to bus routing, repeater insertion, and choice of shielding/spacing rules for signal integrity and performance. We address four basic questions. (1) How should width and spacing be allocated to maximize performance for a given line pitch? (2) For a given line pitch, what criteria affect the optimal interval at which repeaters should be inserted into global interconnects? (3) Under what circumstances are shield wires the optimum technique for improving interconnect performance? (4) In global interconnect with repeaters, what other interconnect tuning is possible? Our study of question (4) demonstrates a new approach of offsetting repeater placements that can reduce worst-case cross-chip delays by over 30% in current technologies.

1 Introduction

With technology scaling, on-chip interconnect becomes an increasingly critical determinant of performance, manufacturability and reliability in high-end VLSI designs. Current and future designs are generally interconnect limited, and the available routing resource must be carefully balanced among signal distribution, power/ground distribution, and clock distribution. Table reproduces several technology projections from the 1997 SIA National Technology Roadmap for Semiconductors [1, 2].

The implications of technology scaling – particularly for system interconnect – are very complicated. Example considerations for a 7-layer metal (7LM) process might include:

- Local interconnect layers (e.g., M1-M3) should generally remain at near-minimum dimensions and pitch in order to achieve routing density (for an example analysis of interconnect density in forthcoming 0.25μm processes, again see [15]). For short lines (e.g., several hundred microns or less), thinner metal offers less lateral coupling capacitance and driver loading, and thus locally improves circuit performance. At the same time, maximum wire width is limited by the aspect ratio upper bound. The resulting thin and narrow wires are highly resistive and also subject to reliability concerns; they are hence unsuitable for global interconnects, power distribution, etc. We also note that layers M2-M3 (and maybe M4) will support a mix of local and "near-global" wiring, e.g., long wires within a single block. The distribution of lengths and performance goals for these signals can vary considerably between designs; since shorter wires are better routed on thinner metal, these design-specific considerations will affect the interconnect.
- Power distribution layers (e.g., M6-M7, maybe M5), which typically also support the top-level clock distribution (mesh or balanced tree), should be as thick as possible for reliability. IR drop and clock skew – as well as robustness under process variations – also suggest the use of thick wire on these layers. Thick wire additionally conserves area, but can suffer from increased lateral capacitive coupling.
- Global interconnect layers (e.g., M4-M6) support inter-block signal runs with length on the order of 3000μm - 15000μm. To satisfy delay and signal integrity constraints, at least three degrees of freedom are available: line width and spacing, repeater insertion, and shield wiring. Repeater insertion shields downstream capacitance and is the canonical means of converting "quadratic" *RC* delay into "near-linear" delay; this technique also improves edge rates and hence noise immunity. When lateral coupling capacitances are large, worst-case "Miller coupling" begins to dominate noise and delay calculations; this is alleviated by increasing the line spacing and/or adding shield wiring (i.e., wires connected to ground), with future techniques possibly including dedicated ground and power planes interleaved with signal layers [10]. Another technique to reduce the lateral coupling capacitance is to interleave signal lines which do not switch at the same signal transistion period. The bus-dominated nature of global interconnects in building-block and high-performance designs only worsens the effects of coupling, since it causes longer parallel runs.
- All layers are subject to mutual pitch-matching, via sizing, etc. considerations. Hence, available widths and spacings on one layer are not independent of the widths and spacings on a second layer.

The above are only a few of the applicable design considerations; the net effect is that balancing interconnect resources is now extremely difficult as designs move into the quarter-micron regime and beyond.

Interconnect Tunning Strategies

At the leading edge of performance, *interconnect tuning* has become a critical degree of freedom in system design. By interconnect tuning, we refer to the *selection by a design team* of line thicknesses, widths and spacings in multi-layer interconnect to simultaneously achieve: (i) distribution (available wiring density) for local signals, global signals, clock, power and ground; (ii) performance (signal propagation delay), particularly on global interconnects; (iii) noise immunity (signal integrity), again particularly on global interconnects; and (iv) manufacturability and reliability (e.g., required margins for AC self-heat or DC electromigration on interconnects, short-circuit power in attached devices, etc.). Today, interconnect tuning is a key activity in most leading-edge microprocessor projects. It is clearly an option whenever the design and fabrication are owned by a single entity; however, for high-volume projects even fabless design houses are exercising increasing influence on vendors' processes [15]. Nevertheless, this topic has received very little attention in the literature, with only a small handful of high-level treatments available.[1]

* Principal author, to whom correspondence should be addressed: Dr. Sudhakar Muddu, Silicon Graphics, Inc., muddu@mti.sgi.com. Andrew B. Kahng is associate professor of computer science at UCLA.

[1] For example, [16] describes a characterization and analysis methodology and the need to break ideal scaling in deep submicron interconnect. [13] is another work that centers on

SIA National Technology Roadmap (1997)							
Year	1995	1997	1999	2001	2003	2006	2009
Minimum feature size (*nm*)	350	250	180	150	130	100	70
High-end on-chip clock frequency (MHz)	300	750	1200	1400	1600	2000	2500
# Wiring layers	4-5	6	6-7	7	7	7-8	8-9
Minimum contacted M1 pitch (μm)	1.0	0.64	0.46	0.40	0.34	0.26	0.19
Metal height/width aspect ratio	1.5:1	1.8:1	1.8:1	2.0:1	2.1:1	2.4:1	2.7:1

Table 1: Selected technology projections from the 1997 SIA NTRS.

This work provides, to our knowledge, the first studies of interconnect tuning in the literature. We center on global wiring layers (e.g., M4 and M5 in a 6LM process), and interconnect tuning issues related to bus routing, repeater insertion, and choice of shielding/spacing rules for signal integrity and performance.[2] (Of necessity, our studies are for now independent of several other issues, e.g., wire tapering and choice of wire thickness.)

We address four basic questions.

1. How should width and spacing be allocated to maximize performance for a given line pitch?
2. For a given line pitch, what criteria affect the optimal interval at which repeaters should be inserted into global interconnects?
3. Under what circumstances are shield wires the optimum technique for improving interconnect performance?
4. In global interconnect with repeaters, what other interconnect tuning is possible?

We answer these questions using technology parameters from a representative 0.25um CMOS process; this matches the process technology context for many current- and next-generation microprocessors. Coupling capacitance studies are performed with the commercial QuickCap 3-D field solver from Random Logic Corporation, and interconnect delay and noise coupling studies are performed with the commercial HSPICE simulator from Avant!/Meta-Software. Of particular interest is our study of question (4): we demonstrate that a new methodology for offsetting repeater placements can reduce worst-case cross-chip delays by over 30% in current technologies, versus traditional repeater insertion methodology.

2 Allocation of Width and Spacing for Given Pitch

Our first study seeks to determine how width and spacing should be optimally allocated for a given line pitch. In practice, the actual line width used is considerably greater than the minimum line width achievable in lithography. Thus, there is freedom to tune the width and spacing once assumptions are in place for line thickness and target line length. We note that because very long inter-block lines will have repeaters inserted regularly (see Section 3 below), the maximum line length of interest is equal to the optimum interval between repeaters; this length ranges between 2500 μm and 5000 μm for global interconnect layers in leading-edge technologies.[3]

We have performed detailed studies of "fast" M3 interconnect with 3.2μm pitch, assuming that M2 crossunders are dense (i.e., can be approximated as a ground plane) [14] and explicitly modeling M4 crossovers.

analysis of a given multi-layer interconnect process, as opposed to the underlying interconnect tuning. [5] and [11] are examples ofsystem-level treatments based on Rent's rule for interconnect length distribution.

[2]Even though the results presented in this paper are for aluminum interconnects, similar techniques can be applied for copper and low K interconnects.

[3]Note that the parameters used in the paper are obtained using drawn deminsions of the transistors. Actual dimensions of transistor widths and interconnect length/width/spacing are 64% shrink of drawn dimensions. But the 0.35 µm process itself refers to actual dimension.

Width,Space (µm)	Coupling Capacitance per µm (aF)				
	Left Neighbor	Right Neighbor	Top Plane	Bottom Plane (ground)	Total
1.0,2.2	25.20	25.61	54.79	46.84	152.66
1.2,2.0	29.00	29.26	56.74	48.22	163.53
1.4,1.8	33.33	33.11	57.76	51.53	177.32
1.6,1.6	38.71	38.60	59.09	51.90	188.41
1.8,1.4	44.75	44.12	60.22	51.52	200.92

Table 2: Summary of M3 coupling capacitances extracted using QuickCap. Bottom M2 is a ground plane; top M4 is populated by crossover lines.

Dielectric modeling is based on actual layer data for a representative 0.25μm CMOS process. The commercial QuickCap tool from Random Logic Corporation was used to perform extraction of coupling and area capacitances, as shown in Table 2. As is typical in such analyses, we assume worst-case coupling, i.e., a total coupling factor of 4.0 (worst-case coupling factor of 2.0 to each of the left and right neighbors of the (victim) line under analysis).

Table 3 shows HSPICE-computed line delays for M3 line lengths ranging from 4000μm to 6000μm. Again, dense M2 is assumed to be a ground plane, and M4 crossovers are modeled explicitly. The Table shows that (width,spacing) = $(1.2, 2.0)\mu m$ gives the best performance for the given line pitch.

3 Bounding the Interval Between Repeaters

A very basic study (in some sense a pre-requisite to all other interconnect tuning) asks how often repeaters should be inserted into global interconnects. This is of course a chicken-egg problem, in that the optimum repeater interval depends on the interconnect tuning, and the interconnect tuning depends on the maximum run ever made without an intervening repeater. However, the following can be noted.

- A body of study shows that repeaters should be inserted at uniform intervals. In other words, there should be a constant interconnect length (or interconnect delay) between each pair of adjacent repeaters; the first and last segments of the path are exceptions because in practice the driver and receiver sizes may not be the same as the repeater size.

  Assuming that the driver size and the receiver size are the same as the size of the repeaters inserted along the path, we calculate the total delay, optimal number of repeaters and optimal distance between the repeaters. The total delay for a path with K repeaters is

  $$T_{tot}^{K} = T_{first\_stage} + (K-1) * T_{Rep\_stage} + T_{Final\_stage}$$

  The delay of the first stage is the total delay from the output of driver to the input of the first repeater, i.e., $T_{first\_stage} = T_{gd} + T_{int}$, where gate load delay is $T_{gd} = R_{rep}\left(C_{int}^{eff} + C_{rep}\right)$, interconnect

Width,Space (μm)	50% Threshold Rise Delay (ps)								
	4000 μm M3 length			5000 μm M3 length			6000 μm M3 length		
	Driver Load Delay	Interconnect Delay	Total Delay	Driver Load Delay	Interconnect Delay	Total Delay	Driver Load Delay	Interconnect Delay	Total Delay
1.0,2.2	106.19	113.99	220.17	132.74	168.36	301.10	159.28	233.09	392.37
1.2,2.0	115.00	100.72	215.73	143.76	149.26	293.02	172.51	207.14	379.65
1.4,1.8	126.61	92.80	219.41	158.27	138.04	296.31	189.92	192.10	382.02
1.6,1.6	138.77	87.12	225.89	173.46	130.04	303.04	208.15	181.41	389.56
1.8,1.4	151.24	82.84	234.08	189.04	124.03	313.08	226.85	173.41	400.26

Table 3: Delay estimates for various M3 line configurations. Driver and receiver buffer sizes: (wp=100μm,wn=50μm). Delay is computed from input ofdriver to input of receiver.

delay is $T_{int} = R_{int}\left(C_{int}/2 + C_{rep}\right)$, and R_{rep}, C_{rep} are repeater output resistance and input gate capacitance. The effective capacitance at the gate output can be approximated as $C_{int}^{eff} = \alpha C_{int}$ where α is a constant between $1/6$ and 1 [7]. Let L_p be the interconnect path length between driver and receiver. Then for optimal placement of repeaters the interconnect length between repeaters is $\frac{L_p}{K+1}$. Therefore, the total delay for the path is

$$\begin{aligned} T_{tot}^{K} &= (K+1) * (T_{gd} + T_{int}) \\ &= (K+1) * R_{rep}\left(\alpha * c * \frac{L_p}{K+1} + C_{rep}\right) \\ &\quad + r * L_p\left(c * \frac{L_p}{2(K+1)} + C_{rep}\right) \end{aligned} \quad (1)$$

where r, c are resistance and capacitance per unit length of the interconnect line. We compute the optimal number of repeaters that minimizes total delay by setting $\frac{\partial T_{tot}}{\partial K} = 0$, and obtain

$$K = \sqrt{\frac{rcL_p^2}{2R_{rep}C_{rep}}} - 1 \quad (2)$$

In the current range of $0.35\mu m$ and $0.25\mu m$ process generations, global interconnects have repeaters inserted with periods ranging from $2500\,\mu m$ to $10000\,\mu m$.

- Repeater insertion is also driven by pure interconnect delay, since larger time of flight implies larger slew time on the transition seen at the receiver. Edges with long slew times cause much larger gate delays, are more susceptible to noise, are more susceptible to process-distribution influenced delay variations, and also increase the short-circuit power dissipation. Even in today's designs, slew times above 600-700 ps cannot be tolerated. Thus, even without the delay minimization objective, edge rate control will force insertion of repeaters.

- In practice, repeaters will be implemented using inverters whenever possible, due to performance and area efficiency.

Table 4 summarizes M3 interconnect slew times for line width $1.0\mu m$ and line spacing $1.2\mu m$ (corresponding to a "dense" M3 routing pitch), and input slew time of 400 ps. All capacitance extractions were performed with QuickCap, and correspond to M4 and M1 as the top and bottom ground planes, respectively. Switching factors range from 4 (both neighbors switching in the opposite direction from the victim) to 2 (both neighbors quiet, or one neighbor switching in the opposite direction and one neighbor switching in the same direction with respect to the victim).[4] We see that the M3 distance between repeaters has an upper bound of $5000\mu m$ due to edge rate considerations alone. Separate studies show that this upper bound on distance between repeaters is essentially unaffected by changes to the driver/receiver sizing or the input slew time.

Driver/Receiver (wp,wn)(μm)	Width (μm)	Space (μm)	Length (μm)	SF	Delay (ps)	Rise Time (ps)	Fall Time (ps)
(130,65)/(130,65)	1	1.1	10000	4	589	1679	1510
(130,65)/(130,65)	1	1.1	9000	4	486	1421	1265
(130,65)/(130,65)	1	1.1	8000	4	393	1187	1044
(130,65)/(130,65)	1	1.1	7000	4	310	975	847
(130,65)/(130,65)	1	1.1	5000	4	172	623	525
(130,65)/(130,65)	1	1.1	10000	3	488	1405	1267
(130,65)/(130,65)	1	1.1	9000	3	404	1193	1066
(130,65)/(130,65)	1	1.1	8000	3	327	1001	885
(130,65)/(130,65)	1	1.1	7000	3	259	828	723
(130,65)/(130,65)	1	1.1	5000	3	147	538	458
(130,65)/(130,65)	1	1.1	10000	2	388	1131	1026
(130,65)/(130,65)	1	1.1	9000	2	323	966	869
(130,65)/(130,65)	1	1.1	8000	2	263	817	728
(130,65)/(130,65)	1	1.1	7000	2	209	682	601
(130,65)/(130,65)	1	1.1	5000	2	120	456	393
(130,65)/(130,65)	1.4	1.6	10000	4	366	1123	980
(130,65)/(130,65)	1.4	1.6	9000	4	303	963	832
(130,65)/(130,65)	1.4	1.6	8000	4	246	818	698
(130,65)/(130,65)	1.4	1.6	7000	4	195	686	578
(130,65)/(130,65)	1.4	1.6	5000	4	111	465	384
(130,65)/(130,65)	1.4	1.6	10000	3	320	992	869
(130,65)/(130,65)	1.4	1.6	9000	3	266	854	740
(130,65)/(130,65)	1.4	1.6	8000	3	217	729	625
(130,65)/(130,65)	1.4	1.6	7000	3	172	615	522
(130,65)/(130,65)	1.4	1.6	5000	3	99	422	352
(130,65)/(130,65)	1.4	1.6	10000	2	275	862	759
(130,65)/(130,65)	1.4	1.6	9000	2	229	746	650
(130,65)/(130,65)	1.4	1.6	8000	2	188	640	553
(130,65)/(130,65)	1.4	1.6	7000	2	150	543	465
(130,65)/(130,65)	1.4	1.6	5000	2	87	382	322

Table 4: Summary of M3 interconnect slew times. M4 is top layer; M1 is bottom layer. Two combinations of width/spacing are shown, along with three different coupling factor assumptions. The input slew time is 400 ps and the output slew times are computed as 10%-90% for rise time and 90%-10% for fall time.

4 Benefits of Shield Wiring

Our third study addresses the question of whether shield wiring is an effective means of improving delay and signal integrity performance of long global interconnects. We note that a number of leading-edge design projects seemingly attempt to use shield wiring in their layout methodol-

[4] When two parallel neighboring lines $L1$ and $L2$ switch simultaneously in opposite directions, the driver of $L1$ sees the grounded line capacitance plus *twice* the coupling capacitance of $L1$ to $L2$. If $L2$ is quiet when $L1$ switches, then the driver of $L1$ sees the grounded line capacitance plus the coupling capacitance to $L2$. And if $L2$ switches simultaneously in the opposite direction, the driver of $L1$ sees only the grounded line capacitance. (In leading-edge processes, *each* neighbor coupling is of the same (and possibly greater) magnitude as the area coupling to ground.) The "coupling factor" or "switching factor" is often given in the range $[0, 2]$, and since most lines have two neighbors, the total coupling factor is in the range $[0, 4]$.

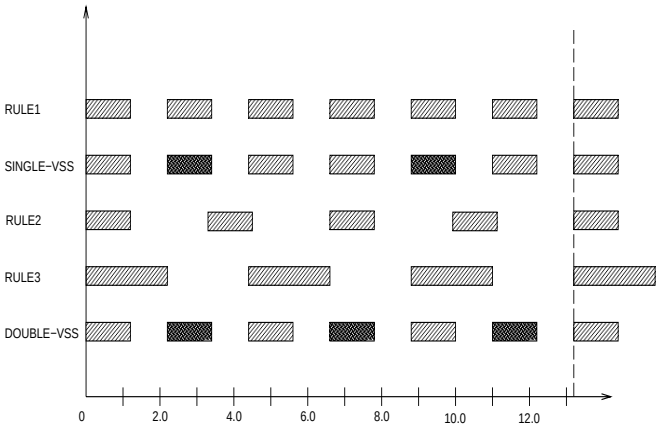

Figure 1: Pitch-matched width-spacing rules. Rule1 allows six lines per 13.2μm; Rule2 and the Single-V_{SS} rule (Rule1 width/spacing, but every third line grounded) both allow four signal lines per 13.2μm; and Rule3 and the Double-V_{SS} rule (Rule1 width/spacing, but every other line grounded) both allow three signal lines per 13.2μm.

ogy (cf. recent capabilities of Avant!'s Aquarius-XO standard-cell router).

We consider various width-spacing rules for M3 interconnect, in order to evaluate the utility of spacing vs. shielding techniques. Our evaluations are with respect to delay only; for all of the configurations, the assumed slew time upper bounds of approximately 600ps imply that noise coupling will not be problematic. Figure 1 contrasts five pitch-matched width-spacing rules:

- **Rule1:** 1.2μm width, 1.0μm spacing
- Single-V_{SS}: 1.2μm width, 1.0μm spacing, with every third line grounded (i.e., every signal line has one grounded neighbor to shield it)
- **Rule2:** 1.2μm width, 2.1μm spacing
- **Rule3:** 2.2μm width, 2.2μm spacing
- Double-V_{SS}: 1.2μm width, 2.1μm spacing, with every other line grounded (i.e., every signal line has two grounded neighbors to shield it)

Again, QuickCap was used to extract capacitive couplings of a given victim line to its neighbor lines and the neighboring top/bottom layers; these results are shown in Table 5.[5] Table 6 shows the delay performance for a 4000μm M3 line, under various bottom ground and top plane configurations.

Our observations include the following:

- The Rule3 rule provides 37% decrease in total delay, but since C_{eff} was not used in the gate load delay computation, actual delay reductions could be even greater.
- The Single-V_{SS} rule is *less effective* than the Rule2 rule; note that the two rules are equivalent in terms of effective routing density.[6]
- The Double-V_{SS} rule gives improved total delays compared with the Rule3 rule, with the rules being equivalent in terms of effective routing density. However, the Rule3 rule yields smaller interconnect delays, so that driver size reductions have greater potential for delay improvement. Thus, the Rule3 rule seems preferable.[7]
- Gate load delays are larger than interconnect delays, suggesting that it is preferable to decrease line widths and increase line spacings. We also note that a dense M4 top layer decreases total delay, and a dense M2 bottom (ground plane) layer decreases total delay for smaller line widths only.

[5] Notice that the Rule1, Rule2 and Rule3 rules have worst-case coupling factors = 4. On the other hand, the Single-V_{SS} rule has worst-case coupling factor = 3, and the Double-V_{SS} rule has worst-case coupling factor = 2.

[6] Our studies have not yet addressed the routing interactions that can potentially affect this analysis. In particular, shield lines may be added to bring power and ground connections to repeater blocks.

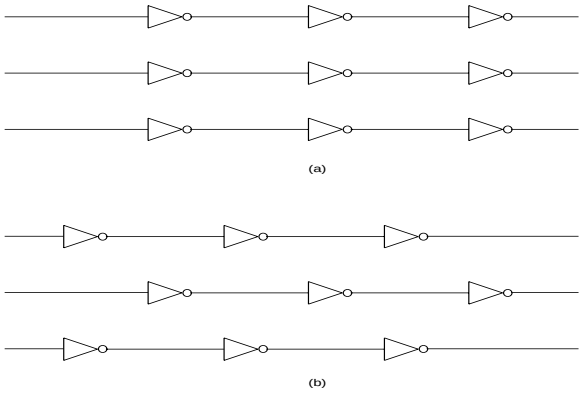

Figure 2: Reduction of worst-case Miller coupling by offsetting inverters. In (a), inverters on the left and right neighbor lines are at phase = 0 with respect to the inverters on the middle line. In (b), inverters on the left and right neighbors are at phase = 0.5.

5 A New Repeater Offset Methodology for Global Buses

Finally, we study another form of tuning that is possible for global interconnects. Our motivations are three-fold: (i) global interconnect is increasingly dominated by wide buses, as discussed above; (ii) in the present methodology, global interconnects are designed in light of *worst-case* Miller coupling; and (iii) in present methodology, a long global bus is routed using repeater *blocks*, i.e., blocks of co-located inverters spaced every, say, 4000μm.

We have proposed a simple method to improve global interconnect performance. The idea is to reduce the worst-case Miller coupling by offsetting the inverters on adjacent lines (see Figure 2). In the previous methodology (Figure 2(a)), the worst-case switching of a neighbor line (i.e., simultaneously and in the opposite direction to the switching of the victim line) persists through the entire chain of inverters. However, with offset inverter locations (Figure 2(b)), any worst-case simultaneous switching on a neighbor line persists only for half of each period between consecutive inverters, *and furthermore becomes best-case simultaneous switching for the other half of the period!*.

To confirm the advantages of this method, the following experimental methodology was used.

- We study systems of three parallel interconnect lines, with lengths either 10000μm or 14000μm. These lines are stimulated by a waveform with risetime = falltime = 200ps. The middle line is considered the "victim" for analysis purposes.
- We model two "technologies" representative of M3 and M4 in an 0.25μm CMOS process. In each technology, line resistance is 50Ω per 1000 μm. In Technology I, capacitive couplings to left neighbor, ground and right neighbor per 1000 μm are respectively 60fF, 80fF and 60fF. In Technology II, capacitive couplings to left neighbor, ground and right neighbor per 1000 μm are respectively 80fF, 160fF and 80fF.

[7] When two buses have activity patterns such that each is quiet when the other is active, then their lines can be interleaved such that they effectively follow the Double-V_{SS} rule. In such a case, interleaving is clearly superior to the Rule3 rule, since the effective routing density is doubled.

M3 Rules	Width,Space (μm)	Ground,Top Planes	Coupling Capacitance per μm (aF)				
			Left Neighbor	Right Neighbor	Top Plane	Bottom Plane (ground)	Total
Rule1	1.2,1.0	Substrate,M4 Line	68.23	68.15	43.68	14.79	195.03
Rule1	1.2,1.0	M2,M4 Line	60.30	60.92	43.96	34.88	202.37
Rule1	1.2,1.0	M2,–	74.67	74.23	–	42.99	192.44
Rule2	1.2,2.1	Substrate,M4 Line	36.87	34.37	58.58	18.07	148.29
Rule2	1.2,2.1	M2,M4 Line	26.96	27.10	58.51	48.72	160.41
Rule2	1.2,2.1	M2,–	42.17	42.43	–	59.15	143.96
Rule3	2.2,2.2	Substrate,M4 Line	35.09	36.50	77.61	22.14	171.52
Rule3	2.2,2.2	M2,M4 Line	26.18	25.61	77.51	67.92	198.82
Rule3	2.2,2.2	M2,–	44.33	43.86	–	73.23	162.14

Table 5: M3 coupling capacitances extracted using QuickCap for various interconnect tuning rules and combinations of bottom and top planes.

M3 Rules	Width,Space (μm)	Ground,Top Planes	50% threshold rise delay (ps)			% Gain w.r.t. Rule1
			Driver Load Delay	Interconnect Delay	Total Delay	
Rule1	1.2,1.0	Substrate,M4 Line	173.04	116.88	289.92	–
Rule1	1.2,1.0	M2,M4 Line	167.84	114.03	281.87	–
Rule1	1.2,1.0	M2,–	178.03	119.62	297.65	–
Rule2	1.2,2.1	Substrate,M4 Line	114.47	84.75	199.22	29
Rule2	1.2,2.1	M2,M4 Line	112.50	83.66	196.16	30
Rule1 with Single VSS	1.2,1.0	Substrate,M4 Line	137.41	97.34	234.75	17
Rule1 with Single VSS	1.2,1.0	M2,M4 Line	136.17	96.66	232.83	17
Rule1 with Single VSS	1.2,1.0	M2,–	139.14	98.28	237.42	16
Rule2	1.2,2.1	M2,–	119.29	87.39	206.68	27
Rule3	2.2,2.2	Substrate,M4 Line	126.91	49.95	176.85	37
Rule3	2.2,2.2	M2,M4 Line	130.08	50.90	180.98	36
Rule3	2.2,2.2	M2,–	130.40	50.99	181.39	36
Rule1 with Double VSS	1.2,1.0	Substrate,M4 Line	99.74	78.11	177.85	37
Rule1 with Double VSS	1.2,1.0	M2,M4 Line	104.34	80.83	185.17	34
Rule1 with Double VSS	1.2,1.0	M2,–	121.14	78.53	199.67	29

Table 6: Delay estimates for a 4000μm M3 line, under various interconnect tuning configurations. Driver and receiver buffer sizes: (wp=100μm,wn=50μm). Delay is computed from input of driver to input of receiver.

- We assume a *period* between inverters (repeaters) of 4000μm. So that HSPICE cannot introduce any error in its RC analysis, we manually distributed the line and coupling parasitics into 40μm segments, i.e., repeaters occurred every 100 segments, and line lengths were 250 or 350 segments. Each segment is modeled as a double-pi model.[8]
- We always place the inverters on the middle line with "phase = 0", i.e., at positions 4000, 8000, ... microns along the line. Inverters on the left and right neighbors are placed according to all combinations of phase = 0, 0.1, 0.2, ..., 0.9 (again with respect to the period of 4000μm). There are 100 different phase combinations. Figure 2 shows the three-line configurations with left/right neighbor phase combinations of (0,0) and (0.5,0.5).
- We stimulate the three lines with the periodic waveform, with the first transition either rising (R) or falling (F). There are eight combinations of directions for the first transisions, i.e., RRR, RRF, ..., FFF.
- Finally, we may offset the input waveforms of the left and right neighbors by -100ps, 0ps or +100ps with respect to the input waveform of the middle line. There are nine combinations of these input offsets.

[8]This segmenting is chosen such that any finer-grain representation does not change the HSPICE-computed delays.

Table 7 shows HSPICE delays for systems of three lines of length 10000 μm, using Technology I, for all combinations of rising (R) and falling (F) initial transition on the input waveform. The Table shows delays for inverter phases (0,0) and (0.5,0.5) on the left and right neighbors of the middle line (phase 0). The effect of Miller coupling is clearly shown.

[9] gives detailed information on experimental results for the worst-case delays (with respect to all eight possible combinations of rising and falling inputs) for the middle line, for each combination of phases for the inverter locations on the left and right neighbor lines. It also provides the results for the same worst-case delays for the middle line, this time taken over all eight rise/fall combinations and all nine combinations of input waveform offsets. In every case, the optimum phase combination is (0.5,0.5), while the traditional phase combination of (0.0,0.0) is actually the *worst* possible. The worst-case delay is reduced by anywhere from 25% to 30% when the repeaters are placed with optimum phase.

6 Conclusions

To our knowledge, this work has provided the first technology-specific studies of interconnect tuning in the literature. We have described experimental approaches to interconnect tuning issues related to bus routing, repeater insertion, and choice of shielding/spacing rules for signal integrity and performance. In particular, four questions have been addressed: (1) How should width and spacing be allocated to maximize

Input waveforms (Left neighbor, victim, right neighbor)	Interconnect Delay (*ns*)					
	Left,right neighbor buffer phases: 0,0			Left,right neighbor buffer phases:0.5,0.5		
	Left neighbor Delay	Victim Delay	Right neighbor Delay	Left neighbor Delay	Victim Delay	Right neighbor Delay
R, R, R	0.361	0.361	0.361	0.510	0.630	0.510
R, R, F	0.428	0.584	0.676	0.533	0.697	0.499
R, F, R	0.546	0.994	0.546	0.483	0.689	0.483
R, F, F	0.676	0.584	0.428	0.499	0.697	0.533
F, R, R	0.676	0.584	0.428	0.499	0.697	0.533
F, R, F	0.546	0.994	0.546	0.483	0.689	0.483
F, F, R	0.428	0.584	0.676	0.533	0.697	0.499
F, F, F	0.361	0.361	0.361	0.510	0.630	0.510

Table 7: HSPICE delays (ns) for three lines of length 10000 μm, using Technology I, for all combinations of rising (R) and falling (F) initial transition on the input waveform. We show delays for inverter phases (0,0) and (0.5,0.5) on the left and right neighbors of the middle line (phase 0).

performance for a given line pitch? (2) For a given line pitch, what criteria affect the optimal interval at which repeaters should be inserted into global interconnects? (3) Under what circumstances are shield wires the optimum technique for improving interconnect performance? (4) In global interconnect with repeaters, what other interconnect tuning is possible? Our answers to these questions are at times quite surprising: in answering (3), we demonstrate that current shielding methodologies may be suboptimal when compared with alternate width/spacing rules, and in answering (4), we propose a new repeater offset technique that can reduce worst-case cross-chip delays by over 30% in current technologies. Ongoing efforts extend our interconnect tuning research to encompass layer thicknesses, more detailed analyses of noise coupling and tuning to meet noise margins, and the delay/noise behavior in emerging technology regimes (Cu interconnect and low-K dielectrics). Finally, we seek to develop more complete full-chip interconnect tuning approaches based on analyses of the interconnect structure, speed target, and power dissipation target for a given design.

REFERENCES

[1] Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*, 1994.

[2] *Personal communication*, SIA NTRS 1997 Revision, Design and Test Technology Working Group (chairs: R. Howard, P. Verhofstadt), 1997.

[3] C. J. Alpert and A. Devgan, "Wire Segmenting for Improved Buffer Insertion", *Proc. Design Automation Conf.*, June 1997.

[4] A. Deutsch, G. V. Kopcsay, C. W. Surovic, B. J. Rubin, L. M. Terman, R. P. Dunne and T. Gallo, "Modeling and Characterization of Long On-chip Interconnections for High-Performance Microprocessors", *final report*, ARPA HSCD Contract C-556003, September 1995. Also appeared in *IBM Journal of Research and Development* 39(5), Sept. 1995, pp. 547-567.

[5] P. D. Fisher, "Clock Cycle Estimations for Future Microprocessor Generations", *manuscript* (to be published), July 1997.

[6] L. Gwennap, "IC Makers Confront RC Limitations", *Microdesign Resources Microprocessor Report*, August 4, 1997, pp. 14-18.

[7] A. B. Kahng and S. Muddu, "Efficient gate Delay Modeling for Large Interconnect Loads", *Proc. IEEE Multi-Chi p Module Conf.*, Feb. 1996.

[8] A. B. Kahng and S. Muddu, "A Glossary on Analysis and Modeling of VLSI Interconnections", *Manuscript*, February 1996.

[9] S. Muddu and E. Sarto, "Studies of Interconnect Tuning for High-Performance Designs", *Patent Application, SGI*, Sep. 1997.

[10] D. P. LaPotin, U. Ghoshal, E. Chiprout and S. R. Nassif, "Physical Design Challenges for Performance", *International Symposium on Physical Design*, April 1997, pp. 225-226.

[11] J. Meindl, "GigaScale Integration: 'Is the Sky the Limit'?", *keynote presentation slides*, Hot Chips IX, Stanford, CA, August 25-26, 1997.

[12] L. Scheffer, "A Roadmap of CAD Tool Changes for Sub-micron Interconnect Problems", *International Symposium on Physical Design*, April 1997, pp. 104-109.

[13] R. F. Sechler, "Interconnect design with VLSI CMOS", *IBM Journal of Research and Development*, Jan.-March 1995, pp. 23-31.

[14] J. Cong, L. He, A. B. Kahng, D. Noice, N. Shirali and S. H.-C. Yen, "Analysis and Justification of a Simple, Practical 2 1/2-D Capacitance Extraction Methodology", *Proc. Design Automation Conference*, June 1997.

[15] L. Gwennap, "IC Vendors Prepare for 0.25-Micron Leap", *Microprocessor Report*, September 16, 1996, pp. 11-15.

[16] S.-Y. Oh, K.-J. Chang, N. Chang and K. Lee, "Interconnect modeling and design in high-speed VLSI/ULSI systems", *Proc. International Conference on Computer Design: VLSI in Computers and Processors*, October 1992, pp. 184-189.