



Getting to the Bottom of Deep Submicron II: A Global Wiring Paradigm

Dennis Sylvester and Kurt Keutzer

University of California, Berkeley
Department of Electrical Engineering and Computer Sciences
dennis@eecs.berkeley.edu keutz@eecs.berkeley.edu

Abstract

Global interconnect is commonly regarded as a key potential bottleneck to the advancing performance of high-speed integrated circuits. Previous work has suggested that local interconnect effects can be managed through a deep submicron design hierarchy that uses 50,000 to 100,000 gate modules as primitive building blocks. This work aims to examine interconnect at the global level to determine if there are any significant roadblocks which will prevent National Technology Roadmap for Semiconductors expectations regarding clock speed from being met. Specifically, the issues of global RC delay, signal time-of-flight, inductance, clock and power distribution, and noise are studied. Results indicate that, while global clock frequencies will necessarily be lower than local clock speeds, NTRS expectations should be attainable to the 50-nm technology generation. Achieving these high clock speeds (10 GHz local clock) will be aided by the use of a newly proposed routing hierarchy which limits interconnect effects at each level of a design (local, isochronous, and global).

1. Introduction

Deep submicron (DSM) effects have been proposed as potential showstoppers to the continuing advancements in integrated circuit performance. Examples of DSM effects include the rising RC delay of on-chip wiring, noise issues such as crosstalk and delay deterioration, and increasing power dissipation. These issues have been addressed in a number of recent works with the general conclusion that interconnect effects will dominate performance in DSM designs.

In a recent tutorial [1], the authors presented an alternative analysis of such DSM effects and analyzed how they are likely to impact future design methodologies. We proposed a new DSM design methodology based on the use of 50,000 to 100,000 gate modules as primitive building blocks. Within these modules, it was shown that interconnect effects will not dominate performance in future high-speed designs. For instance, intramodular interconnect delay comprises $< 40\%$ of the total delay even when considering pessimistic noise assumptions. Likewise, power density of a module remains fairly constant over the range of technologies studied (0.25 to 0.1 μm) meaning power can be expected to scale with chip area. Discrepancies between the results of [1] and that of previous work (e.g. [2]) can primarily be explained by the inclusion of device sizing and scaling wirelengths in the former. Furthermore, the proposed modular design methodology is highly compatible with a reuse oriented design approach in that a wide variety of intellectual property (IP) blocks (e.g. embedded microprocessor cores) can be implemented

within the given size range.

In [1] global wiring was included in the critical path delay analysis by approximating a typical global wirelength (1 cm in 0.25 μm , scaled by 15% / generation) and optimally buffering it with respect to delay. However, a more detailed analysis is necessary since there are a host of global interconnect problems that were not sufficiently addressed there. To motivate this work, we begin by noting that semiconductor processing is advancing at such a rate as to enable the integration of hundreds and even thousands of 50,000 to 100,000 gate blocks at sub-0.1 μm process geometries. Block-level placement and routing of 1000's of such modules under global timing, power, noise, and area constraints is the major design challenge of DSM (see Figure 1).

This companion paper to [1] seeks to analyze and quantify the global impact of interconnect on future high-performance designs as outlined in the 1997 National Technology Roadmap for Semiconductors (NTRS) [2]. Specifically, this work focuses on microprocessors as these designs achieve the highest performance and will meet the limitations of global interconnect first. The main question to be answered is whether or not the new design methodology of [1] will create significant or insurmountable problems in global routing. Other key questions concerning global interconnect to be examined include:

- Over what sized region will we be able to propagate a high-speed signal (> 1 GHz) across chip in a single clock cycle?
 - Including increasing die sizes up to 750 mm^2
- Can a high-speed clock be distributed reliably across a chip in light of increasing die sizes and process variation?
- Does noise at the global level pose a significant signal reliability concern?
- Will inductance result in severely degraded signal integrity?

Each of these topics will be discussed in detail, beginning with primary delay issues such as signal time-of-flight and scaling of global wires. In addition, a new **routing** hierarchy is proposed for DSM designs which complements the *design* hierarchy of [1]. To demonstrate, a representative back-end process for 50-nm microprocessors is suggested.

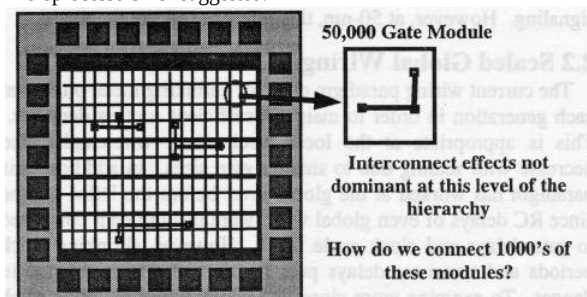


Figure 1. The design challenge in DSM is the assembly of 1000's of 50-100K gate modules considering chip-level interconnect effects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ISPD '99 Monterey CA USA
Copyright ACM 1999 1-58113-089-9/99/04...\$5.00

2. Primary Delay Issues

The foremost problem posed by long interconnect in DSM is that of the reverse scaling properties exhibited by wiring. This well-documented phenomenon implies that continual scaling (i.e. shrinking) of global interconnect, in conjunction with rising die sizes, will soon limit the attainable clock frequencies in a microprocessor. For instance, beginning with the 180-nm technology generation, the NTRS predicts a divergence of global and local clock frequencies due to the impact of global interconnect. In this section we look at the concepts of signal time-of-flight and global conductor dimensions and the constraints they put on global communication.

2.1 Time-of-flight

Larger die sizes and higher clock frequencies predicted in [2] imply that time-of-flight will become an upper bound on speed. The time-of-flight (TOF) for a signal is given by:

$$\text{TOF} = 33.33\sqrt{\epsilon} \text{ ps/cm} \quad (1)$$

Here, ϵ is the dielectric constant of the medium (~ 4.0 for SiO_2). For example, a 750 mm^2 die, as predicted for the 50-nm technology generation [2], cannot support a global clock frequency greater than 5.48 GHz using Manhattan routing techniques. This value is an upper bound since it uses air as a dielectric and the entire clock cycle to traverse the longest potential path. A more realistic value would allot 80% of the clock cycle to this path and use a dielectric constant of ~ 1.5 , resulting in a maximum global clock speed of 3.58 GHz. This speed is near the projected global frequency of 3 GHz, but it is more interesting to look at the impact of TOF on local clocking (isochronous) regions.

As mentioned above, [2] predicts a divergence of global and local clock speeds in future designs. Clearly, due to TOF restrictions alone, an entire 750 mm^2 die could not support global clock frequencies in the range of 10 GHz (see Figure 2). Due to advances in processing, this clock speed will be realizable as approximately the delay through 10 loaded stages in a 50-nm process. Given this divergence of clock speeds, TOF limitations on signal propagation do not play the major role in CMOS designs down to 50-nm; other effects such as RC delay, inductance, and reliable clock distribution will be more limiting factors. Figure 2 reinforces this point by illustrating the relationship between TOF delays and die size. It is also interesting to note that the implementation of new low-k dielectric materials tends to offset the larger die size to provide a fairly constant maximum signal TOF for each technology generation. For instance, we expect a pathological corner-to-corner signal to have a time-of-flight of ~ 220 ps regardless of technology node. Although this value does not increase due to the use of low-k materials, even a constant value will eventually render TOF issues important for global signaling. However, at 50-nm, this effect is still not dominant.

2.2 Scaled Global Wiring

The current wiring paradigm calls for shrinking metal pitches at each generation in order to maintain sufficient routing densities. This is appropriate at the local level where wirelengths also decrease with scaling due to smaller gate sizes. In addition, this paradigm has worked at the global level before the DSM regime since RC delays of even global wires were insignificant compared to gate delays and clock cycle times. However, shrinking clock periods and transistor delays puts the current wiring scheme in danger. To examine more closely, a nominal global wiring pitch of $2 \mu\text{m}$ is used as a point of reference at 250-nm. For each successive technology shrink, the global wiring pitch is reduced by a factor of 0.75. This leads to a final wiring pitch at 50-nm of

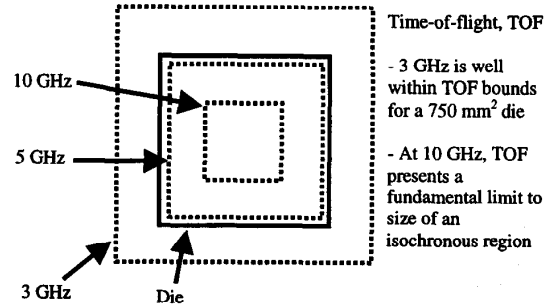


Figure 2. Relationship between TOF delays and die size. The figure focuses on 50-nm microprocessor with $\epsilon = 1.5$.

Technology Generation	Dielectric Constant (ϵ_d)	Resistivity ($\mu\Omega\text{-cm}$)	# of repeaters (W_{min})	# of repeaters ($3 \cdot W_{\text{min}}$)
250-nm	4.0	2.2	14	10
180-nm	3.0	2.2	22	14
130-nm	2.3	2.2	30	22
100-nm	2.0	2.2	50	36
70-nm	1.7	2.2	80	58
50-nm	1.5	2.2	138	98

Table 1. Material properties and optimized number of repeaters for a single corner-to-corner wire for each process.

0.48 μm . Aspect ratio is held at 1.5 for all technologies, which is in keeping with published reports by leading manufacturers [3].

With this wiring pitch, we incorporate the material properties listed in Table 1 to find critical line lengths and optimal buffer sizes. Critical line length is a concept based on the fact that there is an optimal wirelength for each metal level -- wirelengths longer than this should be broken up using repeaters [4]. Since the pathological corner-to-corner wirelength will always be much longer than this critical line length (L_{crit}), we can use these two values to determine the number of repeaters needed to drive such a wire. Results are included in Table 1. Given this information, we now calculate the minimum delay for a pathological wire using repeaters and scaled global wire dimensions. Figure 3 compares this minimum delay for two wire widths to the global cycle time supplied in the NTRS. We see that beyond a certain point the delay actually increases since the wire RC product is rising at the same time that the line length is increasing. Clearly, the scaling of global conductors is not compatible with the expected rise in global clock speeds. This trend has been noticed by industry as well; IBM's 0.18 μm generation is scheduled to have a larger global metal pitch than the previous generation. Figure 4 explores the power ramifications of using scaled global

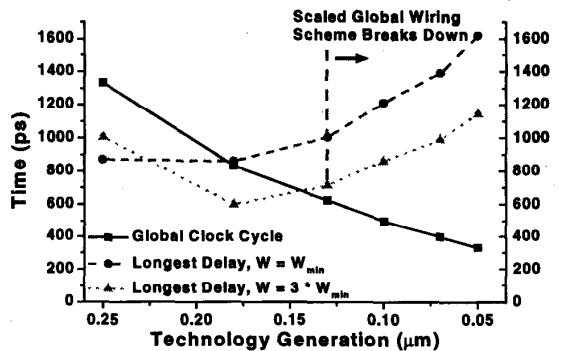


Figure 3. Delays for corner-to-corner wires compared to global clock cycle from [2]. Scaled global wiring is used.

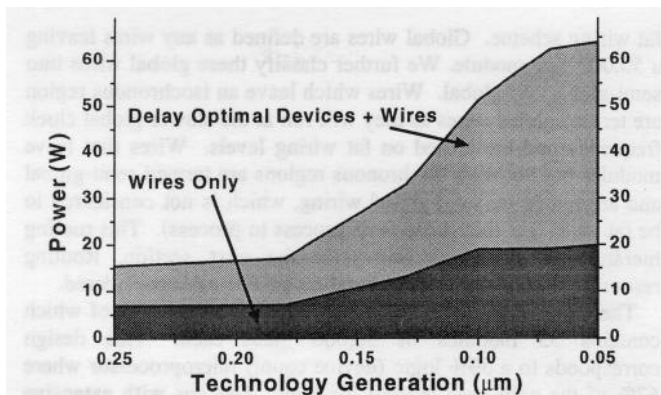


Figure 4. Power for all repeaters and global interconnect (sized $3 \times W_{min}$) where 50% of devices are logic.

wires: L_{crit} drops quickly due to rising line resistance, resulting in huge amounts of repeaters. At 50-nm, this wiring paradigm will consume ~40% of the projected total power for only global interconnect distribution (repeaters + wires) while yielding severely degraded performance.

To help integrate power considerations, a modification is made to Bakoglu's optimal buffer sizing expression [5] to provide an area-optimal buffer size. By multiplying a weighted area function ($W^{1/3}$) by the delay, we obtain a new objective function. Since optimization of delay alone usually results in overly large buffers with high power requirements, we have used the weighted area function to power and delay concerns. The product of T_d and $W^{1/3}$ is then differentiated and the minimum value is found at:

$$W_{opt_area} = \frac{0.541}{R_w C_{in}} \left(\frac{-0.231 R_{dev} C_{out} - 0.126 R_w C_w}{\left(0.053 R_{dev}^2 C_{out}^2 + 0.058 R_{dev} C_{out} R_w C_w + 0.016 R_w^2 C_w^2 + 1.708 R_{dev} C_{in} R_w C_w \right)^{1/2}} \right) \quad (2)$$

This expression gives a smaller value of W than the original formulation in [5]; the delay is consequently higher but the area and power savings are considerable. At line lengths that approach L_{crit} , this formula gives areas that are typically 50 to 65% smaller than [5]. The delay penalty remains under 20% until the line length is about 1/4 of L_{crit} , at which point the delay is not substantial (Figure 5). At L_{crit} , the typical delay penalty is ~11%. As a simpler approximation to (2), we have found that at L_{crit} the W_{opt_area} is 50.8% of W_{opt} from [5] or simply 1/2.

2.3 Fat Wiring

2.3.1 Performance Analysis

The use of fat, or unscaled, wires at the global metal levels was first suggested in [6]. Unscaled wiring has the benefit of a fixed low RC delay at the expense of fewer available routing tracks. In this work, we present a more comprehensive analysis of the performance and routing impact of using such fat wires. Let us first examine the performance aspect of fat wires. Fat wires in this work have a pitch of 2 μm and a thickness of 1.5 μm . The resistance of a clad copper wire at these dimensions is 147 Ω/cm . Capacitance varies from approximately 2 pF/cm to ~0.75 pF/cm in 50-nm technology. Figure 6 explores the maximum reachable distance, L_{max} , for a minimum-pitch fat wire at each generation of interest. L_{max} is defined as the distance that can be traveled in 80% of the global clock cycle predicted in the NTRS. This is compared to the pathological corner-to-corner wirelength, $2 \cdot D_c$. NTRS values for die size are used in the expectation that they will present an upper-bound on chip area according to current design trends. We see that, even with fat wires, corner-to-corner wirelengths cannot be accommodated past 180-nm.

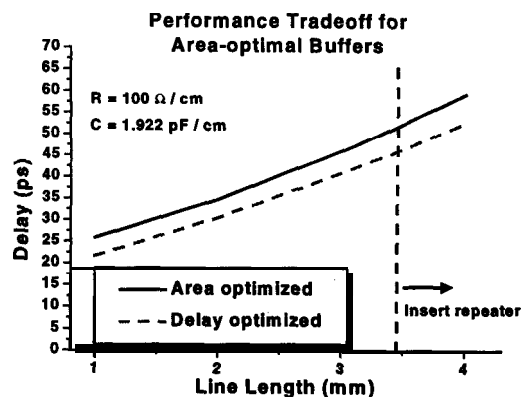


Figure 5. Comparison between new area-optimal repeater size and delay-optimal repeater expression [4].

However, the situation is not as bad as it appears. Based on empirical data and global wirelength models [7,8], we project that the percentage of global wires with lengths $> L_{max}$ is very small, under 2% throughout. Proper floorplanning may be able to further limit these wires. In addition, when using larger than minimum linewidths, L_{max} increases such that, even at 50-nm, it exceeds $2 \cdot D_c$. Thus, for very long wires (a very small portion of the total global picture), wiresizing techniques can be used to maintain acceptable performance at the penalty of increased power. Table 2 demonstrates the effectiveness of wiresizing at 50-nm. By doubling the linewidth from the minimum, L_{crit} rises by 26% which translates directly to L_{max} since repeaters reduce delay dependency to linear. Two scenarios are shown: 1) spacing is kept at minimum, and 2) spacing is equal to linewidth. The second case has severe routing resource penalties with little performance gains. Figure 7 revisits the power issue using fat wires and repeaters. We find that power is actually slightly decreased from the alternative scenario using scaled wires. Furthermore, NTRS global clock frequencies can be met throughout the roadmap using fat wires. Also, the use of area-optimal repeaters rather than delay-optimal (80% area-optimal, corresponding to less critical paths) reduces global interconnect

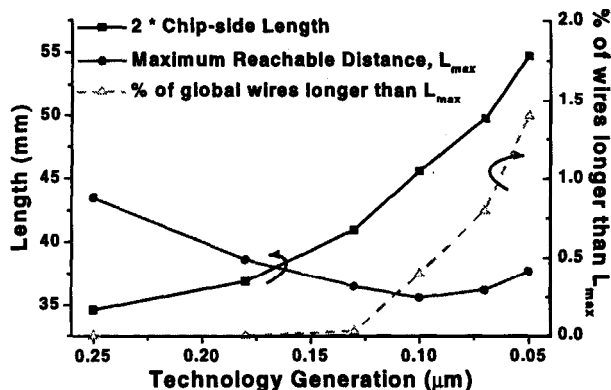


Figure 6. L_{max} becomes less than twice the chip-side length past 130-nm for minimum-pitch fat wiring. Wiresizing allows L_{max} to exceed $2 \cdot D_c$ at all generations.

Linewidth (μm)	L_{crit} ($S = S_{min}$)	L_{crit} ($S = W$)
1	1661	1661
1.5	1917	1991
2	2098	2207
3	2344	2476
4	2504	2639
5	2617	2749

Table 2. Wiresizing leads to increased L_{crit} values (at 50-nm).

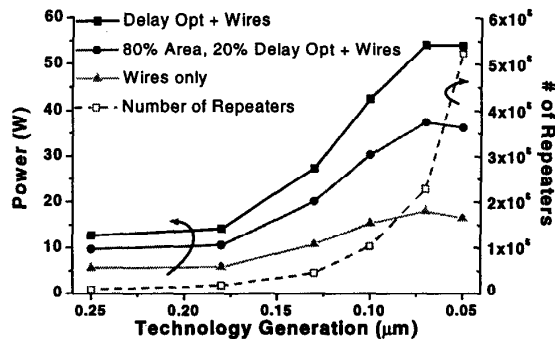


Figure 7. Fat wiring yields lower power and better performance. Area-optimal drivers cut power by >30% and number of repeaters is reduced from scaled wire scenario.

power consumption significantly, by ~30% at 50-nm technology.

2.3.2 Routing Resources

To assess the impact of large-pitch wiring on routing resources, this section introduces several analytical models. In the discussion, we will focus on the logic portion of a design as the wiring requirements for memory are generally much lower than that for logic. Ideal routing capacity is reduced by several factors including power distribution, via blockage, and clock routing.

The power distribution network uses significant routing resources, especially at the top layers. In this work, power grid dimensions are found by limiting the peak IR voltage drop to under 4% of V_{dd} . Analytical expressions are derived in Section 7 to describe the IR drop of an arbitrary layer as a function of metal linewidth. When reliability constraints are met, the percentage of routing resources used is calculated for each metal layer.

It has been estimated empirically that for metal layers with equivalent pitches, an upper layer blocks 12-15% of an underlying layer due to its need to connect to the substrate using vias [6]. However, when larger metal pitches are used on higher levels, the amount of blockage can be reduced if we use fixed-size vias. The relationship between via blockage and metal pitch can be shown to be linear in this case. With a multilevel interconnect system, vias connecting the top layer to the substrate necessarily block all underlying levels. Thus, metal one is blocked by all subsequent layers, resulting in a sizable loss in its routing capacity. Fortunately upper layers have significantly larger pitches than bottom levels, reducing the via penalties associated with multilevel interconnect.

Clock distribution also serves to reduce available routing resources. Due to the regularity of H-tree structures, the total wiring required for such a network can be found fairly accurately. Given the number of clusters in the H-tree (see Section 5), the total wirelength is given by a simple analytical expression in terms of the chip-side length. Additional "within-cluster" routing is approximated using a heuristic that allocates wires from the central driver to the perimeter of the cluster in all directions.

Routing tools cannot fully utilize all the available routing resources for a given design. This is mainly due to the algorithms used within the routing tools. This effect is modeled by first calculating the available routing resources after clock routing, power/ground routing, and via blockages. At this point, the routing area is multiplied by a routing efficiency factor to give the estimated available routing resources. This routing efficiency factor is set at 0.5 in this work, which is based on discussions with industry CAD engineers.

Based on the above models, the wirability of a generic 50-nm microprocessor has been studied to determine the feasibility of the

fat wiring scheme. Global wires are defined as any wires leaving a 50,000 gate module. We further classify these global wires into semi-global and global. Wires which leave an isochronous region are termed global wires as they will run at the slower global clock frequency and are routed on fat wiring levels. Wires that leave modules but not their isochronous regions are termed semi-global and are routed on semi-global wiring, which is not considered to be fat wires (i.e. they scale from process to process). This routing hierarchy is discussed further in the next section. Routing resources for both semi-global and global wiring is considered.

There are found to be 100 isochronous regions, each of which contains 35 modules of 50,000 gates each. This design corresponds to a 50% logic (device count) microprocessor where 67% of the chip area is used for logic. Designs with extensive memory (75% or more) will be more likely to be wirable.

We assume that semi-global routing will use dimensions that are 3X the minimum allowable. For 50-nm, this corresponds to a contacted pitch of 0.45 μm . Aspect ratio is set at two to compromise between resistance and noise effects [9]. Based on Rent's rule [10], an average intra-isochronous wirelength of 1.15 mm is found for fan-out of two. While this value may seem small, it is actually greater than 50% of the isochronous region side-length. With an expected average wiring pitch of $1.5 \cdot P_{\min}$ (conservatively allowing for wiresizing), the total required semi-global routing resources for each isochronous region is 3.32 mm^2 .

Available routing is now determined, starting with 2 levels of semi-global wires. If the requirements exceed capacity, a third semi-global level will be required. Via blockage depends on upper layers which have not yet been assigned – we use final results of fat wiring optimization to yield accurate results. Via blockage for the uppermost layer is 16% and for the underlying layer, 29%. Minimum-pitch wiring is found to be acceptable for power distribution with wiring usage of less than 2%. Clock distribution is performed on fat wiring layers so its impact on semi-global resources is neglected. With a routing efficiency of 0.5 we find these 2 layers to provide 3.84 mm^2 of routing, which is greater than the requirements. Note, however, that significant wiresizing such that the average pitch exceeds $2 \cdot P_{\min}$ would yield an unwirable design.

Let us now examine global routing. Using the same approach as before, (where the global Rent's exponent is slightly smaller due to floorplanning emphasis and potential delay penalties associated with global routing) the average wirelength is calculated to be 7.84 mm (with fan-out of 2). There will be nets with much longer lengths than this, however many signals will only need to travel from one isochronous region to its neighbor. The length of such a net will be in the range of 2-3 mm. Therefore, this wirelength is a conservative intermediate value between relatively rare pathological lines and more typical shorter global wires. Routing requirements are determined with an average wiring pitch of $2X P_{\min}$. We use a wider average pitch here than in the semi-global case because we anticipate the use of shielding wires in order to deal with inductance. This will limit routing density and serve to increase the effective wiring pitch somewhat. Recall that very few (<2%) fat wires will require wiresizing to achieve delay targets. At this point, we determined the need for 2 classes of fat wires. The extreme wires ($L \gg L_{\text{avg}}$) require very fat wiring tracks. However, the shorter global wires also discussed will only waste routing resources if they are routed at these fat levels. So, we allow for the classification of global wires between "shorter" global wires and "extreme" global wires. The former will have greater numbers but shorter average lengths whereas the latter will be very few in absolute numbers but their length can be significant. Given NTRS expectations for 9 metal

layers at 50-nm, four layers remain for global usage (3 for local and 2 for semi-global). We further break this into two fat layers and two layers which will have $\sim 1/2$ the pitch and thickness of the fat layers. Using $P_{fat} = 2 \mu m$, we find the routing requirements to be 656.5 mm^2 by approximating that $2/3$ of the wires will be routed on the lower global layers since their capacity is doubled.

Our analysis indicates that vias block 15% of metal 8, 14% of metal 7, and 27% of metal 6 in this wiring scheme. Power and ground distribution accounts for $\sim 5\%$ on the top level (flip-chip is used with $\sim 10 \text{ mV}$ voltage drop) and 2% on subsequent layers. Clock distribution (both local and global) is found to consume 5% of routing area on the top 2 levels (we estimate that, in order to limit process variation, a shielding plate is used underneath the top level distribution) and none of the other global layers. Summing these contributions, we find a total available routing area of 806 mm^2 . This system is therefore wirable, even in the presence of anticipated shield wires.

This analysis indicate that by adding global wiring layers during each generation (while remaining within the bounds of NTRS expected metal layers), large-scale microprocessors remain wirable at 50-nm (> 1 billion transistors). For instance, six of the nine metallization levels at this process technology should be used solely for global routing, where global routing is defined as routing among 50,000 gate modules.

3. Routing Hierarchy

Our analysis indicates that due to global RC delays as well as time-of-flight considerations, the global clock will necessarily be slower than the achievable local clock frequency. We expect this 2-clock architecture to evolve around the 130-nm generation, which is one generation later (excluding the 150-nm generation) than predicted in the NTRS. The local clock speed will be set roughly by the delay time through 10 loaded gates (approximately 8 to 10 GHz at 50-nm). This will continue to rise as long as faster devices can be made. However, the global clock speed will be set by the propagation delay of the longest global interconnect. Obviously, it is in the best interests of the designer to keep the global clock as fast as possible (or as close to the local clock as possible, in order to reduce latency that will be associated with global accesses). This problem of increasing the global clock speed becomes equivalent to reducing the length of the longest global interconnect. Thus, timing-driven floorplanning will be key in that reducing the pathological wirelength from $2 \cdot D_c$ to D_c will effectively double the global clock speed (since repeaters reduce the delay problem to a linear one).

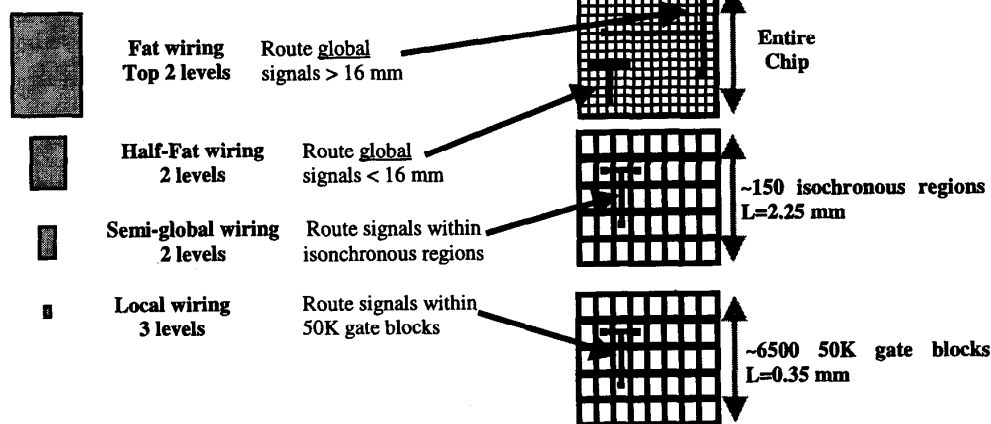


Figure 8. Application of new wiring hierarchy to a 50 nm microprocessor.

What is the size of a locally clocked, or isochronous region at 50 nm? This question can be answered by determining how far we can transmit a signal within a single local clock cycle. Using the top-level fat wiring ($P_{min} = 2 \mu m$), we find that, in 80% of the 10 GHz local clock cycle, L_{max} is about 14 mm. This value corresponds to 16 isochronous regions with an area of 47 mm^2 each. However, a different approach could be used to determine isochronous region size. Using lower levels of metal (not fat wiring), we will obtain a smaller value of L_{max} , leading to more isochronous regions. The advantage here is the exclusion of fat wiring from being used inside of these locally clocked zones, freeing up more routing tracks for longer global wiring.

This point forms the basis for a new wiring hierarchy which complements the envisioned DSM design hierarchy. The design hierarchy at 130-nm and beyond consists of 3 levels – the global level, the isochronous level, and the module level. The optimal wiring hierarchy used to interconnect these designs also consists of 3 levels – global routing (connecting elements at global clock frequency), semi-global routing (connecting modules within isochronous regions), and local routing (connecting gates within modules). In this manner, each level of the wiring hierarchy has a dedicated purpose; to provide connectivity for its corresponding level of the design hierarchy. Figure 8 and Table 3 further explain this new wiring hierarchy. Having outlined our general approach to managing global interconnect, we now consider a number of particular factors that might also inhibit global wiring performance in deep submicron.

4. Inductance

Inductive effects are expected to become more significant in future DSM designs since signal bandwidth is increasing as on-chip rise times decrease. Signal bandwidth is determined by examining the Fourier transform of a digital pulse -- the maximum frequency of interest is related to the edge rate or rise time of the pulse. A signal with rise time T_{rise} has frequency components up to $1/T_{rise}$. This can be conservatively defined as the signal bandwidth. However, a more realistic definition for signal band-

50-nm (μm)	Fat 2 levels	$\frac{1}{2}$ Fat 2 levels	Semi-global 2 levels	Local 3 levels
Pitch	2	1	0.45	0.15
Thickness	1.5	0.75	0.45	0.15
ILD	0.8	0.4	0.18	0.06
Thickness				

Table 3. Back-end process parameters for a 50-nm microprocessor using the proposed wiring hierarchy.

width is given by $0.35 / T_{\text{rise}}$, which corresponds to a cut-off frequency at the -3 dB points [11]. For instance, a 1 GHz signal with 100 ps rise/fall times has a 3-dB bandwidth of 3.5 GHz, which is significantly greater than the operating frequency.

If the product of inductance and angular frequency ($2\pi * 0.35 / T_{\text{rise}}$) is comparable to the line resistance, a first-order statement that inductive effects are important can be made. More accurately, expressions from [12] are used to define a range of line lengths at which inductance should be considered. The expression from [12] is given here and describes an interval of wirelengths where inductive effects are important. This interval is determined by back-end electrical characteristics (R, L, and C per unit length) as well as the signal rise time.

$$\frac{T_{\text{rise}}}{2\sqrt{LC}} < \text{Length} < \frac{2}{R} \sqrt{\frac{L}{C}} \quad (3)$$

Equation (3) assumes that the line is being properly driven -- overdriving RLC lines results in ringing effects and additional delay. We have applied this expression to all generations of technology to determine the relevant wirelengths with regard to inductance. Interconnect characteristics are extracted using a 2-D field solver. Figures 9 and 10 examine the relationship between the intervals found and the critical line lengths. Fat wiring is used at W_{min} . Wider wires have also been studied; their smaller resistance and larger capacitance make the range of significant wirelengths much greater in these cases. Minimum pitch global wiring at 250-nm does not demonstrate inductive effects even for very large spacings to ground (e.g. 20+ pitches). Global wires that are $3*W_{\text{min}}$ see inductive effects when the spacing to ground is larger than $3.5*S_{\text{min}}$. If L_{crit} is smaller than the lower limit of (3), we can safely say that inductive effects are unimportant since signals longer than L_{crit} will not be routed without buffering (for delay reasons). Spacing to ground lines is varied in these figures to show the importance of having a nearby current return path -- this can be accomplished by using shield wires as demonstrated in Figure 11 [13]. Results indicate that the range of inductive effects expands for scaled processes to include more and more of the useful wirelength spectrum for a design. At 50-nm, for example, the use of fat wires will require shield wires within 3 minimum spacings for almost all global signals. This extensive use of shield wires will drastically reduce routing density, further emphasizing the need for added global metallization levels. Finally, it is important to re-emphasize that overdriving global lines (using overly large drivers with small internal impedances) can create inductance problems (e.g. ringing effects, added delay) even for line lengths which do not fall into the aforementioned range of wirelengths. This point will create an additional device sizing constraint in the design phase.

5. Clock Distribution

Another significant issue concerning chip-level interconnect is that of clock distribution. As the clock cycle shrinks, we see a corresponding drop in allowable clock skew. However, larger die sizes mean that a larger overall clock distribution network must be provided. These two points lead to a fine-grain clock network in which the growing network is made up of increasing numbers of shrinking components. In this section, we apply the well-established buffered H-tree to future designs to determine if it can continue to provide low-skew, high-speed clock distribution.

Concentrating on local skew, we use BACPAC (Berkeley Advanced Chip Performance Calculator) to find the required size of an H-tree for a 50-nm design [14]. For a local clock cycle of 100 ps, we allot 5% of this to local skew and 5% to global skew. Modeling of global clock skew is complex since it requires an

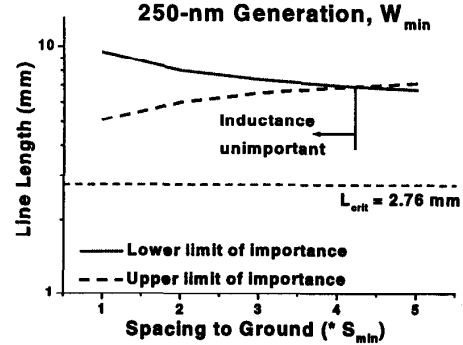


Figure 9. Line lengths of importance for considering inductive effects (global wiring levels).

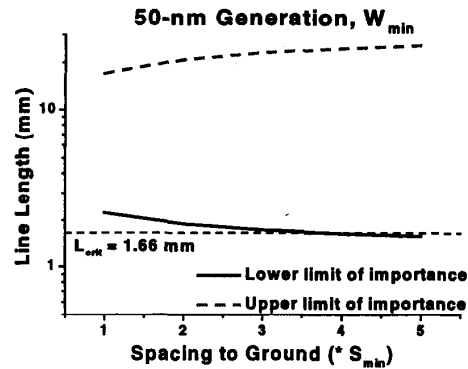


Figure 10. Minimum pitch global wiring at 50-nm requires shield wires within a few pitches to limit inductive effects.

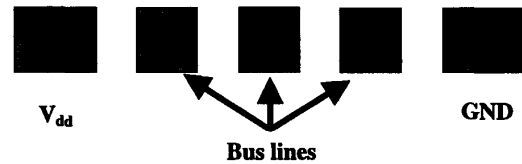


Figure 11. Suggested global routing practice for sub-100-nm technologies. V_{dd} /GND lines supply close current return paths.

estimation of process variation at all intermediate levels of the H-tree. On the other hand, local skew is determined mainly by the size of a cluster (smallest component of an H-tree) and localized process variation at the last level of buffering. Based on the delay model of [15], the expression for local clock skew becomes:

$$T_{\text{skew}} = 0.377(1.1^2 R_{\text{wire}} C_{\text{wire}} L^2 + 1.1 R_{\text{wire}} L C_{\text{latches}}) + 0.693(1.1^2 C_{\text{wire}} L R_{\text{device}}) \quad (4)$$

The factors of 1.1 account for 10% variation in key parameters such as wiring resistance, capacitance, and device resistance. The term C_{latches} accounts for the destination capacitance of the clock driver which consists mainly of clocked transistors in latches. A heuristic is developed to estimate the number of latches along the path of interest (middle of cluster to corner). By setting (4) to 5% of the local clock period, we find the longest possible wirelength that can be driven. The total number of clusters in the clock tree is determined by dividing the total chip area by the cluster area (L^2) and rounding up to the next feasible value (e.g. 16 or 64).

Using this approach at 50-nm, an appropriate H-tree contains 4096 clusters and yields local skew of ~ 1 ps, or 10% of a loaded gate delay. Each cluster has a size of 0.183 mm^2 and contains 1 or

2 50,000 gate modules. This clock tree corresponds to the local clock distribution – the global clock must also be distributed over the entire die. While a clock tree containing 4096 clusters seems complex, the buffered H-tree structure has significant advantages that may allow it to continue as the clock network of choice in DSM. It does not use large amounts of wiring, has relatively low power consumption (compared to grid-based networks), and CAD tools exist to exploit the regularity of its structure in the design phase. An overlay of global and local clock trees is shown in Figure 12, with 256 clusters in the global tree.

From this discussion, it seems possible that local clock skew can be manageable at DSM geometries and clock speeds. However, we have neglected the impact of global skew to this point which may be the most important component of skew due to the large die sizes expected in the future. If global clock skew becomes a bottleneck in chip design, new design styles must be looked at. Currently a potentially exciting new area of research lies in moving the clock distribution network from on-chip to the package level [16]. Specifically, the use of flip-chip packaging allows for easy signal distribution since connection can be made anywhere on the die as opposed to only the periphery [17]. Also, flip-chip packaging has low parasitics (inductance, capacitance) so that the package-to-chip connection is relatively clean. In general, package-level RC parameters are 3 to 4 orders of magnitude smaller than those found in on-chip applications, allowing for easy transmission of digital waveforms over a large area (e.g. 750 mm²) with little attenuation. Using this technique, global clock skew can be minimized while local clock skew is kept low by using smaller clusters within a clock tree.

6. Noise

The issues of $L \cdot di/dt$ noise and crosstalk noise are significant in that, given current design techniques, both issues are expected to become more problematic. Flip-chip packaging presents itself as a partial solution to $L \cdot di/dt$ noise since flip-chip has very low inductive parasitics compared to wirebonding.

Crosstalk noise and delay degradation are also important due to the dominance of coupling capacitance over ground capacitance. Based on the interconnect scaling scenario presented here (aspect ratio of 1.5 for global lines, unscaled minimum pitch) and the analytical crosstalk model of [18], we have found that crosstalk at the global level will not be as significant as the local level due to the use of large repeaters – their capacitance will dampen the effects of coupling capacitance. For instance, the maximum line length of L_{crit} has a crosstalk voltage of 60 to 80 mV at 50-nm depending on whether delay-optimal or area-optimal repeaters are used. This corresponds to 10-13% of the supply voltage, which is not extremely high. Values over 20% are usually considered problematic [9]. However, the use of scaled global wires will lead to larger values of crosstalk since S_{min} will be decreased. This is further evidence that the fat wiring scheme provides optimal

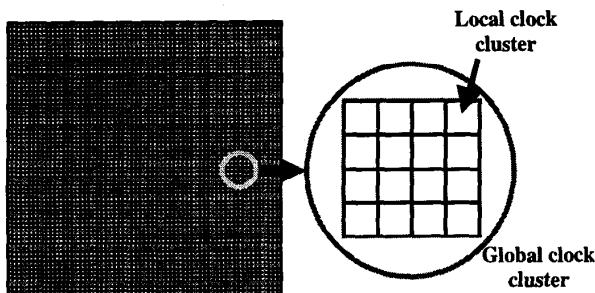


Figure 12. Global/local clock distribution at 50-nm.

performance. Delay degradation of critical timing paths will need to be limited by the use of shield wires, which are also helpful in reducing inductive effects. It is very unlikely, however, that a long net will have two neighbors for its entire run and these signals will switch simultaneously. Even in this instance, we expect only a 30% rise in delay for an optimally driven line of length L_{crit} . Nonetheless, 30% delay variation is unacceptable on critical timing paths, hence the need for shield wires.

7. Power Distribution

Voltage drop in the power distribution networks of large-scale designs is a function of the peak current being drawn from the supply as well as the distribution network resistance. Significant IR drops (e.g. > 5% of V_{dd}) lead to delay variation and reduced noise margins. With rising power consumption yet dropping V_{dd} values, the supply current in future microprocessors will increase quickly. Also, larger wires are needed in order to reduce IR drops along with V_{dd} , which negatively impacts global routing. In this section, we develop analytical models for the voltage drop in a power distribution grid for two packaging technologies to help determine if DC voltage loss problems can be managed in DSM.

Power and ground on a chip are normally distributed using a grid of metal lines on each layer of metal with connections when equipotential lines cross each other. The power grid model of this section is adapted from [19]. The maximum IR drop is calculated for each metal layer and then summed to obtain the total drop from the pads to the silicon level. Our hierarchy consists of a pair of lines (V_{dd} and ground) running parallel at minimum spacing. The distance between lines of the same potential is called the grid pitch. For each grid pitch, there are two lines running the full chip length. We concentrate on the top layer of metal since the majority of the total voltage drop, V_{drop} , occurs there.

Conventional wirebonding constrains power pads to be located at the chip periphery, creating very long lines from the supplies to the middle of the die. Thus, wirebonding results in very wide power distribution lines on the top metal layer hampering the routing of other signals such as clock or global busses. We define the maximum voltage drop on the top layer in this case as:

$$V_{top} = I_{top} R_{top} = J_{avg} \frac{D_c}{2} P_{top} \cdot R_{int} \frac{D_c}{2} = \frac{I_{chip}}{8} P_{top} R_{int} \quad (5)$$

Here we define a total current and multiply by the worst-case resistance to the middle of the die. The current density is determined by calculating the total current on the chip ($Power/V_{dd}$) and assuming it is uniformly distributed. The current distribution area of concern is defined as $D_c/2$ (D_c is the chip-edge length) times the power grid pitch on the top level. Remaining layers in the power grid have a similar expression for voltage drop, except the maximum wirelength that the voltage drop can occur over is reduced to half the pitch of the above layer grid.

Flip-chip technology allows for V_{dd}/GND to be distributed anywhere on the die using solder bumps. If V_{dd} and ground bumps alternate, then the effective distance between power connections to the grid becomes $2 \cdot P_{bump}$, where P_{bump} is the bump pad pitch. The effective length of the worst-case resistive path from pad to underlying level has therefore been decreased from $D_c/2$ to P_{bump} . This reduction corresponds to about a 50X change which can be directly translated to thinner wires and lower voltage drops. The expression for voltage drop on the top layer for flip-chip technology is:

$$V_{top} = I_{top} R_{top} = J_{avg} 2P_{bump}^2 \cdot R_{int} P_{bump} \frac{1}{2} = J_{avg} P_{bump}^3 R_{int} \quad (6)$$

From (5), we see that a reduction in either top-level grid pitch or interconnect resistance is required to maintain a tolerable V_{top} .

Since R_{int} scales as $(1 / \text{linewidth})$, we find that these 2 parameters have an inverse relationship. Hence, reducing line resistance for a fixed pitch (or vice versa) can be taken as the main objective. However, such a reduction requires the use of wider lines and eventually the entire upper layer will be consumed by power distribution routing. In addition, wirebonding pad pitches will have difficulties scaling to meet rising I/O pad number requirements. For these reasons, the continued use of wirebonding techniques is not a scalable concept. On the other hand, flip-chip allows for small IR drops, arbitrary signal distribution to limit global wirelengths, and reduced simultaneous switching noise.

8. Summary and Conclusions

This work aims to thoroughly examine the role of global interconnect in determining future system performance. In particular we aimed to determine whether global wiring issues could be managed in the context of the highly modular approach suggested in [1]. We found that the current practice of scaling global wires is not sustainable beyond the 180-nm generation due to the rising RC delays of scaled-dimension conductors. Furthermore, we have reinforced the notion that there will be a divergence of local and global clock speeds at the 130-nm technology node due to the effects of time-of-flight and wiring RC delays. To combat the RC problem, we recommend the use of fat global wires as described in [6]. In this study we used a comprehensive analysis of routing resources in a 50-nm microprocessor to determine whether the fat wiring scheme (with $P = 2 \mu\text{m}$) is feasible. We found that, by using additional metal layers for global routing, the fat wiring scheme is indeed scalable into DSM. The role of clock distribution in future designs is also studied with the conclusion that the buffered H-tree clock network will enable low values of local skew as long as the number of clusters can be increased. Global skew, while not explicitly modeled, may require moving the global clock distribution network off-chip to the package level where wiring RC is much smaller. Noise issues are also discussed; $L \cdot di/dt$ noise and power supply IR drop become much less significant with the use of flip-chip packaging. A model of IR drop was presented which demonstrated that conventional wirebonding packaging is non-scalable in terms of power supply reliability. Inductive issues were discussed by examining the importance of inductance for various line lengths in different technologies. We found that inductance is becoming a more significant problem, especially when using the fat-wire scheme with low resistance wiring. By using shield wires and trading off routing density, inductive effects should be containable for the most part.

Constructively, we propose a wiring hierarchy which complements the modular design methodology proposed in [1]. This modular methodology proposes the use of 50,000 to 100,000 gate modules of logic to eliminate the impact of interconnect at the local level. These modules are arranged together in isochronous (or locally-clocked) regions which run at a higher clock speed than the global clock. These isochronous regions come together to form the entire design. The wiring hierarchy applies different levels of wiring to route each level of the design hierarchy. Local routing, where minimum pitch is set by lithography capabilities, is used within the modules. Semi-global routing, whose dimensions are a fixed multiple of the local routing for each generation (i.e. semi-global dimensions scale with processes), is used exclusively within isochronous regions. Finally, fat global routing connects these isochronous regions with very long wires. Using this wiring paradigm, we demonstrate a possible back-end structure for a 50-nm microprocessor.

In conclusion, it is clear that managing the aspects of interconnect delay, clock skew, noise, inductance, IR drop, and routing resources is significantly more challenging at the global level than at the module level. Nevertheless, having analyzed the key deep submicron effects we believe that we have outlined a methodology that addresses these effects. To enable the use of the proposed methodology a number of key technologies are required: fat wires; attention to device sizing, buffering, repeater and shield wire insertion; a regular (e.g. H-tree) structure for local clock distribution; and possibly a flip-chip packaging approach for global clock distribution. The absence of any one of these technologies would require a re-evaluation of the methodology, but orchestration of all of these techniques together in the proposed methodology appears to be sufficient to enable designs to meet or exceed projected NTRS speeds for future process generations.

References

- [1] D. Sylvester and K. Keutzer, "Getting to the Bottom of Deep Submicron," *Proc. of ICCAD*, pp. 203-211, 1998.
- [2] National Technology Roadmap for Semiconductors, Semiconductor Industry Association, 1997.
- [3] M. Bohr *et al.*, "A High Performance 0.25 μm Logic Technology Optimized for 1.8V Operation," *Proc. of IEDM*, pp. 847-50, 1996.
- [4] R. Otten, "Global Wires: Harmful?" *Proc. of International Symposium on Physical Design*, pp. 104-109, 1998.
- [5] H.B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, 1990.
- [6] G.A. Sai-Halasz, "Performance Trends in High-End Processors," *Proc. of IEEE*, pp. 20-36, Jan. 1995.
- [7] N. Vasseghi, K. Yeager, E. Sarto, and M. Seddighnezhad, "200-MHz Superscalar RISC Microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 1675-1685, Nov. 1996.
- [8] P. Zarkesh-Ha and J.D. Meindl, "Stochastic net length distributions for global interconnects in a heterogeneous system-on-a-chip," *Proc. of VLSI Symposium on Tech.*, pp. 44-5, 1998.
- [9] D. Sylvester, C. Hu, O.S. Nakagawa, and S.Y. Oh, "Interconnect Scaling: Signal Integrity and Performance in Future High-Speed CMOS Designs," *Proc. of VLSI Symposium on Tech.*, pp. 42-3, 1998.
- [10] W.E. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Trans. on Circuits and Systems*, v. 26, pp. 272-277, Apr. 1979.
- [11] S.-Y. Kim, N. Gopal, and L. Pileggi, "Time-Domain Macromodels for VLSI Interconnect Analysis," *IEEE Trans. on CAD*, v. 13, pp. 1257-1270, Oct. 1994.
- [12] Y.I. Ismail, E.G. Friedman, and J.L. Neves, "Figures of Merit to Characterize the Importance of On-Chip Inductance," *Proc. of DAC*, pp. 560-565, 1998.
- [13] Y. Massoud, S. Majors, T. Bustami, and J. White, "Layout Techniques for Minimizing On-Chip Interconnect Self-Inductance," *Proc. of DAC*, pp. 566-571, 1998.
- [14] <http://www-device.ccs.berkeley.edu/~dennis/BACPAC>
- [15] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Trans. on Electron Devices*, v. 40, pp. 118-124, Jan. 1993.
- [16] Q. Zhu and S. Tam, "Package Clock Distribution Design Optimization for High-Speed and Low-Power VLSI's," *IEEE Trans. Comp, Pack, and Manuf Tech.*, v. 20, pp. 56-63, Feb. 1997.
- [17] R.R. Tummala and E. Rymaszewski, *Microelectronics Packaging Handbook*, Van Nostrand Reinhold, New York, 1989.
- [18] O.S. Nakagawa, D. Sylvester, J.G. McBride, and S.-Y. Oh, "Closed-form modeling of on-chip crosstalk noise in deep submicron ULSI interconnect," *HP Journal*, pp. 39-45, Aug 1998.
- [19] W.S. Song and L.A. Glasser, "Power Distribution Techniques for VLSI Circuits," *IEEE Journal of Solid-State Circuits*, v. 21, pp. 150-156, Feb. 1986.