B8.

On-Chip Decoupling Capacitor Optimization for High-Performance VLSI Design

Howard H. Chen and Stanley E. Schuster

IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598, U.S.A.

ABSTRACT

This paper describes the on-chip power bus modeling and switching noise analysis for high performance circuit design, and the methodology to optimize the placement of on-chip decoupling capacitors. The switching noise is analyzed at both the package level and the chip level. An equivalent circuit which consists of time-varying resistors, loading capacitors, and decoupling capacitors, is used to simulate the switching activities of functional blocks. Both the resistive and inductive voltage drops on the power bus are modelled to identify the hot spots on the chip and ΔV across the chip. Based on the noise analysis results, a decoupling capacitor insertion algorithm is proposed to determine the amount of decoupling capacitance needed to keep the power supply voltage within specification, and optimize the final size and location of on-chip decoupling capacitors.

1. Introduction

As VLSI sub-micron technology advanced in recent years, the state-of-the-art interconnection feature size can be reduced to 1 μm or less. The smaller wire spacing, combined with the longer wire length in large chips, the faster switching speed, shorter cycle time, and smaller power supply voltage, have led to significant noise problems in today's high performance circuits. In particular, the switching noise problem, which traditionally only occurred on the package level, can no longer be ignored on the chip level.

The switching noise is caused by changes in current (ΔI) through various parasitic inductances. The simultaneous switching of I/O drivers and internal circuits can increase the voltage drop on the power supply by an amount of $\Delta V = \sum L\Delta I/\Delta t$, where L is the effective wire inductance of power busses, ΔI is the peak current, and Δt is the rise time. This power supply noise not only will introduce additional signal delay, but also may cause false switching of logic gates. To resolve this problem, decoupling capacitors are often added to keep power supply within specification, provide signal integrity, and reduce EMI radiated noise. For low-frequency ΔV problems, it may be adequate to use only the off-chip decoupling capacitors. However, for high-frequency ΔV problems, the on-chip decoupling capacitor is more effective due to its proximity to the switching activities. A design rule which limits the number of simultaneously switching drivers can also be used to contain the voltage excursions.

Although many techniques were used to characterize and reduce the simultaneous switching noise on the packages and modules, the on-chip power distribution and switching noise problem has not been addressed until recently. For today's 300 MHz CMOS RISC Microprocessor design [1], as much as 160 nF on-chip decoupling capacitance is used to control the power-supply noise. Therefore it is imperative to accurately estimate and optimize the use of on-chip decoupling capacitors for high-performance design. In this paper, we will first describe how to model the on-chip power bus structure and switching activities. The resistive (*IR*) and inductive (*LdI/dt*) voltage drops are considered at both the chip level and package level, so that we can correctly identify the hot spots on the chip, and ΔV across the chip. If excessive switching noise is present at the local hot spots, an iterative improvement procedure will be used to estimate the on-chip decoupling capacitance needed to keep Vdd within specification. The additional decoupling capacitors will then be inserted inside the macros, or placed adjacent to a macro by a decoupling capacitor insertion algorithm which minimizes the total chip area.

2. Power Bus Modeling

To analyze the DC and AC voltage drop on the power busses, an equivalent RLC network is constructed for the on-chip multilayer power bus structure (Fig. 1). The nominal resistance at $25°C$ is determined by each layer's sheet resistance R_s and the width of the power bus ($R_{25} = R_s/width$). At a operating temperature of $85°C$, the resistance increases to $R_{85} = R_{25} \times (1 + T_C \times (85 - 25)) \times (1 + 10\%)$, where T_C is the temperature coefficient, and an additional 10% is added to account for the electromigration induced resistance increase over the lifetime of the device. The total capacitance for the power bus consists of three components: the area capacitance C_{area}, the fringe capacitance C_{fringe}, and the line-to-line capacitance $C_{line-line}$. The area capacitance is the parallel plate capacitance to the wiring planes above and below. The fringe capacitance is the capacitance from the left and right edges of the wire to the wiring planes above and below, based on semi-cylindrical approximation. The line-to-line capacitance is the coupling capacitance between adjacent wires on the same wiring plane. The modeling of wire inductance, however, is more complicated and cannot be represented by simple formula such as $LC = \epsilon_r/c^2$. Depending on the assumptions of ground planes and whether adjacent wires provide the return current, the on-chip inductance can vary by as much as 50x. For a periodic Vdd/Gnd structure, we use the *PROPCALC* program [2] to calculate the propagation characteristics, assuming that the mesh plane on the MCM module constitutes the ground plane. Mutual inductance is also considered if the Vdd and Gnd busses are in close proximity of each other on the same wiring plane or between adjacent planes.

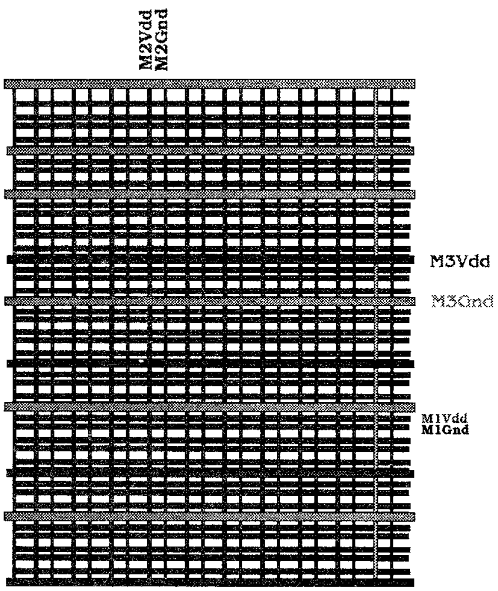

Fig. 1. Multilayer power bus structure.

3. Switching Noise Analysis

After we calculate the resistance, capacitance, and inductance for each power bus segment, an equivalent RLC power-bus network can be generated (Fig. 2). The mesh grid of horizontal and vertical power busses will subdivide the chip into small areas, and the switching activities in each area can be represented by an equivalent circuit with time-varying resistors and capacitors (Fig. 3). The equivalent switching circuits are then attached to the corresponding points on the power bus to model the switching activities.

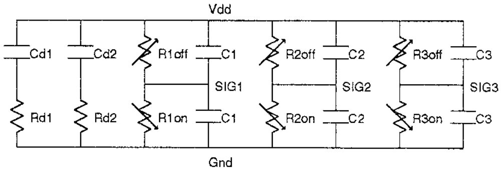

Fig. 3. Equivalent switching circuit.

The loading capacitance for the equivalent circuit is calculated by $C_L = P_{area}/2V^2f$, where P_{area} is the estimated power for the corresponding area, V is the power supply voltage, and f is the clock frequency. When the circuit is turned on, the time-varying resistance will be set to R_{on}, where $R_{on}C_L$ = switching time constant. When the circuit is switched off, the time-varying resistance will be set to R_{off}. Since not all circuits will switch at the same time, the loading capacitance C_L can be further divided into subcircuits $C_{L1}, C_{L2}, C_{L3}, \ldots$, where $\sum C_{Li} = C_L$, to simulate the distributed switching activities. The various switching patterns are controlled by switching on/off $R1$, $R2$, $R3$, ... at different times. Fig. 4 shows the signal waveforms ($SIG1$, $SIG2$, $SIG3$) of a typical 3-stage switching circuit and the corresponding noisy power supply. These waveforms provide a better resemblance to the real switching activities than a simple sine-squared or triangular-shaped current source would have generated.

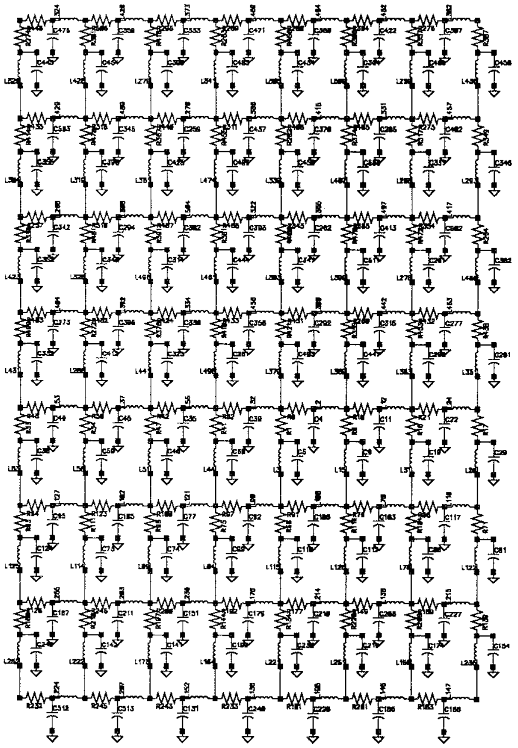

Fig. 2. Equivalent RLC power bus network.

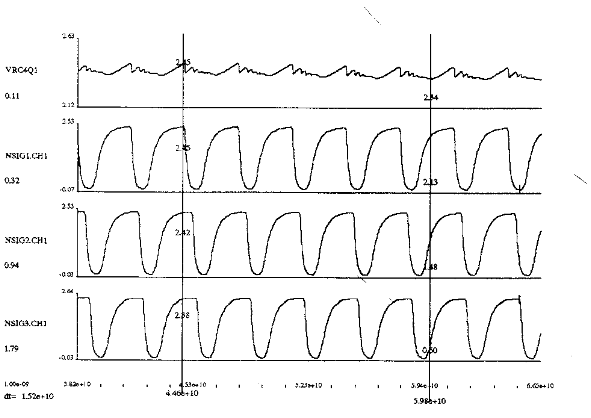

Fig. 4. Noisy power supply due to signal switching.

During the nonswitching period, circuits which are connected to Vdd and Gnd can provide the built-in on-chip decoupling capacitance. This on-chip decoupling capacitor is represented by R_{d1} and C_{d1}, where $R_{d1}C_{d1}$ = built-in decoupling time constant. If additional on-chip decoupling capacitors are needed, they can be placed either inside the macro or on the periphery adjacent to a macro. The additional decoupling capacitor can be properly modelled with its own time constant $R_{d2}C_{d2}$ to determine the minimal size and optimal location, subject to floor-planning constraints.

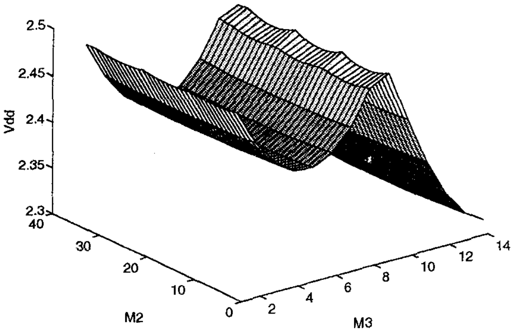

Fig. 5. 3D switching noise analysis.

To display the ground bounce effect and identify the hot spots on the chip, a 3-dimensional graphics program is used during the switching noise analysis. Fig. 5 shows the power distribution for one macro area, which includes 16 C4's (8 for Vdd and 8 for Gnd) [3]. For illustration purposes only, the C4's are directly connected to the power supply and assumed to supply constant voltage. Due to the wider M3 power busses we use in the horizontal direction, the voltage variation is most noticeable along the narrower M2 power busses in the vertical direction.

Table I shows the power supply voltage drop under different conditions. When the entire loading capacitance C_L switches simultaneously, the power supply voltage Vdd will drop from 2.5V at the C4, to 2.26V in the worst case. When the loading capacitance C_L is distributed into 5 different stages, $C_{L1} - C_{L5}$, which switch at different times, the Vdd drops to 2.29V. If we assume the built-in decoupling capacitance to be equal to twice the switching capacitance, Vdd will drop to 2.30V. For comparison purposes, if the power bus inductance is not taken into account, Vdd will only drop to 2.39V. Therefore both the *IR* drop and $\Delta I(LdI/dt)$ noise play a major role in the chip-level switching noise analysis, and must be included in the power bus modeling.

A common mistake which has often been overlooked during the chip-level noise analysis is the absence of a package-level model. If a chip is connected to an MCM or PCB package, the power supply voltage at the C4's can no longer be maintained constant. In Table I, when the on-chip power bus model are connected to the package-level power distribution model at the C4's, Vdd will drop from 2.50V to 2.28V at the package level, and down to 2.19V at the chip level. Therefore, a complete analysis which considers package inductance on the multichip module must be conducted to account for voltage drops on both the package level and chip level.

TABLE I
Vdd Excursion from C4 to Circuits

Assumptions	V(C4)	V(ckt)
Chip level, Simultaneous switching	2.5000	2.2605
Chip level, Distributed switching	2.5000	2.2976
Chip level, Built-in decoupling	2.5000	2.3066
Chip level, RC model (no inductance)	2.5000	2.3963
Chip-MCM, Package inductance	2.2801	2.1987

4. Hot Spots and Differential Noise

One of the major concerns during switching noise analysis is the voltage differential between various locations on the chip. We are concerned with not only the steady state noise of the hot spots, but also the transient noise when circuits switch from one power level to the other. To examine the differential noise between different units on the chip, we partition the chip site into 9 (3x3) regions. The circuits in each region are switched from 20% idle power to 100% full power. Assuming a power supply voltage of 2.5V for the 0.25 μm CMOS technology [4], we measure the transient voltage and the steady state voltage at both the lumped C4 (I/O) location in each region and the local power bus inside the circuits. Table II shows the Vdd distribution when the flip chip C4 technology is used to provide the on-chip power supply. If the C4's are replaced by peripheral I/O's, the minimum Vdd in the center region will drop to 1.964V during transient state and 1.998V during steady state (Table III). Therefore, beyond the I/O density advantage, the use of C4's for high performance design provides significant leverage on noise reduction, especially as the chip size and power increase.

After identifying the hot spots on the chip and their corresponding switching noise, a hierarchical procedure is used to estimate the total on-chip decoupling capacitance needed. This hierarchical procedure first estimates the decoupling capacitance needed for each one of the 9 (3x3) regions to keep the power supply within specification. Then within each region, the total estimated decoupling capacitance for this region will be redistributed among various hot spots by an iterative improvement method. The final decoupling capacitance allocation A_i for each functional block b_i will be totalled, then optimized in various sizes and shapes in the following floor-planning procedure.

5. Decoupling Capacitor Optimization

Since the power supply voltage directly affects the driving capability and signal delay of VLSI circuits, most designs now require ΔV to be contained within 10% of Vdd. To achieve this goal, decoupling capacitors are added to minimize the switching noise. For high performance circuits with a cycle time of 5 *ns* or less, it is estimated that as much as 10% of the chip area may be needed to serve this purpose. Therefore, it is important to estimate and

TABLE II
Vdd Distribution with Complete C4's

Transient Vdd-Gnd in 3x3 chip site (Complete C4's)						
Vdd	C4	Circuit	C4	Circuit	C4	Circuit
Min	2.2362	2.2076	2.2266	2.2039	2.2341	2.2110
Max	2.6934	2.6876	2.6958	2.6896	2.6942	2.6884
Min	2.2180	2.1930	2.2119	2.1908	2.2164	2.1952
Max	2.6946	2.6876	2.6958	2.6889	2.6952	2.6885
Min	2.2323	2.2046	2.2227	2.2009	2.2303	2.2080
Max	2.6942	2.6881	2.6963	2.6899	2.6949	2.6889
Steady state Vdd-Gnd in 3x3 chip site (Complete C4's)						
Vdd	C4	Circuit	C4	Circuit	C4	Circuit
Min	2.3912	2.3683	2.3906	2.3688	2.3910	2.3689
Max	2.6210	2.6176	2.6222	2.6187	2.6213	2.6179
Min	2.3930	2.3708	2.3930	2.3716	2.3929	2.3713
Max	2.6217	2.6180	2.6228	2.6191	2.6220	2.6184
Min	2.3916	2.3687	2.3911	2.3693	2.3915	2.3694
Max	2.6201	2.6165	2.6211	2.6176	2.6203	2.6169

TABLE III
Vdd Distribution with Peripheral I/O's

Transient Vdd-Gnd in 3x3 chip site (Peripheral I/O's)						
Vdd	I/O	Circuit	I/O	Circuit	I/O	Circuit
Min	2.2741	2.1440	2.2748	2.0681	2.2714	2.1568
Max	2.6732	2.4835	2.6726	2.4626	2.6741	2.4882
Min	2.2647	2.1083	2.2993	1.9644	2.2616	2.1310
Max	2.6677	2.4704	2.6456	2.4411	2.6690	2.4768
Min	2.2674	2.1499	2.2664	2.0999	2.2648	2.1610
Max	2.6733	2.4873	2.6732	2.4691	2.6739	2.5038
Steady state Vdd-Gnd in 3x3 chip site (Peripheral I/O's)						
Vdd	I/O	Circuit	I/O	Circuit	I/O	Circuit
Min	2.3905	2.1962	2.3910	2.0997	2.3904	2.2191
Max	2.6223	2.4190	2.6222	2.3135	2.6228	2.4428
Min	2.3963	2.1380	2.4101	1.9986	2.3956	2.1691
Max	2.6200	2.3531	2.6010	2.2049	2.6209	2.3853
Min	2.3920	2.2167	2.3921	2.1322	2.3919	2.2371
Max	2.6216	2.4384	2.6220	2.3464	2.6219	2.4594

allocate the area needed for on-chip decoupling capacitors during the floor-planning stage.

The floor planning of flexible decoupling capacitors is restricted by the topological and ordering constraints of the preplaced functional blocks. Given the relative placement of a set of functional blocks B, we can generate two directed acyclic graphs (G_H, G_V), where G_H is the horizontal constraint graph and G_V is the vertical constraint graph [5]. For each block $b_i \in B$, there is a corresponding node in both G_H and G_V. The chip boundaries are represented by the *LEFT* and *RIGHT* nodes in G_H, and the *TOP* and *BOTTOM* nodes in G_V. If (b_i, b_j) is an edge in G_H, then b_i is to be placed to the left of b_j. If (b_i, b_j) is an edge in G_V, then b_i is to be placed below b_j. The weight x_i of node b_i in G_H represents the x-dimension (width) of the block, while the weight x_{ij} of edge (b_i, b_j) in G_H represents the horizontal spacing between adjacent blocks b_i and b_j. Similarly, the weight y_i of node b_i in G_V represents the y-dimension (height) of the block, while the weight y_{ij} of edge (b_i, b_j) in G_V represents the vertical spacing between adjacent blocks b_i and b_j. As decoupling capacitors fill the empty spaces where $x_{ij} > 0$ or $y_{ij} > 0$, additional nodes b_j' and edges b_{ij}' may be introduced dynamically to represent the pseudo blocks of decoupling capacitors.

After the constraint graphs are defined, the problem of decoupling capacitor optimization can be formulated as follows. Given the dimension (x_i, y_i) of each block b_i, find an optimal dimension $(x_i + \Delta x_i, y_i + \Delta y_i)$ for block b_i, and the dimension (x_i', y_i') of pseudo block b_i' in the adjacent empty space, such that $\Delta x_i \geq 0$, $\Delta y_i \geq 0$, and $(x_i + \Delta x_i) \times (y_i + \Delta y_i) - x_i \times y_i + \sum(x_i' \times y_i') \geq A_i$, where A_i is the area of additional decoupling capacitance needed for block b_i. If $L(G_H)$ is the length (total weight) of the path from *LEFT* to *RIGHT* in G_H and $L(G_V)$ is the length of the path from *BOTTOM* to *TOP* in G_V, then $L(G_H) \times L(G_V)$ is the total chip area which must be minimized.

The following algorithm describes how to size and place the on-chip decoupling capacitors within a given floor plan.

1. Sort all blocks by their estimated decoupling capacitance allocation A_i in descending order.
2. For each block b_i,
    - Fill empty surrounding space by introducing pseudo block $b_i'(x_i', y_i')$, or expanding physical block $b_i(x_i + \Delta x_i, y_i + \Delta y_i)$, without increasing $L(G_H)$ or $L(G_V)$, until A_i is entirely allocated, or no more empty space is available.
    - Update A_i by subtracting the area of empty space used for decoupling. Update G_H and G_V.
3. If chip size is fixed, stop.
4. Sort all blocks by their updated decoupling capacitance allocation A_i in descending order.
5. For each block b_i,
    - Fill empty surrounding space by introducing pseudo block $b_i'(x_i', y_i')$, or expanding physical block $b_i(x_i + \Delta x_i, y_i + \Delta y_i)$, without increasing $L(G_H)$ or $L(G_V)$, until A_i is entirely allocated, or no more empty space is available.
    - Update A_i by subtracting the area of empty space used for decoupling. Update G_H and G_V.
    - Expand physical block $b_i(x_i + \Delta x_i, y_i + \Delta y_i)$, with minimal increase of $L(G_H)$ or $L(G_V)$, until A_i is entirely allocated. Update G_H and G_V.

The optimization of on-chip decoupling capacitors involves an iteration process between circuit simulation and floor planning. Given the specifications and location of each functional block, the circuit simulator will analyze the switching noise on the power bus, identify the hot spots, and determine the amount of decoupling capacitance needed. The floor planner, in turn, translates the amount of decoupling capacitance into physical area, and determines its aspect ratio and location. The positions of neighboring blocks may also be affected, if there is not enough room to embed the decoupling capacitors. The added decoupling capacitors are then modelled with proper time constants and simulated with the new floor plan until ΔV is contained.

The exact placement of decoupling capacitors will be refined after the physical layout for each macro is complete. It may be possible allocate a certain portion of the decoupling capacitor inside the macro, if space is available. A transistor-level simulation can also be performed, with the real switching patterns for each macro, to verify that the ΔV results coincide with those derived from the equivalent circuit model during floor planning.

6. Experimental Results

The on-chip power bus and switching circuit models have been used in conjunction with the package-level MCM power distribution model to analyze the switching noise problem for high performance VLSI design. The analysis maps the entire chip floor plan, where each functional unit has its own power specification and switching pattern, to the power bus structure, and identifies the hot spots where decoupling capacitance may be needed. Fig. 6 shows the transient Vdd distribution when C4's are used to provide the on-chip power supply, and 5% of the total chip area is available for decoupling capacitance. Since our chip size is fixed, the on-chip decoupling capacitors can only be inserted in empty spaces inside the macros, or adjacent to the macros, to minimize the noise. If the chip is not optimized with the additional decoupling capacitance, the worst case transient Vdd will drop from 2.123V to 1.992V with the C4's, and drop from 1.739V to 1.676V without the C4's (Table IV). The circuit simulation time for a 20-cycle whole chip noise analysis is 140 minutes on an IBM R/S 6000 workstation with a memory requirement of 692M using the optimized L/U factorization solution method. The same analysis will require 272 CPU minutes, but only 76M memory, if row-wise Gaussian elimination is used as the solution method.

TABLE IV
Minimum Transient Vdd Variation

Minimum Transient Vdd	Low	High
C4 with on-chip Decap	2.123V	2.286V
C4 without Decap	1.992V	2.198V
Peripheral I/O with Decap	1.739V	2.303V
Peripheral I/O without Decap	1.676V	2.234V

7. Conclusions

With the advent of smaller circuit geometry and shorter cycle time, the on-chip noise problem will become a major concern for high-frequency VLSI circuit design. The switching noise analysis described in this paper can identify the hot spots on the chip where the most significant Vdd drops occur, and estimate the amount of decoupling capacitance needed to minimize the noise. A decoupling capacitor insertion algorithm which minimizes the total chip area will then be used to optimize the size and location of decoupling capacitors. The on-chip decoupling capacitor optimization problem has introduced a new set of floor-planning constraints for today's high performance design, and these new constraints will present many challenges for future research in this area.

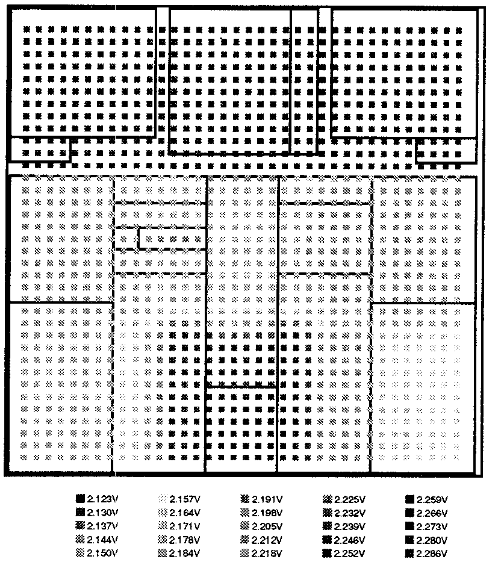

Fig. 6. Transient Vdd distribution with complete C4's.

8. References

[1] W. Bowhill, *et al.*, "A 300 MHz 64b quad-issue CMOS RISC microprocessor," in *Proceedings, International Solid-State Circuits Conference,* February 1995, pp. 182-183.

[2] B. Rubin, "An electromagnetic approach for modeling high-performance computer packages," *IBM Journal of Research and Development,* vol. 34, no. 4, p. 585-600, July 1990.

[3] L. Miller, "Controlled collapse reflow chip joining," *IBM Journal of Research and Development,* vol. 13, no. 3, pp. 239-250, 1969.

[4] B. Davari, *et al.,* "A high performance 0.25 *μm* CMOS technology," in *Proceedings, International Electron Devices Meeting,* December 1988, pp. 56-59.

[5] G Vijayan and R. Tsay, "A new method for floor planning using topological constraint reduction," *IEEE Trans. Computer-Aided Design of ICAS,* vol. 10, no. 12, pp. 1494-1501, December 1991.