PHYSICAL DESIGN CHALLENGES FOR PERFORMANCE

INVITED PAPER

David P. LaPotin  Uttam Ghoshal  Eli Chiprout  Sani R. Nassif
IBM Austin Research Laboratory, Austin, TX, U.S.A.
{dpl, ghoshal, chiprout, nassif}@austin.ibm.com

ABSTRACT

Recent trends in high performance microprocessor design suggest that complex gigahertz processors based on deep-submicron CMOS technologies will be practical in the near future. It is also certain that integration complexity will result in ever-increasing demand for interconnection connectivity and bandwidths. Front-end chip planning, back-end interconnect design, and global electrical analysis issues will be at the forefront.

1. INTRODUCTION

Future deep-submicron CMOS fabrication processes will enable designs containing tens of millions of logic devices and on-chip clock rates approaching one gigahertz. Obtaining the highest performance design points will require an intimate understanding of the interrelationship between fabrication technology, circuit design techniques, and physical layout. Complete electrical analysis in electromagnetic terms is difficult and expensive once a chip is completed. Yet, with increasing operating frequencies and smaller scaling, electrical analysis has been brought to the forefront with the impact of high frequency effects on digital design, e.g. ringing, crosstalk, false switching, and pattern-dependent delay. This paper presents one vision of the highest leverage CAD issues and required innovation to meet next-generation technology and performance challenges.

2. PHYSICAL ISSUES

We believe that interconnection delay and signal integrity issues will be managed by controlled-impedance interconnection structures [1], interconnection layout rules (both design rules and topological), and structured circuit design. Controlled impedance striplines with orthogonal X-Y wiring layers can be designed with layout techniques similar to those practiced in multichip module system design. Dedicated wiring layers for power and global signal management, and possibly reducing signal routing capacity, will necessitate the need for more advanced and aggressive routing, partitioning, and placement algorithms. Differential routing can be employed to minimize global signal coupling. Integrating cells and modules at the global level should be easier because the input and output load characteristics of macros can be well characterized in terms of the impedance characteristics of the interconnections. Physical design tools exploiting the regularity and structure of wide data buses and I/Os for achieving low-voltage swings, crosstalk suppression, and $df/dt$ control will be vital for high performance. Methods to incorporate quick but accurate estimation of electromagnetic effects such as line-to-line coupling and line inductance will have to be accommodated in future analysis algorithms. The choice of circuit families and machine organizations will be closely coupled to physical issues, such as partitioning, placement, and global routing, where structured, regular-layout solutions will dominate.

Manufacturing process variations, which exhibit themselves as spatial variations in parameters such as $L_{	ext{eff}}, V_t$, and interconnect capacitance, can profoundly influence chip timing. It is not uncommon to have intra-die variations account for 50% or more of the total variability. This implies that timing uncertainty due to process fluctuations can be longer from one path, to another, within a chip, than it would be for the same path across multiple chips. Existing performance-driven PD tools unrealistically assume perfect path tracking.

In order to properly handle the effect of process variations, a new generation of modeling and analysis tools is required. These tools must explicitly deal with: (1) the spatial relationship among active (device) and passive (interconnect) components, e.g., distance and orientation; and (2) the implied statistical behavior of the corresponding physical and electrical quantities, e.g., $L_{	ext{eff}}$ and line capacitance. Examination of these models may lead to new insights and PD algorithms.

We have begun to see inroads of noise-aware algorithms into physical design, particularly at the routing level. These early approaches should be extended to include physical information about noise coupling and timing, and logic-orthogonality information, in order to choose low-noise paths with minimal pessimism. Pre-layout and post-layout simultaneous driver, receiver, and wire sizing to address timing; power; electromigration, inductance, and other concerns will also become important.

An important field of exploration is early physical-design estimation and global chip planning. Beyond traditional floorplanning schemes, it would be helpful to have techniques to project forward using a given design point and a particular design style to obtain figures of merit as to the resulting PD impact. Key open issues are the effect of architectural and partitioning decisions on back-end PD, obtaining early estimates of interconnect lengths and timing arrival times, the density of wiring and transistors, power...
and thermal analysis, inductance issues, shielding versus spacing tradeoffs on global nets, and chip wiring-layer utilization and composition.

Buffer insertion has become an important mechanism to deal with interconnect delay. However, buffer insertion is typically applied at higher levels in the design process using wiring-delay estimates. It would be more effective for buffering algorithms to deal directly with the physical layout either to propose, pre-layout, ways of inserting white space for buffer insertion during layout, or post layout, to automatically search for locations where buffers may be physically inserted. Across chip length variations (ACLV) and thermal balancing issues, both affecting buffer delay and clock skew, must be considered.

Static timing techniques are commonplace for critical path optimization, repowering, and final timing checkoff. At the transistor level, current approaches rely on identifying channel-connected components, constructing a timing graph, and finding the longest path from primary output and/or latch to primary input and/or latch. As clock frequencies increase and logic levels reduce, accuracy issues are a major concern. Specifically, enhancements to consider larger clusters beyond channel-connected components, device modeling accuracy, assumptions made during pattern dependent timing simulation, path sensitization, manufacturing variations, and false path issues all need to be investigated.

3. GLOBAL SIGNAL PROPAGATION

Distributing gigahertz clocks globally across the chip will also be challenging. The worst-case clock skew is dependent on the variations in interconnect delay and buffer skews. Controlled-impedance transmission line techniques result in time-of-flight delays that are less dependent on both metal and dielectric thickness. Clock buffer layout and placement will have to null proximity effects leading to channel length variations, threshold voltage variations, and temperature variations. Temperature effects are especially significant for clocks and global interconnections in high-performance designs where global variations of ten degrees Centigrade are not uncommon. For example, the sector buffers in a clock distribution network operate at different local surface temperatures depending on the sector load capacitance. In addition, metal interconnection resistivities are also sensitive to temperature variations (4-6% per degree Centigrade).

Another issue is the distribution and placement of high-power circuits, particularly along critical timing paths. Localized hot spots can degrade performance due to increased junction temperatures. Placement cost functions may need to consider both global power and thermal densities in order to reduce temperature induced performance variability. Finally, global clock optimization and skew management have a major influence on performance; current design methodologies throw away performance by not simultaneously optimizing the clock network and circuit together.

Design and optimal placement of decoupling capacitors to reduce current transients in the power distribution network has to be investigated in detail. Currently, decoupling capacitors are sized and placed manually, and the necessity for them is roughly estimated by back-of-the-envelope type calculations. Electrically, they are typically constructed as a distributed capacitance, where the effectiveness is reduced by parasitic affects as well as package inductance. The more decoupling capacitance the better methodology may result in underdamped oscillations in the power grid and may actually degrade the performance. More rigorous approaches to decoupling capacitor analysis and placement are needed.

Chemical mechanical polishing is in wide use by all major semiconductor companies for back-end planarization. To minimize etching variations during manufacturing, unused wiring tracks are routinely filled with metal, which may or may not be connected to power rails. When the metal fill is connected, the capacitance of adjacent signal lines rise; when the fill is floating, adjacent signal coupling is increased. Moreover, the effect of this phenomenon is not considered in chip design, since parasitic extraction and electrical timing analysis are done prior to the filling process. Specific analysis and recommendations are required.

4. INDUCTIVE EFFECTS

On-chip interconnects running at high speed have begun to exhibit inductive effects. Long, wide and low-loss lines, driven by large, high-current drivers can exacerbate inductive effects. Further, it has been shown that, while skin effect does not seem to come into play due to the small thickness of the metal lines, there could be proximity effects where the return path of the driven current would vary with frequency, leading to frequency dependent inductive effects.

To date, most physical design algorithms that deal with delay constraints such as floorplanning, placement, and wiring, use interconnect delay estimates based on RC abstractions. One such abstraction is Elmore’s model. However, due to the increase in inductive effects, if not addressed by specific design rules (wiring topology and loading restrictions), PD algorithms will have to, in the future, contain some estimates as to the effect of inductance on the wiring delay. Inductive effects will mean that the estimates will no longer be of a monotonic waveform, as in the RC case.

Another impact on PD design methods will be inductance avoidance. Placement of ground return planes, routing, and wire sizing of signal and clock lines should be done with inductance in mind. For example, inductance is reduced if a ground return path is provided nearby. This can be in the form of either a nearby ground line, $V_{dd}$ line, or a dedicated metal plane. Also, inductance can be reduced by decreasing the line width thereby increasing resistance of the line. Proximity effects can be reduced by avoiding high resistance thin lines next to low resistance wide lines that give a variable current return path. Mutual inductance between lines can also be an issue. Future PD systems will have to include inductance analysis and avoidance mechanisms.

5. CONCLUSIONS

Future process technologies, aggressive circuit techniques with possibly greater noise sensitivity, and ever increasing clock frequencies will drive new tool development in the upcoming years. High speed CMOS applications will require an intimate understanding of the interrelationship between fabrication technology, circuit design, physical layout, and electrical analysis. The traditional physical-design scenario of back-end placement and routing will be augmented to a more analysis driven, higher-performance paradigm. Early chip planning tools in conjunction with interconnect planning and detailed electrical analysis will be a prerequisite for high performance.

REFERENCES