Thermal Aware Buffer Insertion in the Early Stage of Physical Designs

Jaehwan Kim, Byung-gyu Ahn, Minbeom Kim, and Jongwha Chong

Abstract—Thermal generation by power dissipation of the highly integrated System on Chip (SoC) device is irregularly distributed on the intra chip. It leads to thermal increment of the each thermally different region and effects on the propagation timing; consequently, the timing violation occurs due to the misestimated number of buffers. In this paper, the timing budgeting methodology considering thermal variation which contains buffer insertion with wire segmentation is proposed. Thermal aware LUT modeling for cell intrinsic delay is also proposed. Simulation results show the reduction of the worst delay after implementing thermal aware buffer insertion using by proposed wire segmentation up to 33% in contrast to the original buffer insertion. The error rates are measured by SPICE simulation results.

Index Terms—Buffer insertion, thermal variation, cell delay, library characterization

I. INTRODUCTION

Continuous development of CMOS technology has realized SoC designs which is the high level integration of a variety of intellectual property (IP) cores into a single chip. SoC designs are being settled as a main design issue, and have many advantages such as greater functionality with a simpler hardware design, lower system cost, and more reduced time to market than the board level systems; however, power density and operation frequency are significantly increased [1, 2]. Because these increments lead to power dissipation and heat generation on the chips rise, the thermal problem is emerged as one of the most important issues in SoC designs. Temperature impacts on resistivity, definitely increases it, so that the resistances of interconnected wires and transistors are increased [3]. In the result, additional delay is induced especially in high performance SoC circuits; consequently, the management of thermal effect is necessary to overcome such degradation of circuit performance. In general, temperature is non-uniformly distributed on the chips due to the irregular CMOS switching activity and the different operation modes from various functional blocks [3]. Thermal effects on the timing variation have been studied [4] that if a driving cell and/or wire are effected by the thermal effects, the position of an inserted buffer is moved while the previous buffer insertion methods are inserting a buffer in the middle of a wire or buffers with the regular length of a wire; therefore, additional buffers are needed because the shifted buffers drive a longer wire as well as K. Sundaresan et al. analyzed the impact of Joule heat and thermal gradients on timing violation [5]. To alleviate timing violation occurred by temperature, the delay optimization technique like buffer insertion is essential. When implementing buffer insertion, the numbers and positions of buffers are carefully dealt with, according to the result of recent study in [6], 43.6% of repeaters are inserted in nets of the inter block level, and 70% of repeaters are occupied of the total number of cells in case of the intra block; therefore, predicting the thermal dependent resistances in the early stage of physical design and planning to reserve the area of the buffer spaces considering temperature are necessary. For the accurate estimation, buffer insertion planning with the accurate
thermal aware timing budget by using thermal aware delay model and wire segmentation stage is proposed before place & route (P&R).

The cell delay of buffer is needed for buffer insertion and it is used as LUT matched by input transition time and output capacitance load. The spice simulations make LUT for every cell and thermal is one of the most important input factors. As thermal variations occur, the new spice simulations are needed for new cell characterization and it takes much time. Cell library only served LUT of the cases for best, typical and worst corner. We propose fast cell characterization method without every spice simulation for thermal variation. This paper also studies the distribution of temperature on a chip grid by grid and invests the changed resistance of nets and buffers to be placed in each thermal grid cells. For simulation, buffer insertion is implemented using by Van Ginneken algorithm modified delay model and wire segmentation for thermal efficiency.

Section II shows the methods of buffer insertion in the floorplan stage. Section III shows the buffer insertion methods and cell characterization. Section IV and V show the simulation results and conclusion.

II. FLOORPLAN LEVEL BUFFER INSERTION

In case that a cell drives a long wires or many sinks, the propagation delay of such driving cell is effectively reduced by inserting buffers. Buffer planning in the early stage of physical design estimates the area of the buffer space to be inserted after P&R is highly required because the blocks and cells work as the obstacles from the buffers, as well as the large number of buffers are inserted in the advanced and complex SoC devices so that the buffer planning in the floorplan stage should be implemented in advance.

1. Buffer Insertion

For timing estimation in floorplan, it is assumed that the nets are interconnected with among the block pins which are placed the edge of the blocks and connected with the closest cell of a source and sinks respectively as shown in Fig. 1. Global routing is performed based on Minimum Spanning Tree (MST) in all connected inter block nets and network topology is extracted. In given routing trees, buffer insertion of Van Ginneken algorithm [7] based on dynamic programming is used that finds the optimal position of the segmented candidate nodes by selecting the optimal slack and the smallest load capacitance from the sink to the source. Our work includes the thermal effects into this buffer insertion method.

2. Delay Model

To find the optimal position of buffers, the delay model including the propagation delay of wires and the driving delay of cells is needed to calculate the optimal slack (required arrival time – delay). \( \Pi \) model based on Elmore delay model is used for wire in the original buffer insertion method, i.e. when a signal moves from the vertex \( i \) to \( j \), the edge \( e \) = \( (v_i, v_j) \) between the vertex is modeled as

\[
D(e) = R(e) \left[ \frac{C(e)}{2} + C(V_j) \right]
\]

where \( C(V_j) \) represents the downstream capacitance at vertex \( V_j \), and if the buffer type \( B_k \) is inserted at \( V_r \), the buffer delay is modeled as in

\[
D(v_i) = R(B_k)C(V_j) + K(B_k).
\]

where \( R(B_k) \) is the driving resistance of the buffer \( B_k \) and \( K(B_k) \) is the intrinsic delay of buffer \( B_k \).

III. THERMAL AWARE BUFFER INSERTION

State of art SoCs have a lot of functional blocks and multi voltage levels for operation of the blocks. These features of the SoC devices make the non-uniform
distribution of power consumption and induce the thermal gradients variably ranged on a chip, eventually the characteristics of the chip can be changed caused by the temperature. Ajami et al. stated that it is a safe assumption that the capacitance per unit length is independent of thermal variations along the wires while the resistance is dependent of them [4]. Unless considering the changed resistance of interconnected wires and driving cells by temperature, it is possible to misestimate the delay and insert the insufficient buffers. The lack of the number of buffers leads to the occurrence of the timing violated paths; therefore, delay calculation considered the characteristics of the thermal effect on a chip is required for effective buffer insertion from the floorplan stage. In this study thermal aware delay model and segmentation for buffer insertion proposed in [8] are extended and described in part 1 and 2.

1. Thermal Aware Delay Model

Under the assumption of that resistances are changeable by temperature while capacitance remains unchanged, to apply increased resistances caused by temperature into the buffer insertion algorithm while calculating RC delay, resistances have to be replaced with them of included the temperature effects based on following equation:

\[ R' = R[1 + \alpha(T' - T)] \]  \hspace{1cm} (3)

where \( R \) is a resistance in standard temperature (T) and T is assumed at 293.15 K, \( R' \) is an increased resistance caused by temperature (\( T' \)), and \( \alpha \) is a temperature coefficient. Assumption that edge (e) is passing through the number of thermal grid cells, and a signal is propagated from vertex i (\( v_i \)) to vertex j (\( v_j \)). To calculate RC delay in thermal variations, applying the increased resistances into Elmore delay model so that temperature dependent Elmore RC delay model becomes

\[ D'(V_i) = \sum_{k=1}^{n} r_k l_k \left[ \frac{C(e)}{2} (L - \sum_{k=1}^{n} l_{k-1}) + C(v_j) \right] \]  \hspace{1cm} (4)

where \( n \) is the total number of thermal grid cells overlapped with e, \( k \) is a thermal grid cell from vertex i, \( r_k \) and \( l_k \) are unit resistance and length of a \( k_{th} \) thermal grid cell respectively, \( C(e) \) is the wire capacitance, \( L \) is the total length of a net, and \( C(v_j) \) is the load capacitance. If buffer is inserted at vertex i, thermal dependent buffer delay \( D'(V_i) \) becomes

\[ D'(V_i) = R'(B_k)C(v_i) + K'(B_k). \]  \hspace{1cm} (5)

where \( R'(B_k) \) and \( R(B_k) \) are an increased resistance and a resistance of a buffer in \( k_{th} \) thermal grid cell respectively, and \( K'(B_k) \) is the thermal dependent intrinsic delay of the buffer, and they can be defined as

\[ R'(B_k) = R(B_k)[1 + \alpha(T' - T)], \]  \hspace{1cm} (6a)

\[ K'(B_k) = K(B_k)[1 + \beta_{in}(T' - T)]. \]  \hspace{1cm} (6b)

where \( K(B_k) \) is an intrinsic delay of a buffer, \( \beta_{in} \) is the temperature coefficient of the intrinsic delay. The coefficient is decided by the buffer size and the value is referred from [9].

Contrast to the original delay calculation, proposed one is a type of distributed RC model to include the thermal effects on a chip, i.e. resistances of wires and cells are recalculated based on (3), so the resistance per unit length (\( \mu \)) and the driving resistance of the inserted buffers are differently distributed along with each thermal grid cells. For the increased resistance caused by temperature, the interconnect delay and driving resistance of the inserted buffers are calculated for finding the optimal slack, e.g. when the source and sink blocks are arranged on a chip as shown in Fig. 1, the number of thermal grid cells overlapped the net is ten; thereby, the maximum number of the wire resistance of the unit length and the driving resistance of buffers are ten. Because each of the thermal grid cells has their own thermal information that already estimated, the wire resistance of the unit length and the inserted buffers in each grid cells can be recalculated, so the interconnect delay and the buffer delay can be calculated to find the optimal slack in thermal variations.

2. Thermal Aware Wire Segmentation

Buffers are selectively inserted of the candidate nodes, and the nodes are come from segmented wires divided by a net. The number and position of inserted buffers are dependent on those candidate nodes. For the most
effective use of thermal aware delay model in buffer insertion included thermal variations, it is important to include the thermal effect while segmenting wires.

Alpert et al. presented wire segmentation algorithm that find the optimal length of both ends of a net and optimal number of the inserted buffers between the ends of net with the same length [10]. If using the segmentation in thermal buffer insertion, the delay of each segmented wire becomes different. Each segmented wire has the same delay but different length as segmenting a net by the ratio between resistances at standard temperature and the increased temperature as follow

\[ l'_{\text{seg}} = \frac{r_k}{r_k} l_{\text{seg}}. \]  

(7)

where \( l'_{\text{seg}} \) is the length of segmented wire in thermal variations if \( l_{\text{seg}} \) is the length of conventional segmentation.

There is a candidate node in the branching point, as shown in Fig. 2(a). A net with 10 candidate nodes is illustrated in Fig. 2(b), length-based wire segmentation is shown in Fig. 2(c), and delay-based wire segmentation is represented in Fig. 2(d). The propagation delays of delay of point 1 (\( d_1 \)) to point 5 (\( d_5 \)) are the same.

3. Thermal Aware Cell Delay LUT Modeling

Technology library includes cell delay information for buffer insertion. Cell delay information is given as LUT for a number of different input slews and load (output) capacitances as shown in Fig. 3. Delay tables are generated using a detailed transistor-level circuit simulator SPICE. Many input factors including temperature are needed for SPICE simulation and all simulation for these factors is impossible because it needs much time. Actually, technology library gives delay tables for best, typical and worst corners. In this research, cell delay LUT modeling methodology is proposed for thermal variation aware buffer insertion.

The interpolation between two temperature corner’s LUT generates new delay LUT for dedicated temperature. Because each LUT has different LUT point about input slew and output capacitance as shown in Fig. 3(a) and (b). It means that the input slews are \( a, b, c \) and \( d \) for the LUT in Fig. 3(a) while the input slews are \( a, b', c \) and \( d' \) for the LUT in Fig. 3(b). As in Fig. 3(b), it is the reason that the interpolation between point \( a \) and point \( c \) to calculate the delay for point \( b \) and the interpolation between point \( e \) and point \( g \) to calculate the delay for point \( f \) are needed. After this process synchronizing input slew and output capacitance between two LUTs, we get delay table for dedicated thermal using the interpolation between LUTs as in Fig. 3(c). This LUT generation method reduces simulation time with plus or minus five-percent error because every SPICE simulation is not necessary and accuracy. The buffer insertion using these LUTs have higher efficiency.

IV. SIMULATION RESULTS

Buffer insertion included thermal effects which are...
changed resistances of driving cells and wires are implemented. Initially receiving the netlists of MCNC benchmarks as inputs and generating the floorplan results using by a tool ParquetFP developed in [11]. To extract the thermal distribution from the floorplanned MCNC benchmark circuits, Hotspot developed in [12] which is a tool estimates temperature in the floorplan stage is used with the 64 * 64 grid mode to analyze the thermal grid by grid.

Our target CMOS technology is 65 nm and all parameters are referred to the Predictive Technology Model (PTM) in [13]. According to PTM, wire resistance per unit length is 0.074 $\Omega/\mu m$ and its capacitance is 0.118 fF/$\mu m$. A fixed buffer type that has resistance 363 $\Omega$, and its load capacitance and intrinsic delay are 23.4 and 36.4 respectively at the standard temperature is used. The estimated results of floorplanned blocks with the thermal profile maps are shown in Fig. 4, and the range of the distributed temperature and their related resistances of each benchmark circuits are represented as shown in Table 1. Standard temperature is set to 293.15 K. Temperatures are increased in all circuits. Particularly, in case of ami33, because the circuit has small size but lots of blocks and high routing density, the range of temperature variation is the widest of the benchmarks, it leads to the most increments of the resistances and the variation range of them.

Once the thermal gradients on a chip are estimated based on grid by grid, Three methods of buffer insertions are implemented: conventional buffer insertion based on dynamic programming, the method applied thermal aware delay model, and the method applied thermal aware delay model and wire segmentation. 2 cases of those simulations are implemented by using a tool, Fast Buffer insertion (FBI), developed in [14], and we modified the tool by ourselves to apply the proposed delay model and wire segmentation, so that the resistance variation by distributed temperature is reflected in buffer insertion algorithm. For the first case, critical paths which have the longest time to propagate a signal from a source to sinks, typically long nets and nets have many sinks, are initially settled 10% of the total nets in each benchmark circuits. Both the conventional buffer insertion and the buffer insertion applied thermal aware delay model are implemented to verify how many critical paths more occur and how many buffers are more necessary for timing closure by the induced critical paths. The comparison results of both buffer insertion methods are shown in Table 2. More critical paths are induced in all of benchmark circuits implementing a buffer insertion included the effects of thermal variations than without included the effects. According to this result, if buffers are inserted in only room temperature, i.e. when inserting buffers without considering the effects of increased resistances caused by temperature along the interconnect lines, timing estimation can be failure due to the induced critical paths.

For the second case, a net of each benchmark circuit which has the worst delay is implemented by the three buffer insertion methods, after that how changed the number of buffers and propagation delay are verified. All implemented nets consist of a single source and many sinks and have long length. The comparison results are shown as Table 3. It is identified that the propagation delay is worse due to the thermal effects by the first case. When implemented by both thermal aware buffer insertions, more buffers are inserted than the conventional method whether thermal aware delay model and the proposed wire segmentation are applied to the conventional buffer insertion method; however, the propagation delay can be reduced by using the proposed wire segmentation with the same number of buffers as only applied thermal aware delay model. As a result of that, our proposed wire segmentation can search the temperature optimal position of buffers. The delay
reduction of the most benchmark circuits between wire segmentation applied method and the conventional method are 9 to 13%; particularly, a type of benchmark circuits such as ami33 that has small area but lots of blocks and high routing density is highly recommended to use thermal aware buffer insertion with the proposed wire segmentation because propagation delay can be reduced by approximately 33%, in contrast to reducing delay under 13% in the other benchmarks. This result shows temperature highly effects on the high integrated circuits.

For the accuracy of our study, we compare estimated delay and SPICE simulation. The error rate of thermal aware buffer insertion applied wire segmentation for SPICE simulation is about 5% in average.

Table 4 shows the cell intrinsic delay table. The SPICE simulation indicates timing results for dedicated input transition and output capacitance under specific temperature. We use Samsung logical library [15] and we denotes the SPICE timing value as 1 because it is confidential. The results of interpolation are calculated

| Table 1. Initial values and variation range of increased temperature and wire resistance of each benchmark circuit |
|---|---|---|---|---|---|---|---|---|
| ami33 (area: 1184*1176) | ami49 (area: 6370*6412) | apte (area: 7602*6844) | hp (area: 3304*3304) | Xerox (area: 4641*4893) |
| Initial value | 293.15 | 0.074 | 293.15 | 0.074 | 293.15 | 0.074 | 293.15 | 0.074 |
| Variation range | 396.17 – 871.07 | 0.104 – 0.24 | 310.15 – 375.63 | 0.079 – 0.098 | 303.61 – 335.13 | 0.077 – 0.086 | 306.63 – 479.14 | 0.078 – 0.128 |

| Table 2. Increased critical paths and the resultant number of inserted buffers on chips distributed temperature |
|---|---|---|---|---|
| ami33 (total nets: 123) | ami49 (total nets: 408) | apte (total nets: 97) | hp (total nets: 83) |
| Critical Paths buffers | Critical Paths buffers | Critical Paths buffers | Critical Paths buffers |
| Standard temperature | 12 | 109 | 41 | 337 |
| Distributed temperature | 19 | 198 | 63 | 561 |
| Increased rate [%] | 37 | 45 | 35 | 40 |

| Table 3. Comparison the number of inserted buffers and delay after implementing three buffer insertion methods and delay reduction between the conventional buffer insertion and thermal aware buffer insertion applied wire segmentation |
|---|---|---|---|---|
| Conventional buffer insertion | Thermal aware buffer insertion | Thermal aware buffer insertion applied wire segmentation | Error rate [%] (SPICE simulation) |
| ami33 | 23 | 841.26 | 29 | 647.12 | 29 | 562.63 | 5.4 |
| ami49 | 26 | 1900.78 | 33 | 1771.46 | 33 | 1653.97 | 4.7 |
| apte | 7 | 1078.67 | 11 | 1014.78 | 11 | 991.26 | 3.6 |
| hp | 6 | 605.16 | 9 | 569.83 | 9 | 547.51 | 5.3 |
| xerox | 21 | 824.95 | 24 | 776.79 | 24 | 748.59 | 6.1 |

| Table 4. Comparison the proposed LUT generation and SPICE simulation for Samsung cell library |
|---|---|---|
| Specific Temperature with input transition and output capacitance | Proposed interpolation | SPICE |
| Cell 1 | 0.964 | 1 | 3.6 |
| Cell 2 | 0.977 | 1 | 2.3 |
| Cell 3 | 1.028 | 1 | 2.8 |
| Cell 4 | 0.981 | 1 | 1.9 |
| Cell 5 | 0.989 | 1 | 1.1 |
| Cell 6 | 1.015 | 1 | 1.5 |
| Cell 7 | 1.042 | 1 | 4.2 |
| Cell 8 | 0.954 | 1 | 4.6 |
| Cell 9 | 1.058 | 1 | 5.8 |
| Cell 10 | 0.974 | 1 | 2.6 |
by proposed LUT generation and the circumstance is same with SPICE simulation. The results are denoted as ratio of calculated delay to SPICE simulated delay. The error rate of proposed LUT generation method is under 5.8%.

V. CONCLUSIONS

This paper presents thermal aware buffer insertion in early stage of physical design with the proposed delay model and the thermal aware wire segmentation. Thermal aware LUT modeling for cell intrinsic delay is also proposed. Simulation results show the length of wire and the range of temperature are heavily effect on the propagation delay, so that critical paths are induced 35 to 58% in the simulations. Buffer insertion by applied the proposed delay model and wire segmentation can produce the delay reduction up to the 33%. The error rate of LUT modeling to SPICE simulation is under 6%.

ACKNOWLEDGMENTS

This work was sponsored by ETRI SW-SoC R&BD Center, Human Resource Development Project. This research was also supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2012-H0301-12-1011).

REFERENCES


Jaehwan Kim received the B.S. degree in Computer Science from Hanyang University, Seoul, Korea, in 2006. He is currently in the unified course of M.S. and Ph.D. degree in Electronics and Computer Engineering from Hanyang University. His current research interests are VLSI & CAD as well as SoC design methodology, especially 3D IC physical design methodology.

Byung-Gyu Ahn received the B.S. degree and the M.S. degree in Electronic and Computer Engineering from Hanyang University, Seoul, Korea, in 2003, and 2005 respectively and the Ph.D. degree in 2012. His current research interests are in the physical design automation of VLSI circuits with a special emphasis on clock/power network synthesis, timing analysis, and 3D IC design methodologies.

Minbeom Kim received the B.S. degree in Electronic Engineering from Wonkwang University, Iksan, Korea, in 2008. Currently He is working toward the M.S. degree in Electronics and Computer Engineering at Hanyang University, Seoul, Korea. His current research interests are the physical design of VLSI circuits, especially timing analysis in 3D IC.

Jongwha Chong received the B.S. and the M.S. degree in Electronics Engineering from Hanyang University, Seoul, Korea, in 1975, and 1979 respectively and the Ph.D. degree in Electronics & Communication Engineering from Waseda University, Japan, in 1981. Since 1981, he has been a professor of the Department of Electronics Engineering, Hanyang University. From 1979 to 1980, he was a researcher in C&C Research Center of Nippon Electronic Company (NEC). From 1983 to 1984, he was a visiting researcher in the Korean Institute of Electronics & Technology (KIET). In 1986 and 2008 respectively, he was a visiting professor at the University of California, Berkeley, USA. He was the chairman of CAD & VLSI society in 1993 and President in 2007 respectively at the Institute of the Electronic Engineers of KOREA (IEEK). He was the Director of Institute of Information and Communication Center in 1997 and Dean of Graduate and Undergraduate School of Information and Communications in 1999 at Hanyang University. He is currently the President of KIEEE (Korean Institute of Electrical and Electronics Engineers) and the Chairman of Fusion SOC Forum. His current research interests are of SoC design methodology including memory centric design, indoor wireless communication SOC design for ranging and location, Video system and Power IT system.