A Floorplan-based Planning Methodology for Power and Clock Distribution in ASICs

Joon-Seo Yim, Seong-Ok Bae

DSP Group, Information Technology Lab.
LG Corporate Institute of Technology
16, Woomyeon-Dong, Seocho-Gu,
Seoul, 137-140, Korea
e-mail: jsyim,sobae@lotus.lgcit.com

Chong-Min Kyung

Department of Electrical Engineering
KAIST
373-1, Kusong-Dong, Yusong-Gu,
Taejon, 305-701, Korea
e-mail kyung@ee.kaist.ac.kr

Abstract

In deep submicron technology, IR-drop and clock skew issues become more crucial to the functionality of chip. This paper presents a floorplan-based power and clock distribution methodology for ASIC design. From the floorplan and the estimated power consumption, the power network size is determined at an early design stage. Next, without detailed gate-level netlist, clock interconnect sizing, the number and strength of clock buffers are planned for balanced clock distribution. This early planning methodology at the full-chip level enables us to fix the global interconnect issues before the detailed layout composition is started.

1 Introduction

As the CMOS technology enters the deep submicron(DSM) design era, the interconnect density combined with the increase in wiring levels and the growth in chip size make the interconnect design the most challenging area in DSM technology. An interconnect design methodology considering cross-coupling noise, signal integrity, IR-drop, clock skew, antenna, and hot electron effect, therefore, becomes crucial to the working silicon[1, 2, 3, 4].

In the DSM process, the metal width tends to decrease with the length increasing due to the complex system integration into single silicon. Therefore, the resistance along the power metal line increases. In addition, due to the nonlinear scaling of threshold voltage compared to power supply voltage, IR-drop becomes more serious. Due to the IR-drop, supply voltage at the transistor cell is no longer ideal reference, which weakens the driving capability of logic gates, slows down the circuit, and reduces the noise margin. Typically, it is known that 5% drop in supply voltage increases the overall delay as much as 15% or more. Delay in clock buffers has been known to be increased by above 100% clock due to the poorly controlled IR-drop. Such an increase in delay is critical when we are managing clock skew within 100ps. Once the noise margin drops below the budgeted amount, typically 10%, the design is not guaranteed to operate properly[5, 6].

Most of designs, therefore, limit the IR-drop on the chip within 10% of the total supply voltage. For the cost performance chip working more than 100 MHz, it is estimated that as much as 10% of silicon area may be needed to serve the global power supply network. It is, therefore, important to estimate and allocate the area needed for the power distribution network during the early floorplan stage. Especially, in low cost ASIC designs, the number of metal layers is limited to reduce the number of masks and the process cost. It is, therefore, impossible to assign entire single layer metal to the power/ground distribution network on the contrary to the high-performance microprocessors[7]. The available routing resources, therefore, must be carefully balanced among the signal, power/ground, and clock distributions.

The skew in the clock edges resulting from power supply noise, process variation, and interconnect RC delay introduce uncertainty in the synchronizing element's timing which further limit the performance and can lead to the functional failure due to the hold time violation. As a rule of thumb, a clock skew budget is usually determined by $skew = \left[\frac{Cycle\_Time}{20}\right]$[8, 9]. Various efforts such as H-tree, binary tree, Steiner tree, and optimal buffer/wire sizing schemes have been proposed to minimize the clock skew[8, 10, 11, 12]. In the complex chip designs with many functional blocks, the clock tree is not well balanced with automatic clock tree synthesizer. Balancing the top-level clock skew is, therefore, known as circuit designer's task requiring much engineering skill and experiences rather than just CAD tool's support.

As the design complexity exceeds the several million transistors on chip, it is reasonable to start looking for IR-drop problems and clock skew issues earlier in the design cycle. Currently, most of commercially available tools are focused on the post-layout verification when the entire chip design is completed and detailed layout and current information are known. But the power network and clock skew problem at the final stage are usually very difficult or expensive to fix. We, therefore, propose a floorplan-based planning methodology which progressively refines the layout as the design progresses for both the power and clock distribution. The proposed power and clock distribution methodology was applied to H2SD480i, a HDTV single chip MPEG-2 audio/video/system decoder, which receives digital transport packet data from VSB/FEC block or IEEE 1394 port, parses the TP packets, and decodes video and audio bit streams, developed at LG CIT[13]. This chip was fabricated in 0.35μm 4 metal layer process and tested successfully in the working system.

In section 2, we describe the power distribution methodology with experimental results. Section 3 deals with the clock distribution methodology.

2 Power Distribution

The optimization of power distribution network involves an iteration process between simulation and resizing. Given the current specification and location of each functional blocks,

DAC 99, New Orleans, Louisiana
(c) 1999 ACM 1-58113-109-7/99/06..$5.00

the circuit simulator analyzes the switching noise on the power metal, identifies the worst case IR-drop, and determines the width of power metal. In the traditional approaches, both the global and local power grid are extracted from the actual layout after the P&R is completed, and the current source can be modeled at the transistor or gate level. It takes large simulation time to simulate and analyze such full-chip netlist. Traditional approach, therefore, limits the application of this approach to a block-level rather than full-chip level. Moreover, for the ECO(Engineering Change Order) of the power distribution network, it needs expensive P&R efforts as shown in Fig. 1(a). On the contrary, in the proposed approach, the global power trunk and block level local power network are planned from the floorplan with the area-based current estimation. The power width can be easily changed during the optimization loop as shown in Fig. 1(b). In optimizing the power distribution network, full chip is modeled by a resister network and current sources. They are described in the following sections.

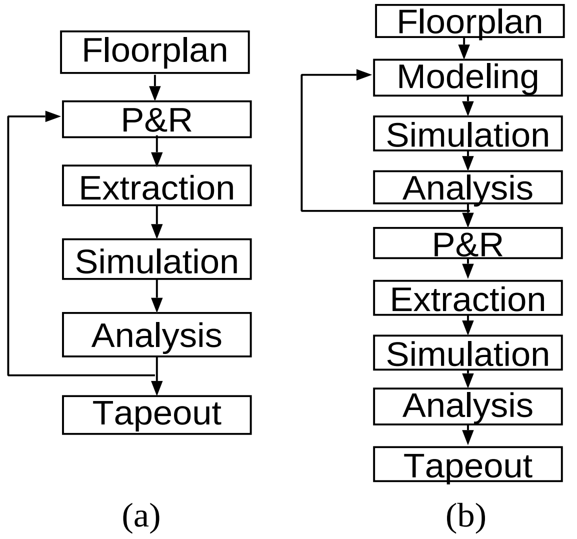

Figure 1: *Comparison of two power distribution approaches between (a) traditional and (b) proposed approach.*

2.1 Power Network

In order to reduce the complexity of full chip analysis, a hierarchical approach is used to build the power network model. At the full-chip level, a global power trunk is generated to subdivide the chip into several blocks as shown in Fig. 2(a). All the switching activities within one block are lumped together, and adjacent cells are connected by the global power trunks. At the block level, where local hot spots are located, a finer grid will be generated to model the detailed power network structure. The width of global power trunk is determined first, then the block-level vertical power refreshes and horizontal power busses are determined. Initially, the width of power trunk is determined by 15% of block size.

The standard cell block is modeled by a resister network as shown in Fig. 3. The number of the power refresh and their width will be determined iteratively. The size of horizontal power bus is pre-determined by the standard cell layout. The number of rows are determined by the block size with 30% overhaed for the routing track. After we calculate the resistance for each power segment, an equivalent SPICE netlist is generated.

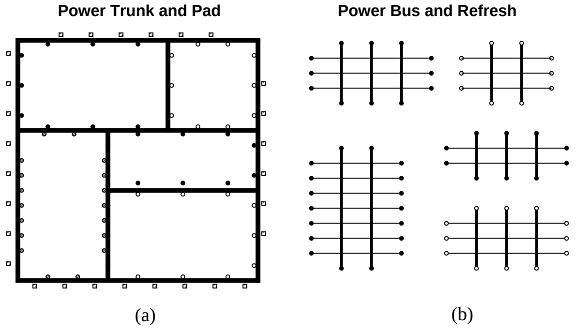

Figure 2: *From the floorplan, two resistive networks are generated : (a) full-chip level global power trunk and Vdd/Gnd pad (b) block-level power grid consisting of power bus and refresh.*

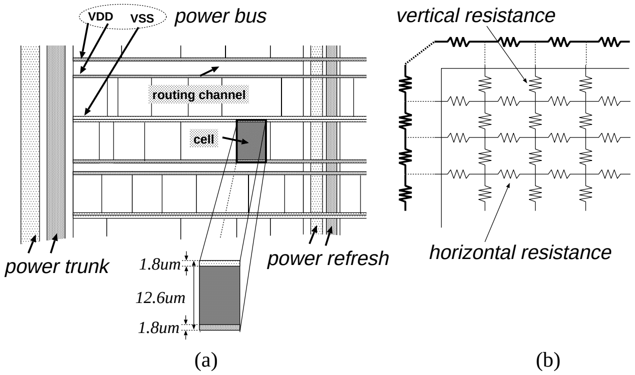

Figure 3: *(a) Schematic view of standard cell block and (b) its power grid model : Power bus(refresh) is modeled by horizontal(vertical) resistance.*

2.2 Current Source

Finding the worst case IR-drop depends on the maximum current flowing in the power supply network. Each point in the power grid may exhibit a worst-case IR-drop for different vectors. The maximum current in the block depends on the timing of the circuit and current drawn by individual gate. Even though the simulation approach is more exact than the static approach, several sets of test vector do not guarantee to cover the worst case IR-drop. In addition, long test vector sequences need long simulation time. There should be trade-off between the speed and the accuracy. In determining the maximum current at the planning stage, the extreme accuracy is not needed. In our approach, the summation of each gate's current is used to obtain block current for early analysis.

Initially, the current is estimated from the block size and the transition activity. To model the current drawn by the chip, a simple area-based DC estimates of the current is used. Usually, the maximum instantaneous current is expected to be 3–7 times the average current[6]. That is, $average\ current = \alpha \times block\ size \times transition\ activity$. $peak\ current = \beta \times average\ current$. This peak current is evenly distributed among the current source corresponding to each power grid in the standard cell as shown in Fig. 4. As the design proceeds, more exact values are obtained using the power simulator.

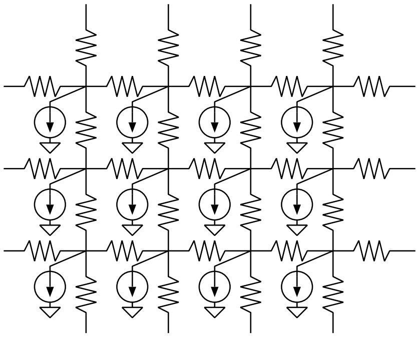

Figure 4: *Current source model* I_{grid} *corresponding to 4×3 grid standard cell with 4 vertical power refreshes and 3 horizontal power busses.* $I_{grid} = I_{block}/(number\ of\ current\ sources)$.

Since the power supply voltage in one region is affected by the switching activities in the neighboring regions, the equivalent circuits for each functional block are connected to the adjacent global power trunk to ensure the accurate analysis results.

2.3 Design Flow

Fig. 5 shows the design flow of power distribution network. In optimizing the power network, three *perl* scripts are developed. Reading the floorplan description file including the geometries of PAD, trunk, refresh, and block-level power consumption, *netlister* generates the resister network corresponding to the power trunk and current source for each power grid network. After the HSPICE simulation is finished, *analyzer* reports the IR-drop value for each block and displays the hot spot in the chip. *optimizer* iteratively updates the width of power trunk and refresh for the next SPICE simulation. The iteration stops when the maximum IR-drop voltage is within the given IR-drop budget.

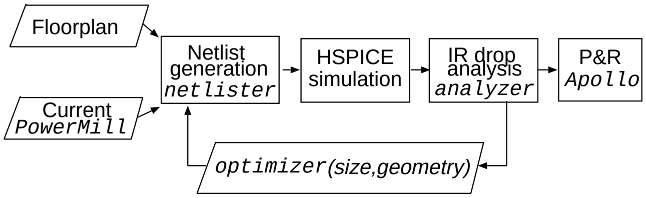

Figure 5: *Design flow of the power distribution network.*

2.4 Experimental Results

Fig. 6 shows the IR-drop value for the initial floorplan. The 8×8 mm^2 chip is subdivided into 6 blocks, corresponding audio decoder, display controller, system parser, IDCT(Inverse Discrete Cosine Transform), VLD(Variable Length Decoding), motion compensation, inverse quantization, and main control block. For the initial power distribution with 3.3V power supply voltage, the worst IR-drop value is 0.45V, which is larger than the budget. The width of the power trunk has a big influence on the IR-drop as shown in Table 1. The effect of the power refresh width on the maximum IR-drop is small as shown in Table 2.

With total average current 2.7W, the peak power of the chip 8W(three times the average power of 2.7W) is used for the IR-drop analysis. By dividing the total current by the estimated chip area, we get the current per unit area. For the final power distribution network, the IR-drop value is shown in Fig. 7. The worst case IR-drop is 0.302V at the center of chip. The main power trunk width is sized to 120μm and the width of power refresh is 60μm or 30μm.

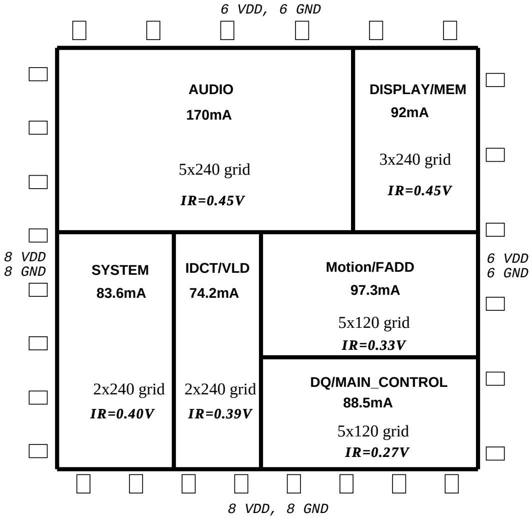

Figure 6: *Resultant IR-drop with initial power distribution, where the power trunk width is 120μm and the power refresh width is 30μm. V × H grid for each block implies that there are V vertical power refreshes and H horizontal power busses.*

Table 1: *IR-drop vs. the width of power trunk*

trunk (μm)	60	80	100	120	140
IR-drop(V)	0.53	0.43	0.38	0.30	0.29

Table 2: *IR-drop vs. the width of power refresh with 100μm power trunk.*

refresh (μm)	20	30	60
IR-drop(V)	0.38	0.35	0.34

3 Clock Distribution

Clock is distributed in the hierarchical way as shown in Fig. 8. The hierarchy consists of the following five level clock drivers.

- L1 : PLL's output driver
- L2 : central driver at the chip center
- L3 : region driver for each block
- L4 : mezzanine driver between the region driver and the final driver
- L5 : final driver which actually drives the flip-flops

At the full chip viewpoint, level 2 clock balancing is more important than the detailed level 3, 4, and 5 clock skew minimizing. Level 2 clock distribution tree looks like global H-tree as shown in Fig. 8. The central L2 clock buffer drives five L3 region drivers, called *NW(north west), NE(north east), SW(south west), and SE(south east).* Within the block level, commercial clock tree synthesis tool is used.

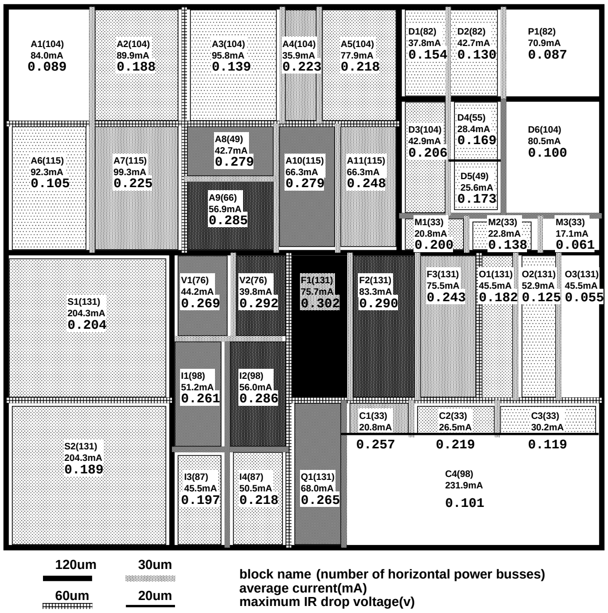

Figure 7: *Final IR-drop map where the worst case IR-drop is 0.302V at block F1(at the center).*

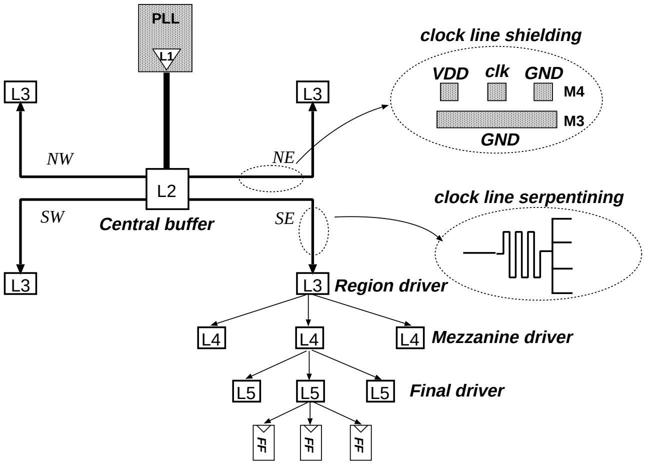

Figure 8: *Hierarchical clock distribution scheme.*

3.1 Top Level Clock Tree

In DSM interconnect-driven designs, the clock loading is limited by interconnect RC rather than the gate capacitance of flip-flops. Therefore, exact three-dimensional RC extraction considering both the inter-wire and the inter-layer coupling is crucial for the exact skew analysis[14, 15, 16]. At first, large data are extracted from the actual layout with variable width and space using 3D extractor Raphael[16]. The coupling capacitance between the table is linearly interpolated using the following equation:

$$Cap_{coupling} = (c_1 + c_2 * W + c_3 * \frac{1}{S} + c_4 * \frac{W}{S}) * L * f_{cc}$$

where W is the width of wire, S is the space between the wire, L is the length of wire, and f_{cc} is the coupling factor. Note that the coupling capacitance between a pair of parallel wires is proportional to their overlapped length, and is inversely proportional to their separating distance. Linear extrapolation is used for width and space that exceed values in the table.

Balancing the clock at the top level more depends on the optimal interconnect parameters such as wire length, width, and space. As the wire width increases, the resistance decreases, while the capacitance increases. Therefore, for a given wire space, the interconnect delay decreases and then increases as the wire width increases - *i.e.*, it has an optimal point as shown in Fig. 9(a). The effect of the inter-wire space on the performance is shown in Fig. 9(b). Widening the distance between the wires helps minimizing the cross-coupling delay. Therefore, the interconnect delay decreases as the space between wires increases. For the same pitch, increasing the wire width (and, therefore, decreasing the wire space) tends to increase the interconnect delay as shown in Fig. 9(c). Therefore, for a given pitch, it is better to increase the wire space than the wire width.

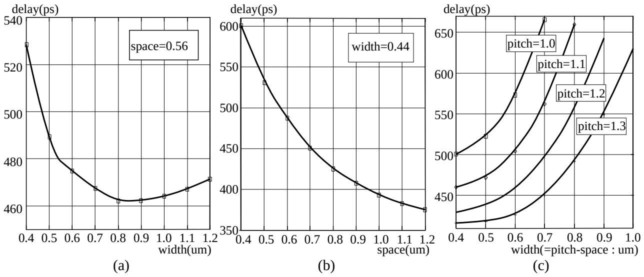

Figure 9: *The interconnect delay for variable wire width and space : (a) delay vs. width (b) delay vs. space (c) delay vs. width for the same pitch.*

For a driver-limited case, as we increase the driver size, the delay improves, but for the RC-limited case(*i.e.*, for long running wire), the performance is not improved even though we increase the driver size. By carefully placing repeating buffers in RC-limited path, it may be possible to reduce the overall interconnect delay. Since a buffer introduces some intrinsic delay, the length of the associated interconnect needs to be chosen such that the use of repeating buffers provides a less total delay than the signal delay without buffers. The wire and central L2 buffer size are iteratively determined to balance the clock delay for each region. For the fine tuning of the clock skew, dummy loading and clock serpentining are used for the unbalanced blocks or interconnect as shown in Fig. 10.

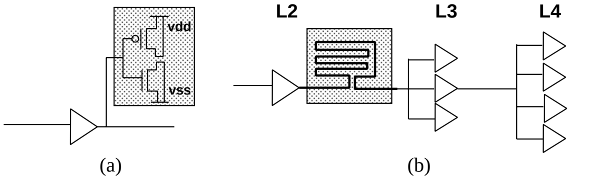

Figure 10: *Load balancing by (a) dummy cell and (b) wire serpentining.*

3.2 Block Level Clock Tree

To balance the clock, we need to know the number of flip-flops and their positions. In the early planning stage before the detailed clock tree synthesis and actual P&R are started, the exact number of clock buffers, flip-flops, and their positions are unknown. Fish-bone based clock tree examples using Apollo clock tree synthesizer(CTS) from Avant![17] are shown in Fig. 11. As an engineering approximation, we might assume that flip-flops and clock buffers are uniformly distributed within block. Fig. 12 shows that our approximation is very similar to the final clock tree.

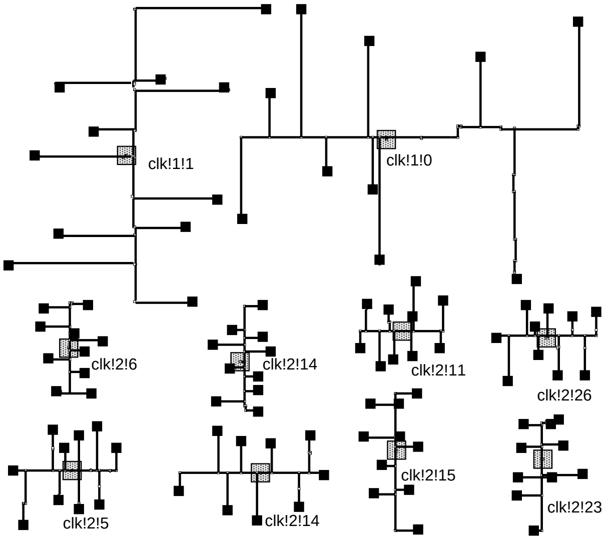

Figure 11: *Example clock tree from Apollo CTS*

The block level clock network is generated based on the model as shown in in Fig. 13. Each L_i clock buffers drive L_{i+1} clock buffers within sub-region. The clock tree modeling script generates SPICE netlist including L3,L4,L5 buffers and flip-flops. To reduce the simulation time, only the longest and the shortest RC paths are included in the netlist and the other loads are modeled by equivalent capacitances. Given the block size and the number of flip-flops, the wire and the gate loading are automatically determined. Then optimal number of clock buffers and drive strength are iteratively determined by the simulation as shown in Fig 14. When the clock skew is larger than the budget①, it requires top-level adjusting② or block-level adjusting③ according to the mismatch amount. After the P&R, post layout skew optimization④ follows.

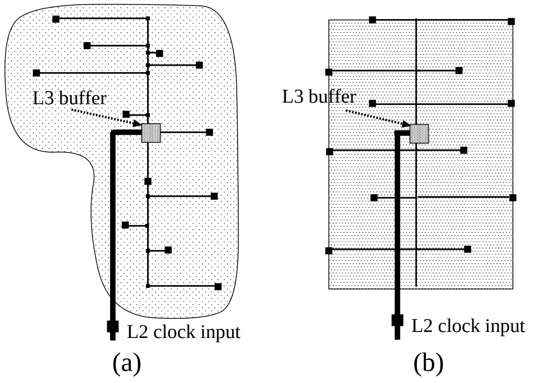

Figure 12: *Block-level clock tree model : (a) a synthesized clock tree by Apollo and (b) an estimation model based on uniform distribution.*

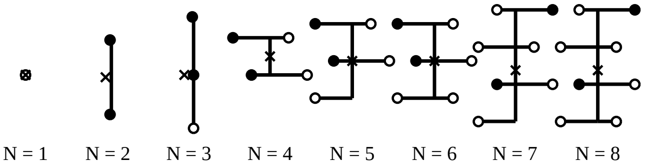

Figure 13: *Block level clock tree model : Only two buffers with black circles corresponding to the minimum and maximum delay are needed for the clock skew report. N is the number of (i+1)-th level clock buffers. Symbol* `X` *is input node, where the previous i-th level clock net is attached. Black circle(●) represents the nearest and the farthest buffer or flip-flop from the input node. White circle(○) is the load to be modeled by equivalent capacitance.*

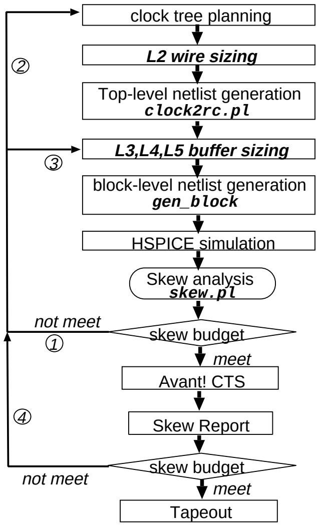

Figure 14: *Design flow of the clock distribution network*

3.3 Experimental Results

In this experiment, clock tree and buffers are optimized to meet the skew budget within $200ps$. Most of clock skew results from L2 network. Therefore, L2 balancing is the most important among all the clock levels. L2 clock routing and clock line width and spacing are shown in Fig. 15. Since the wire capacitance dominates the gate capacitance in L3 clock tree, the block size gives large impact on the clock delay rather than the number of flip-flops. Henceforth, AUDIO block requires three L3 buffers even though the number of flip-flops is small. In passport ASIC library, there are four buffer cells with different strength such as `buf_d1`, `buf_d3`, `buf_d7`, and `buf_dA`. In most cases, the strongest buffer, `buf_dA` is selected. For small blocks such as VLD, display, memory controller, `d3` and `d7` buffers are used to intentionally increase the clock delay.

Table 3 shows the clock balancing results. For the standard cell area excluding memory and the number of flip-flops, the final clock buffer strength, the number of clock buffers, and the resultant minimum/maximum clock delay are shown for each block. Fig. 16 shows the final chip layout. This chip was fabricated using 0.35μm CMOS technology and works successfully in decoding the full HDTV bit stream.

Table 3: *Block-level clock buffers and skew result : full chip skew 0.209 ns (= 2.277-2.068). FF=*$N_3 \times N_4 \times N_5 \times N_{FF}$

block	size(μm)		# of	clock buffer strength			# of buffers or FF's				full chip delay(ns)	
	x	y	FF's	L3	L4	L5	N_3	N_4	N_5	N_{FF}	min	max
AUDIO	4500	1476	1883	A	A	A	3	6	10	11	2.123	2.262
SYSTEM	1470	4400	3328	A	A	A	3	10	10	11	2.090	2.256
VLD	1496	1257	713	7	7	7	1	8	9	10	2.218	2.277
IDCT	1500	2264	1000	A	A	A	1	10	10	10	2.166	2.237
DISPLAY	1500	2500	2004	A	7	7	2	10	10	10	2.160	2.214
MEM_control	2500	305	200	7	7	7	1	2	10	10	2.114	2.136
FADD	2700	1354	880	A	A	A	1	8	10	10	2.116	2.199
MOTION	1530	2164	1406	A	A	A	1	11	11	12	2.068	2.187
MAIN_control	3711	930	718	A	A	A	1	8	9	10	2.154	2.217
DQMD	520	930	389	3	3	7	1	7	7	8	2.215	2.234

4 Conclusion

In this paper, we proposed the *floorplan-based early planning methodology* of the clock and power distribution. The power and clock distribution network can be determined accurately enough at an early design stage thus avoiding the expensive iterations including placement and routing.

By optimizing the geometry and the width of power trunk and refresh, we were able to limit the maximum IR-drop within 10% of power supply voltage. In addition, we reduced the clock skew within 200ps in 0.35 μm CMOS technology by optimizing the interconnect size and buffer strength iteratively.

As future works, area-based current estimation model could be replaced by more exact dynamic current model from the full-chip multi-cycle power simulation. In addition, the clock skew minimization targeting for the design using multiple clocks with different frequency becomes crucial for the *system-on-chip* design.

Acknowledgement

One of authors, Yim, cordially expresses his gratitude to Ed Pak who provides many helpful discussions at SandCraft, deep submicron study group in KAIST, and Sangjin and Yonghwan in LG semicon for their help.

References

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

[2] William E.Guthrie *et al.* "Noise and Signal Integrity in Deep Submicron Design", *Proc. 34th DAC*, pp.720–721, 1997

[3] David Blaauw, "IR-Drop Analysis Signal Net Noise Analysis", *Proc. 34th DAC*, Tutorial, 1997

[4] Howard H. Chen and David D.Ling, "Power Supply Noise Analysis Methodology for Deep-Submicron VLSI Chip Design", *Proc. 34th DAC*, pp.638–643, June, 1997

[5] G.Steele *et al.*, "Full-Chip Verification Methods for DSM Power Distribution Systems", *Proc. 35th DAC*, pp.744–749, June, 1998

[6] A.Dharchoudhury *et al.*, "Design and Analysis of Power Distribution Networks in PowerPC Microprocessors", *Proc. 35th DAC*, pp.738–743, June, 1998

[7] P.E.Gronowski, "High-Performance Microprocessor Design", *IEEE JSSC*, Vol.33, no.5, pp.676–686, May 1998

[8] Y.Shimazu, "High Speed Clock Design", *ASP-DAC Tutorial*, pp.40–53, 1997

[9] H.Fair and D.Bailey, "Clocking Design and Analysis for a 600MHz Alpha Microprocessor", ISSCC Digest of Technical Papers, pp.398–399, 1998

[10] M.Edahiro, "Delay Minimization for Zero-Skew Routing", *ICCAD-93*, pp.563-566, 1993

[11] J.G.Xi *et al.*, "Useful-Skew Clock Routing with Gate Sizing for Low Power Design", *Proc. 33th DAC*, pp.383-388, 1996

[12] C.P.Chen *et al.*, "Fast Performance-Driven Optimization for Buffered Clock Trees Based on Largrangian Relaxation", *Proc. 33rd DAC*, pp.405-408, 1996

[13] "H2SD480i HDTV all-format single chip decoder for HDTV Settop box & PC Add-on card for HDTV receiving", Rev.0.3, Preliminary Specification, LG CIT, 1998

[14] J.Cong *et al.*, "Analysis and Justification of a Simple, Practical 2 1/2-D Capacitance Extraction Methodology", *Proc. 34th DAC*, pp.627–632, June, 1997

[15] F.Dartu and L.T.Pileggi, "Calculating Worst-Case Gate Delay Due to Dominant Capacitance Coupling", *Proc. 34th DAC*, pp.46–51, June, 1997

[16] "Raphael NES", TMA, 1997

[17] "Apollo Fundamentals Training Guide", Avant!, 1997

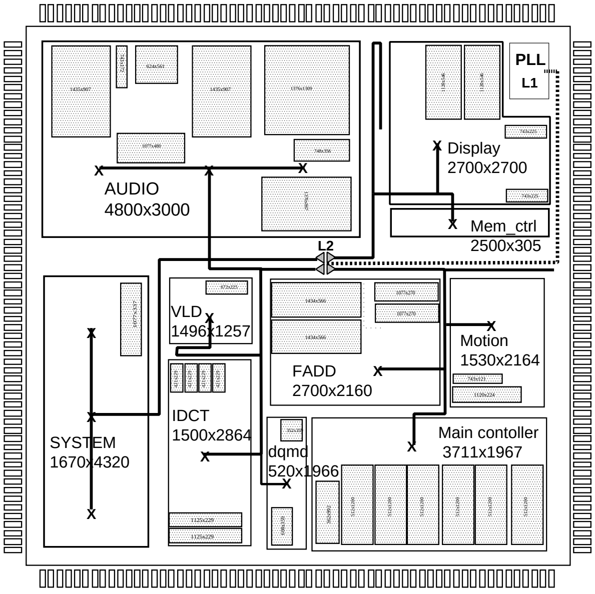

Figure 15: *Level 2 clock interconnect is sized to width=0.7μm and space=1.0μm. X denotes L3 clock buffer. Shaded blocks are memory where clock nets are not routed.*

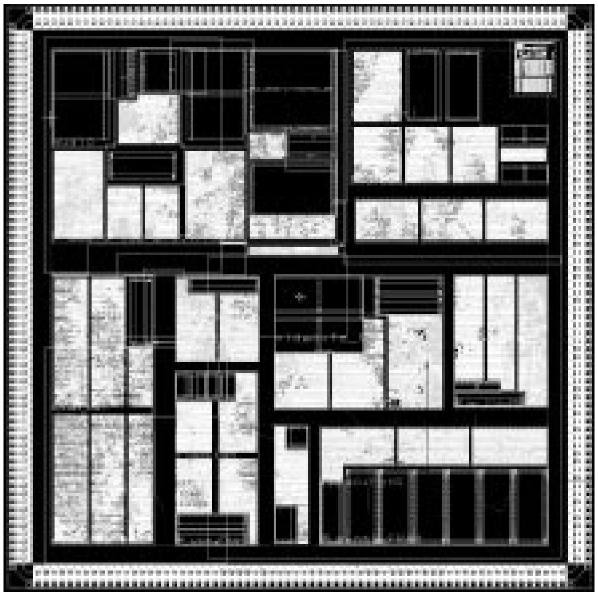

Figure 16: *Plot of final chip layout*