Timing And Power Optimization In Digital VLSI Circuits

G Poornima
D*ept.of ECE, BMSCE*
Bengaluru, India
gpoornima.ece@bmsce.ac.in

Chandrika Uttarkar
D*ept. of ECE, BMSCE*
Bengaluru India
uttarkar.chandrika@gmail.com

Shiv kumar Tomar FET, MJP
D*ept.of ECE, Rohilkand university,*
Bareilly, India
shivktomar@gmail.com

***Abstract*— Ever year microprocessor core is shrinking in size owing to the advancement in technology and decrease in technology node. The new challenges include more stringent and rigorous performance targets compared to previous project. Performance targets include timing, power as well as area optimization. With every new design of microprocessor core there can be a requirement of new techniques to be added to the existing once. The main goal is to improve Timing as well as Power so as to achieve higher quality of design through cell resizing, logic optimization, clock tuning, CTS, dual/quad insertion. Operating frequency and voltage are fixed for each project. Activity Factor is a constant quantity depending on the design. ssThe optimization methods mentioned in this work, focuses on reduction of dynamic capacitance and the resistance to reduce the dynamic power of the circuit. These techniques mentioned in this paper are tested and results analyzed**

Keywords—Timing convergence, logic optimization, Power optimization, Clock tree synthesis, clock pull, Dual/Quad insertion technique, merging of RCB/LCB's.

I. INTRODUCTION

The continuous growth in the semiconductor industry demands new techniques to balance the complexity in designs. Advanced microprocessor design is also one of the emerging technology in which different Intellectual property blocks (IP blocks) are obtained from different vendors and combined with the heart of microprocessor [1]. Fig. 1.1 shows the high level diagram of multicore microprocessor chip. The hardware of advanced microprocessor design basically consists of different IP blocks for some specific functions, bus interface circuits, memory blocks and other standard library cells. Its software components includes real-time operating systems, device drivers and library functions of standard cells. All the characteristics of the standard cells are derived from the standard cell libraries.

A microprocessor can typically have multiple cores that fetch, read, decode, execute, and dispatch microprocessor instructions. Apart from the core, the processor SoC has an on-chip high-level cache, I /O and memory controller, and an integrated graphics device. Each SoC and subsequent core is designed for specific market requirements. These goals are periodically different which is depend on focusing area and market competition. Intel has always had upper hand over its competitor when it comes to performance in terms of frequency. However, there is a need to increase the power performance of Intel processors. In this paper, we are aiming to improve the power efficiency of the device.

In Core designs bottom to top approach is followed. To make design and performance analysis easy, it is necessary to divide the design into smaller parts. As a result, core has to be divided into clusters at top level. The clusters are nothing but parts of processor performing the instruction fetch and also it's decoding. Some other major cluster functions include execution of instructions, operations on memory cache and interface. Different clusters are then split into sections. Sections are branched out into number of units. Integer number execution, generation of address, floating point arithmetic, etc. are what is called units and finally the leaf cell in hierarchy is called Functional Unit Blocks (FUB) shown in figure 1.1.

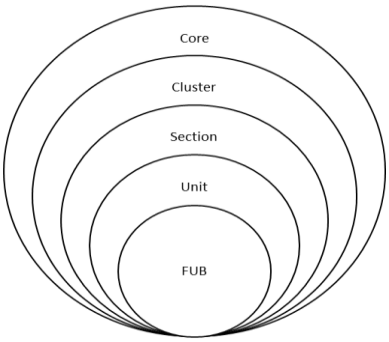

Figure 1.1 Core Design Hierarchy

The most basic cell in the hierarchy is what we call as the Functional Unit Block (FUB). Examples of FUB are multipliers, adders, dividers, repeaters and register files. These FUBs are designed individually first and are later combined in hierarchical way. All designs and implementations are performed at this FUB level. Integration of the various functional unit blocks is done at the section level. FUB is the smallest part of the microprocessor core that can be characterized by RTL code.Since, FUBs are smaller in size they provide advantage in design convergence. The main advantages with functional unit blocks include reduced complexity of core design, reusability of blocks, ease of study in terms of frequency and power [2].

II. LITERATURE REVIEW

Processors are the central component of computers, but in today's world, we can find processors embedded in many other devices such as cars, consumer electronic devices, games and many SoC devices. The architecture of processors has evolved continuously when compare to previous decade. For example, Intel recently released a new microprocessor almost every year [3]. This evolution is motivated primarily by two factors: Moore's Law (transistor scaling) and workload evolution. Optimize performance by introducing a new feature set on a single chip after every process changes. In accordance with Moore's law, this has magically caused chip frequencies to rise and prices to drop [4]. An operating frequency changes with each versions of design. As frequency increases number of negative paths in the design increases, for making these negative paths to positive paths a technique called Static Timing Analysis (STA) is applied. Techniques for making negative margin paths to positive margin are explained in [5] [6].

The power management of the chip is becoming a big problem. Because of increased design complexity, which occurs below nodes less than 100 nm. In addition, leakage current plays an important role in low-power VLSI devices. Leakage and dynamic power consumption in submicron technology is an essential design parameter as it dissipates a significant portion of the overall power consumption. Leakage and dynamic power reduction are emerging as the main goals of power management technology described in [7], which are used in VLSI circuit design, various methodologies, strategy and low power circuit design to increase the battery life of mobile devices. Dynamic Power Management is the most efficient way to change when device status is not operating at full speed or full capacity [9].

Perhaps one of the best-known low-power technologies, clock gating is very effective in reducing the power consumption of digital and VLSI circuits. The designer goes through several iterations to optimize the power to achieve the power budget. Although power must be optimized at every stage of the design flow, optimization at the initial design stage has the greatest effect on reducing power [10]. The goal of the clock gating technique is to paralyze the transition to a portion of the clock path (e.g., flip-flop, clock network and logic) under certain conditions calculated by the clock gating circuit, explained in [11].

The clock tree synthesis responsible for high performance. The clocking network must implement a special trade-off, active buffer distribution in designs and technologies that switch skew sensitivity, clock insertion delay and simultaneous power loss. CTS is a layout technique that generates a dynamic buffer distribution from a source pin to several receiver pins which are clearly explained in [12].

III. PROBLEM SATEMENT

Before starting the optimization experiment, the basic unit is to understand the overall structure of the functional unit block. In the design, we need to identify the power hungry area and the timing critical path. Sometimes the timing critical path can also lead to increased power. Once the circuit is analyzed, we need to understand how to optimize power and timing, and how to use various methods to converge timing and power gain.
The work includes

1. Analysis and Optimization of power for the functional unit blocks
2. Analysis and Optimization of timing to improve frequency performance

To get high quality performance from the design, it is necessary to make sure that various optimization techniques used to achieve timing convergence and gain power, should not have any negative impact on the overall efficiency of the design.

IV. TIMING OPTIMIZATION TECHNIQUES

Timing Analysis is the process to check the speed requirements of the chip at the given frequency Kartschoke & Hojat (2001) in [6]. To guarantee the proper working of the circuitry, the largest path delay in the circuit must be less than the system clock cycle time. Each sequential has certain timing requirements like setup and hold that must be met in order to guarantee the correct operation of the circuit, and the goal of the timing analysis is to check whether they are met. This analysis is manual and which is time consuming stage in the design process. In figure 2.1 shows that when setup and hold violation occurs.

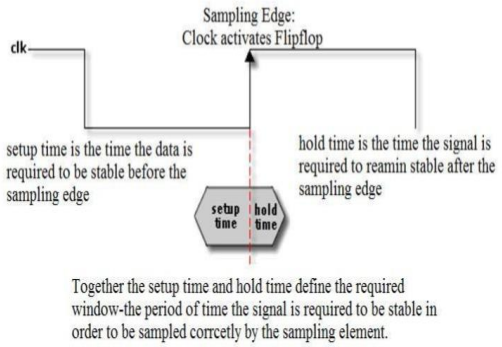

Figure 2.1 : Timing Checks

A. Sizing of Standard cell

Semi-custom design technology contains a standard cells based on requirements. It consist of standard cell library contains high quality standard cells which are accurately characterized and modelled for required functionality. Each cell is defined with worst case rise and fall delay, and proper selection cell size helps in timing convergence. As we know there are two types of Timing violation one is setup violation another one is Hold violation.

For Setup violation, cell has to be upsized then delay of the cell will become less and also slope will be improved. Hence the total path delay will be less.

For Hold Violation, Cell has to be downsized this increases delay of the cell and overall path delay. In standard cell library there are some cells which are High threshold voltage. These cells are slow due to the high Threshold voltage, which helps to fix hold violations.

B. Logic Optimization

There are some special cells characterized in the library to simplify the circuit design and to save the power and area. These type of cells are designed with more than one gate inside a cell. For example, a multiplexer designed with pass transistor logic will have poor driving strength and more prone to noise at input. So, generally inverters are added at output and input of multiplexer to address these issues. Hence, different multiplexer cells are designed by using only-pass gate logic, inverter-pass gate logic or inverter-pass gate inverter logic. Also, frequently used 2 to 3 level gate logics are merged into single cell, to save area. Example AND-OR logic, NAND or NOR Latch, AOI (AND-OR-inverter), OAI, etc.

Sometimes Logic optimisation will also help in Timing convergence. If a path consisting of many logic elements we can optimise it by bypassing any element based with the number of fan out it has. This eliminates a stage delay of that particular element in-turn reducing the path delay. So this helps in fixing setup violation.

V. POWER OPTIMIZATION TECHNIQUE

Power optimization is the most aspect of designing a processor as the lower technology nodes rely on power for their durability. As technology is entering into nanometer world, dimension of the device also shrinks, complexity level of the design increases and thereby increase in power dissipation. The system on chip integrates advanced IC design concepts. The SoC includes memory, cores, analog circuitry. Dynamic power consumption increases significantly with frequency. Power consumption can be classified into dynamic power consumption and static power consumption [13].

This work focuses on the reduction of active and dynamic power of the design. Resistance of the cells is a major contributor to the active power. The dynamic power is given by the following equation

$$P_{avg} = P_{Dynamic} + P_{Static} \quad (1)$$

The dominate part of P_{avg} is the dynamic power, $P_{Dynamic}$, caused by $P_{switching}$ and $P_{short-circuit}$. Here we emphasize on the minimization of $P_{switching}$ which can be expressed by the following equation:

$$P_{switching} = AF * C_{out} * V_{dd}^2 * f \quad (2)$$

Where AF is activity factor and Cout is an output capacitance, Product of AF and Cout is a dynamic Capacitance (C_{dyn}).

Equation 2 states that frequency and power are directly proportional. For any device to work at its best performance frequency plays very important role. So if in case increase in frequency for device to work at faster rate power consumption of that device is more. So balancing the frequency as well as power is a biggest task, hence we are considering other factors like AF and Cout to optimising the design for good performance with low power consumption. Based on AF and Cout the following techniques are used to improve performance at full chip level [14].

A. RTL Clock Gating

RTL clock gating is a power gain technique in Front-end. RTL clock gating can be applied by identifying the sets of flip-flops or latches sharing a common enable. This enable is used to gate the input clock in the design which is then connected to all the flip flops or latches in the design. Therefore, if a flip-flop/register bank which shares a common enable are implemented using RTL clock gating, then the flip-flops will ideally consume zero dynamic power as long as this enable value is zero. The enable gates the clock when it is false, which ultimately stops clock which is driving the FF/register when it is not needed. There are many ways to gate a clock [2], and one common way is the latch based clock-gating, pictured in Figure 6.1. Latch based clock gating is preferred because the High enable can come anytime in AND/OR based clock gating and it may not match with the High of clock, this may cause different duty cycle in output and also glitches. It's already known that how dangerous the glitches are in the clock signal. Thus, when enable is passed through a latch it is assured that the input of AND gate will toggle according the same clock duty cycle and thus there will be no variation in the duty cycle of output. Note that the same enable can be replicated to the similar group of registers/FF thus a single enable latch can be used to clock gate multiple registers/FF.

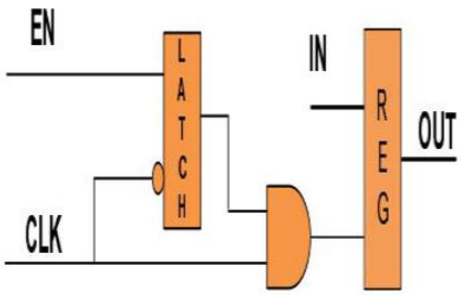

Figure 2.2: Latch based clock design

B. RCB/LCB merge

In all microprocessor designs, the clock topology is divided into three segments: global clock distribution, local clock distribution, and local clock distribution. Global clock distribution is responsible for clock synthesis by on-die PLL (Phase Locked Loop) and distribution of core clock. The global clock is also called as reference clock for all the sequentials in the microprocessor. This clock is basically present at the section level from where it is distributed to the FUB level. Each FUB has its own Regional clock distribution topology which further feeds the local clock distribution. Finally, the sequentials receives the clock from local clock distribution level in the FUB. There are multiple clock buffers used in these topologies. The main task of these topologies is to gate the clock at different levels and

reduce the AF according to the enables in the design. The main focus in this technique will be on minimizing the load capacitance on the high AF nets i.e. at the regional distribution level. The objective is to minimize the use of regional clock buffers in the design and then also provide the same quality of timing and gating for the next level. By this the load capacitance will be reduced on the High AF clock nets which will directly impact in the reduction of dynamic capacitance of the design.

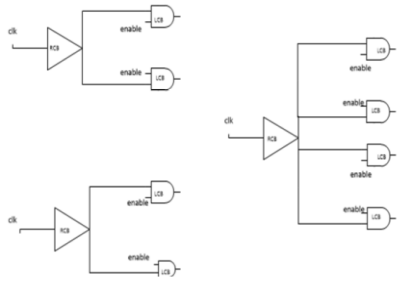

Figure 2.3: RCB merge

C. Clock Tree Synthesis

A clock tree is a distribution of clock network within a system or hardware design. Clock tree synthesis is a physical design implementation flow can help in avoiding serious issues like excessive power consumption, Routing congestion, It includes clocking components and devices from clock source to destination, Hence complexity depends on the number of clocking components used. We know that at lower metal layer resistance will be more hence clock is routed in a higher metal layers. There are several techniques used to achieve better performance. Clock should arrive exactly at same time to all the clocked elements. H-bridge CTS is one which is widely used optimization technique. Here Fish bone structure is used for clock tree synthesis which is shown in figure 3.1

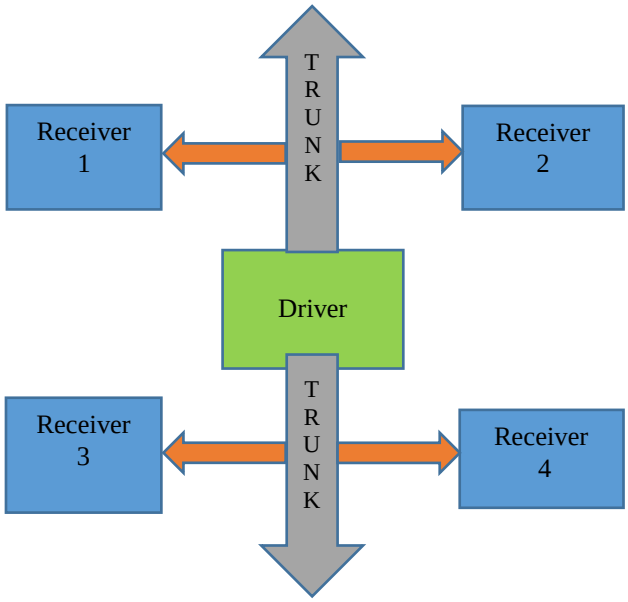

Figure 3.1: Fish bone structured Clock tree

The Fish bone structured clock tree routing is shown in figure 3.1 it is also called trunk and spine structure. Here the driver has 4 receivers and the driver is placed at the centre of all the 4 receivers, these receivers are equidistant from the driver. “Trunk” shown in figure 3.1 is a Routing layer which runs through all the four receivers. These 4 receivers are then connected to the trunk by lower metal layers called “Spine”. This configuration is used when driver has multiple receivers.

D. Dual/Quad insertion technique.

The Dual/Quad insertion technique also called multi-bit clustering is not only used to save power but also save area of the block. Insertion technique is a conversion of neighbouring single sequential cells into dual sequential or quad sequential. If more than one flops/latch are in the same neighbourhood, they can be combined or clustered to prevent giving different clock to each of them. This saves power as most power is consumed by clock. The insertion techniques are classified into two types, they are

- Paired Sequential insertion
- Unpaired Sequential insertion

In paired Insertion technique, when two latches shared by a same clock and are located in the same hierarchy then they can be merged. If two single latches are merged then this technique is called Dual paired insertion technique which is shown in figure 3.2 similarly if two dual latches are merged it is called as quad paired insertion technique.
In unpaired Insertion technique, when two latches shared by same clock and are located in different hierarchy then these latches are pulled out of the hierarchy to merge. If two single latches are merged then this technique is called Dual unpaired insertion technique which is shown in figure 3.3 similarly if two dual latches are merged it is called as quad unpaired insertion technique.

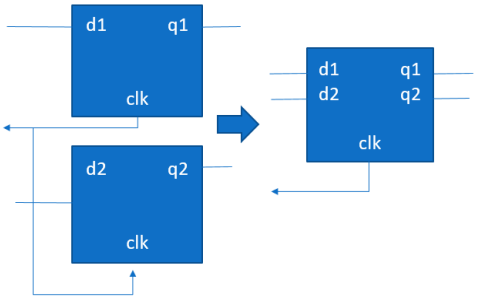

Figure 3.2: Paired Dual Latch

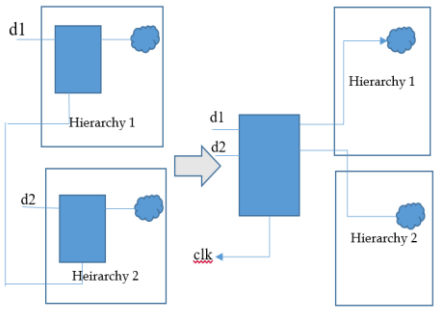

Figure 3.3: Unpaired Dual Latch

E. Clock path Optimization

This technique involves multiple techniques which collectively reduces the dynamic capacitance in the clock path by reducing the AF in the clock nets, by doing optimal placement to reduce the length of clock routes and by removing unwanted buffers or reducing logic in clock path. The main focus is to reduce the dynamic power by reducing the effective dynamic capacitance at the sampling clock nets which are accumulating high delays and also have good setup margins in those paths. Buffers can be removed directly from such paths which will directly reduce the dynamic capacitance almost equal to gate capacitance of the buffer removed. The other main point of focus to reduce the overall dynamic capacitance is reducing AF. The objective is to reduce AF of the clock as early as possible in the design and doesn't let the high AF clock propagate much deeper in the design. This will reduce the nets with high AF and also if low AF clock nets contains high load capacitance in the path, the overall dynamic capacitance contribution of that path will remain low only. Thus, reducing AF of clock nets is very much beneficial for overall dynamic capacitance reduction.

F. Clock pull/ Clock multiplication

Clock pulling is the process of decreasing clock delay. Clock pulling is generally known as clock tuning where a clock buffer is removed if it is redundant, and make sure that removal of clock buffer should not affect the rest of design. Reducing delay in the clock path fixes min violation.

Clock multiplication is another technique used to reduce power in most of design where two clock buffer are merged which means instead of 2 clock buffer we can replace it by 1 clock buffer with bigger size. Here CB1 is driving 2 clock buffers CB2 & CB3 respectively, This 2 clock buffers can be replaced by 1 clock buffer I.e., CBa, this CBa drives further receiver. This also fixes min violation.

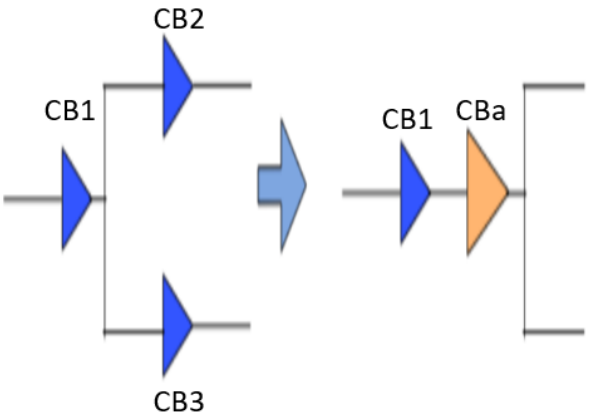

Figure 3.4: clock multiplication

VI. RESULTS

Timing analysis is done by ensuring that the design with good results are obtained for given frequency of operation. The techniques for making negative margin to positive margin are explained in this paper and the results are shown in table 1. Here setup margin and hold margin for all the paths are checked. Similarly for power optimization checked with the power hungry areas and applied techniques like clock tuning, CTS, Dual/quad insertion, clock path optimization to those power hungry areas. The previous generation module (functional unit block) is the starting point in our design which will be called "reference design" in the result section. Optimization is done for the current generation processor will be called "Current design" in the result section.

A. Timing Results

Analysed the Timing critical Paths, Static Timing Analysis is applied to those timing critical paths ensuring that the design will obtain good results.

The setup margin and hold margin with respect to reference design and current design are tabulated. WNS refers to the worst Negative slack which means worst negative margin in the design. Negative path refer to the number of paths which are having negative margin. Internal paths are those which are located within the functional unit block. External path is that the path is travelling from one functional block to other adjacent block (As already explained in section II smaller leaf cell is called FUB, FUBs are placed adjacent to each other form a section).

Distribution of paths for "Reference" and "current" design for setup margin and Hold margin is shown in table 1. The internal and external wns of MAX is -31ps & -42ps for reference design and is reduced to 0ps & -10ps respectively in current design. Similarly MIN is also reduced to -10ps and -16ps respectively.

Table 1: Setup and Hold margin in "current" and "reference" design.

			Previous Design	Current Design
MAX/setup	WNS in ps	Internal	-31	0
		external	-42	-10
	Number of negative paths	Internal	284	0
		external	721	58
MIN/hold	WNS in ps	Internal	-56	-10
		external	-68	-16
	Number of negative paths	Internal	863	72
		external	1023	200

The below graph represents the setup and hold margin with respect to the reference and current design.

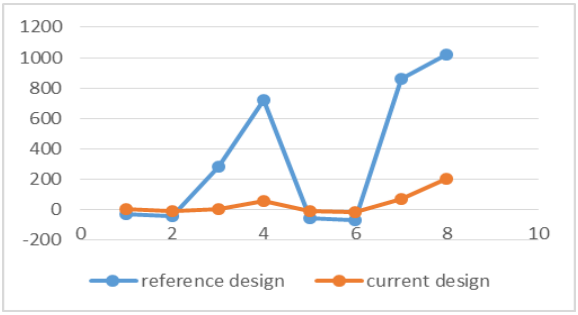

Figure 4.1: Timing convergence graph

B. CTS Result

The Fish-Bone Clock tree synthesis is shown in figure 3.1 whose results are tabulated in table.2. Due to CTS routing the negative margin has improved by 18ps, earlier the net was taking 27 as RC delay but after CTS the delay is reduced to 10.

Feature	**Margin in picoseconds**	**Delay of concerned Net**	**Device delay**
Normal routing	-20	27	102
CTS routing	-2	10	100

Table 2: clock tree synthesis

C. Multi-bit clustering Result

The dual/quad insertion or multi-bit clustering technique is shown in figure 3.2 and 3.3 respectively. The latches or flops can be merged together. The below table gives the information of multi-bit clustering done in the "current" Design.

Sequential Type	**Count**	**Average AF**
Quad Latch	80	13.72
Quad Flops	11	1.80
Dual Latch	121	20.75
Single Flop	13	2.22

Table 3: Summary of multi-bit clustering.

D. Power Results

The figure 4.1 shows usage of power in terms of clock power and data power, Analyzed the power by comparing reference and current design. In reference design, the total power is 18.968 mw and clock power contribution is 10.4%. Efforts have been made to reduce clock power.

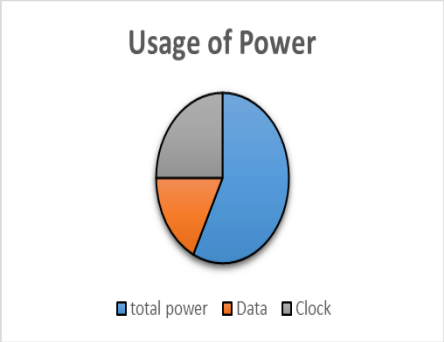

Figure 4.2: usage of power

Optimization techniques are applied to reduce power. Distribution of power in Current design is shown in Figure 4.1. The clock power has reduced from 10.4% to 7.2% of overall power through optimization.

The Dynamic power is responsible for total power of a full chip, it is caused due to switching activity of the cells. The results are shown in table 3.

Table 3: Comparison of power in "Reference" and "Current" designs

	Total dynamic power(mW)	**Clock power (mW)**
Before Optimization	18.968	10.46
After Optimization	15.758	7.251
Power saving	3.21	3.21

The Above table shows the details of reduction in power through optimization. Here optimization refers to the techniques which are explained in this paper along with the results. The clock power plays important role in reduction of dynamic power the Activity factor of clock is more hence the switching activity of clock is proportional to the dynamic power, hence the target is to reduce the dynamic power. Here the gain in dynamic power is 3.21mW.

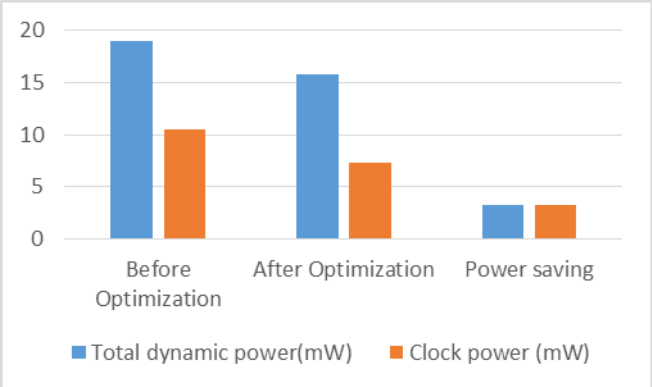

Figure 4.3: Utilization of power before and after optimization

VII. CONCLUSION

The digital circuits were designed at particular frequency, with each versions of design. Operating frequency changes with every generation of microprocessor. In this study, all the optimization techniques related to power and timing were discussed in detail and results were tabulated. Some new timing optimization techniques were introduced in this study which can be performed in VLSI circuits to give the required frequency push. The major target is to reduce dynamic power, this is caused by the simultaneous switching of clock because the Activity Factor of clock is high. Hence the techniques which are explained in this paper targets to reduce dynamic power. Experiments for power and timing optimization are performed successfully with significant gain in setup/hold margins and overall reduction in power.

Clock Tree Synthesis, RCB/LCB merge, Dual/Quad insertion, Logic Optimization all these techniques can be taken to higher level in future. Power reduction as well as timing convergence can be achieved by applying techniques explained in this paper, also can be used separately or together to get gain in power depending on the requirement.

VIII. FUTURE SCOPE

Each new project will be an improvement over its predecessor. This is achieved by adding newer RTL features and further technology scaling. This requires more efforts in terms of performance optimization and better quality checks. Other important factors are noise and its effect on the signals. Extensive work can be done in this field to achieve immunity from noise. Moreover, accounting for on-chip variation to get best picture of timing and making an OCV aware clock tree could be another research area that could be looked into. Timing of the critical paths, speed-path analysis and debug are other factors that could be a plan of future works.

REFERENCES

[1] Anshumali, K., Chappell, T., Gomes, W., Miller, J., Kurd, N. & Kumar, R. (2010), "Circuit and process innovations to enable high-performance, and power and area efficiency on the nehalem and westmere family of intel processors.", Intel TechnologyJournal.

[2] Intel encyclopedia, "http://www.intelpedia.intel.com" in *Intel Corporation Ltd*

[3] Bajkowski, M., Pham, G. & Vepadharmalingam, M. (2010), Optimum leakage dynamic array design, in 'VLSI Design Automation and Test (VLSI-DAT), 2010 International Symposium on', IEEE, pp. 262–265.

[4] Moore's Law and Intel Innovation

[5] Davit musayelyan, vazgen melikyan and meruzhan martirosyan in "Multi-voltage and Mulit-threshold low power design techniques for ORRCA processor on 14nm technology.

[6] Linumon Thomas, V Kiran. "Timing convergence techniques in digital VLSI designs" , 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017

[7] Kartschoke, P. & Hojat, S. (2001), 'Techniques that improved the timing convergence of the gekko powerpc microprocessor', pp. 65–70.

[8] S. K. Varadharajan and V. Nallasamy, "Low power VLSI circuits design strategies and methodologies: A literature review," 2017 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2017, pp. 245-251.

[9] Koichi nose and Takayasu sakurai, "Analysis and future trend of short circuit power". IEEE transaction on computer aided design of integrated circuits and systems, vol 19 No 9. september 2000

[10] Jitesh shinde. Dr.S.S Salankar Clock Gating –A Power Optimizing Technique for VLSI Circuits, Department of Electronics & Telecommunication Engineering J.L.Chaturvedi College of Engineering, Nagpur.

[11] Li-Chuan Weng, XiaoJun Wang and Bin Liu, "A survey of dynamic power optimization techniques," The 3rd IEEE International Workshop on System-on-Chip for Real-Time Applications, 2003. Proceedings. Calgary, Alberta, Canada, 2003, pp. 48-52

[12] Joe burkis "clock tree synthesis for high performance ASICs" IEEE conference on application specific integrated circuits division, 1991.

[13] Zhao, P., McNeely, J., Kuang, W., Wang, N. & Wang, Z., "Design of sequential elements for low power clocking system", IEEE Transactions on very large scale integration (VLSI) systems, 2011, pp. 914–918.

[14] Koichi Nose and Takayasu sakurai "Analysis and future trend of short circuit power" IEEE transactions on computer-aided design of integrated circuits and systems, vol. 19, no. 9, september 2000.