Design Methodology of Ultra Low-power MPEG4 Codec Core Exploiting Voltage Scaling Techniques

Kimiyoshi Usami, Mutsunori Igarashi, Takashi Ishikawa, Masahiro Kanazawa,
Masafumi Takahashi, Mototsugu Hamada, Hideho Arakida,
Toshihiro Terazawa and Tadahiro Kuroda

Toshiba Corporation
580-1, Horikawa-cho, Saiwai-ku, Kawasaki 210, JAPAN
Phone: +81-44-548-2346

kimiyoshi.usami@toshiba.co.jp

1. ABSTRACT

This paper describes a fully automated low-power design methodology in which three different voltage-scaling techniques are combined together. Supply voltage is scaled globally, selectively, and adaptively while keeping the performance. This methodology enabled us to design an MPEG4 codec core with 58% less power than the original in three week turn-around-time.

1.1 Keywords

Low power, voltage scaling, design automation, synthesis, placement, flip-flops, level converters, MPEG4, codec

2. INTRODUCTION

Mobile computing devices have been enjoying considerable success in the consumer electronics market. Market requirements are pushing the performance of those devices even higher, and the battery life even longer. In the near future, a wireless multi-media terminal on which we can enjoy moving pictures and stereo sound will emerge. MPEG4 is a standard on video and audio coding/decoding for mobile computing environment. Clearly, major requirements for MPEG4 codec LSI's are low power for longer battery life and high performance.

In CMOS LSI's, scaling supply voltage V_{DD} is very effective because power is reduced quadratically with V_{DD}. However, scaling the voltage poses an problem of degrading the performance. The following techniques have been reported which reduce power by voltage scaling without degrading the entire performance: The first is a technique in which supply voltage is globally scaled with scaling threshold voltage Vth, such as VTCMOS [2]. The second is a Dual-V_{DD} approach in which the reduced V_{DD} is selectively applied to non-critical paths [5][6]. The third is a variable supply voltage (VS) scheme in which V_{DD} is controlled adaptively using an on-chip DC-DC converter [3]. However, there have been few papers which present a methodology of making these techniques work together and show the level of power reduction we could reach.

In this paper, we propose a novel design methodology in which Dual-V_{DD}, VTCMOS and VS techniques are incorporated into a fully automated design flow. We discuss not only the total effectiveness in power reduction but also the contribution of each technique through an actual design of an MPEG4 codec core. Section 3 presents a background of the three techniques. Section 4 describes a design methodology we propose. Section 5 summarizes the feature of an MPEG4 codec core. Section 6 presents results from applying our design methodology to the core.

3. BACKGROUND

In this section, we present features of Dual-V_{DD}, VTCMOS and VS techniques in detail.

3.1 Dual-V_{DD} Approach

In the Dual-V_{DD} approach, a couple of different supply voltages are used. The reduced voltage (VDDL) is applied to the circuit on non-critical paths, while the original voltage (VDDH) is applied to the circuit on critical paths. So far, a Clustered Voltage Scaling (CVS) technique [5] and an Extended Clustered Voltage Scaling (ECVS) technique [6] have been reported as Dual-V_{DD} approaches. These techniques do not change the critical path delay, resulting in keeping the entire circuit performance. The Dual-V_{DD}

1-58113-049-x-98/0006/$3.50

circuit structure resulting from this technique is shown in Fig.1.

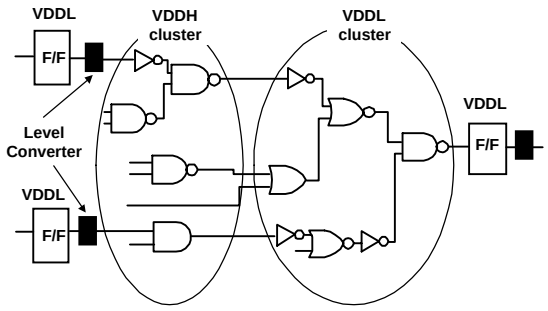


Figure 1 Dual-V_{DD} Structure

Level converters are inserted at the interface between the VDDL circuit and the VDDH circuit. The CVS technique can be further improved by optimizing the insertion points of level converters, leading to an ECVS technique. The ECVS technique allows us to increase the share of VDDL circuits, resulting in saving more power. An algorithm to quickly synthesize the clustered structure with dual supply voltages is presented in [6]. Layout architecture in the Dual-V_{DD} approach is depicted in Fig.2. VDDH cells and VDDL cells are placed in different rows in this architecture [6] .

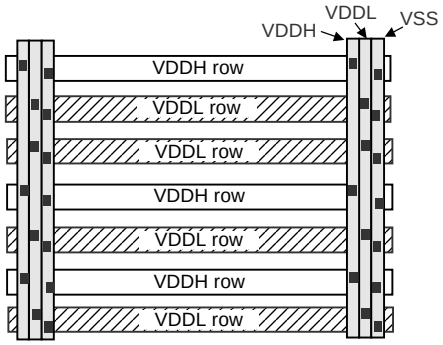


Figure 2 Row-by-Row Dual-V_{DD} Layout Architecture

The Dual-V_{DD} approach reduces power by utilizing excessive slack in a circuit. In actual designs, critical paths are only small portions of the circuit and excessive slack remains in the rest of the circuit. Designers have to struggle with critical paths so they meet the timing constraints. This work usually takes up their design time, resulting in excessive slack left untouched even if it remains. In the Dual-V_{DD} approach such excessive slack is automatically detected and utilized for power reduction.

3.2 VTCMOS

The VTCMOS technique [2] dynamically controls effective threshold voltage by applying substrate bias to MOS transistors. In the active mode, transistors are made to operate at low V_{DD} with low Vth. Meanwhile, in the standby mode, the effective threshold voltage is made to be larger by applying substrate bias to block the leakage current. Transistor performance in the active mode is kept as that in the conventional design by utilizing low V_{DD} and low Vth.

3.3 VS Scheme

In the VS scheme, the minimum supply voltage at which the speed requirements are just satisfied is generated on a chip. Major components in this scheme are an on-chip DC-DC converter and a replica circuit of the critical path [3]. Delay of the replica circuit is dynamically monitored and fed back to the DC-DC converter. The DC-DC converter is adaptively controlled such that it generate the supply voltage at which the delay of the replica circuit matches the required clock period.

Power savings in the VS scheme come from the following: An LSI chip is usually designed with a margin taking into account the voltage fluctuation of the external supply. In other words, a chip is designed so that it can operate even when the supply voltage goes down from V_{DD} to V_{DD} -V'. From the viewpoint of power consumption, however, the power is wasted when the chip operates at higher voltage than V_{DD} -V'. In the VS scheme, the chip is made to operate at the minimum supply voltage satisfying the speed requirements, resulting in saving power.

3.4 Potential Advantages by Combining Three Techniques

All of the three techniques described above reduce power by scaling the supply voltage. However, the targets they aim at are different. The VTCMOS and VS techniques aim at scaling supply voltage of critical paths, while they do not care excessive slack remaining in non-critical paths. On the other hand, the Dual-V_{DD} approach targets at reducing the voltage of non-critical paths, while it does not touch the critical paths. From these observations, it is anticipated that by combining the three techniques, power is reduced both at the critical paths and at the non-critical paths.

The combination of the three techniques has another big advantage. An advantage of implementing a DC-DC converter on a chip can be shared with the three techniques. In VTCMOS and the VS scheme, the minimum supply voltage satisfying the speed requirements is needed. In the Dual-V_{DD} approach, the reduced supply voltage VDDL is required. Those supply voltages are generated with DC-DC converters on a chip. This fact is a big advantage for the system design because a single external voltage source is just enough for the chip.

4. DESIGN METHODOLOGY

We present a new design flow in which Dual-V_{DD}, VTCMOS and VS techniques are combined together. Potential abilities of the three techniques are fully drawn out in the flow. We also present approaches for shortening turn-around-time (TAT) and those for reducing area overhead in the flow. Those approaches include novel circuit technology and CAD algorithm.

4.1 Design Flow

Figure 3 shows the design flow of our voltage scaling approaches. To avoid increasing design complexity for generating Dual-V_{DD} netlist, we separate the design flow into two major steps: One is an ordinary single-voltage design step and the other is a Dual-V_{DD} design step. At the first step, we perform logic synthesis using RT level description and timing constraints assuming a single supply voltage. Generated mapped netlist is sent to the second step. At the second step, we perform Dual-V_{DD} structure synthesis and layout using the same timing constraints as that of the first step.

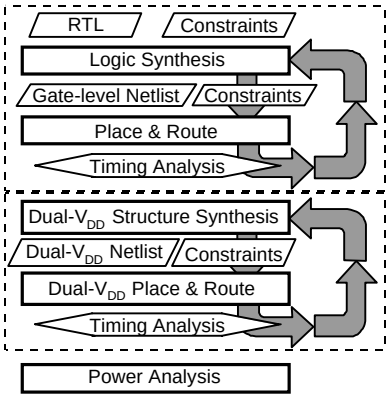


Figure 3 Design Flow

The VTCMOS technique allows the logic circuit in the core to operate at reduced supply voltage with the same performance as the original voltage. In our application of this paper, we determined the supply voltage of the circuit in the core as 2.5V while the original voltage as 3.3V. This is because the combination of 0.2V threshold voltage and the 2.5V supply voltage provides the same performance as that of the 0.55V threshold voltage and 3.3V supply voltage.

In the Dual-V_{DD} structure synthesis, we chose VDDL as 1.75V at which the power gets the minimum when VDDH is 2.5V. Layout results from the Dual-V_{DD} place and route and the layout data of the control circuits for the VS scheme are merged into the final layout data.

4.2 Approaches for Shortening TAT

In logic synthesis to generate single-voltage mapped netlist, we used a cell library which includes a flip-flop with level conversion function. We present why this flip-flop contributes to shortening the design TAT. Detailed features of the flip-flop will be described in the next section. Register elements in RT level description are mapped into those flip-flops even in single-voltage netlist in logic synthesis. The use of flip-flops with level conversion function is very helpful for minimizing the change between the netlist with single voltage and that with Dual-V_{DD}. This is because there is no need to insert level converters at the outputs of those flip-flops, resulting in minimizing the change in gate count and path delay. This allows us to utilize the placement result of the single V_{DD} layout as initial floorplanning information at Dual-V_{DD} layout, leading to shortening TAT.

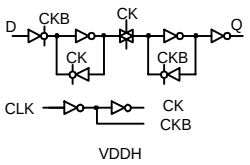
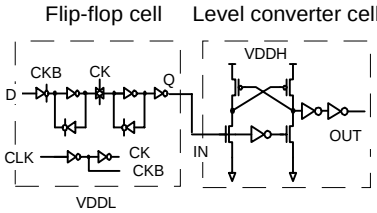
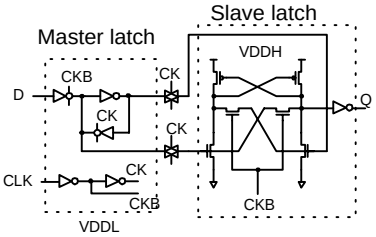
4.3 Approaches for Reducing Area Overhead

Area penalty in the Dual-V_{DD} approach is reported as 15% [6], while penalties in VTCMOS and a VS scheme are reported as less than one percent [2][3]. This motivated us to study and develop new techniques for reducing area overhead in the Dual-V_{DD} approach. Major components of the area overhead are as follows [1]: The first is an increase in cell area resulting from inserted level converters. The second is an overhead caused by layout constraints such that an N-well of VDDH cells should be isolated from that of VDDL cells. The third is an overhead resulting from additional power lines for VDDL. It is also reported that the first two components are dominant: the first component causes area penalty by 10% while the second does by 5%. We present an approach for reducing the area penalty of inserted level converters by using a novel flip-flop circuit. For reducing the overhead caused by layout constraints, we propose a new cell-placement algorithm.

4.3.1 Development of Flip-Flop with Level Converter

The CVS and ECVS structure described in Section 3 minimizes the number of level converters. The number is at most the number of flip-flops because a level converter is basically inserted at the output of a flip-flop. However, in the case of a circuit with a lot of flip-flops, inserted level converters cause a significant area penalty. We propose an approach which reduces the penalty by using a flip-flop circuit with a level-conversion function (FFLC). The circuit is depicted at the right-most column in Table 1. A master latch of the FFLC is the same as a conventional flip-flop, while a slave latch has a level-conversion function. The master latch operates at VDDL, so that it can receive voltage swing of VDDL at the data-input D and the clock-input CLK. The flip-flop circuit works as follows: When CLK is low, data is loaded into the master latch. When CLK goes high, the data in the master latch is transmitted to the slave latch through transmission gates. Level conversion is performed in the slave latch and the data is output to Q. Since the transmission gates are cut off between both latches when CLK is low, the gate of an NMOS transistor connected to the ground in the slave latch becomes floating. To avoid this problem, NMOS transmission gates controlled by CKB are provided. These transmission gates turn on when CLK is low, so that data is statically held in the slave latch.

We compared area penalties among three cases as shown in Table 1. A conventional flip-flop is chosen as the base of comparison. As shown in the second column, area increases by 64% in the case of inserting a level-converter cell at the output of a flip-flop. Meanwhile, area penalty of the FFLC is only 21%. Therefore, by using the FFLC we can reduce the area overhead in the CVS or ECVS structure.

	Conventional flip-flop	Insertion of level-converter cell at output of flip-flop	Flip-flop with level-conversion function (FFLC)
Circuit structure			
Area *	1	1.64	1.21
Power *	1	0.74	0.63
CLK-to-Q delay *	0.24 ns	0.59 ns **	0.42 ns

* Results on area and power are normalized by those on the conventional flip-flop. Power and delay were simulated at VDDH=2.5V, VDDL=1.75V, |Vth|=0.2V.
** CLK-to-OUT delay

Table 1 Comparisons on flip-flops

The FFLC has significant advantages also in power and performance. Data shown in Table 1 were obtained using a SPICE simulation at VDDH of 2.5V, VDDL of 1.75V, and threshold voltage of 0.2V. It is found that the FFLC reduces power to 63% compared to the conventional flip-flop. This results from the fact that the master latch in the FFLC operates at VDDL while the entire circuit of the conventional flip-flop operates at VDDH. CLK-to-Q delay of the FFLC is larger only by 0.18ns than that of the conventional flip-flop. Compared to the case of inserting a level-converter at the output of a flip-flop, the FFLC is superior also in power and performance. Thus, we have chosen to use the FFLC in the design.

4.3.2 Improvement in Cell Placement Algorithm

In the Dual-V_{DD} layout, we place cells considering timing constraints and routability as we do in an ordinary single-voltage design. In addition, the placement constraint is added to each cell corresponding to its supply voltage. In the layout architecture shown in Fig.2, voltage assignment to rows significantly affects the layout quality.

In this paper, we propose an improved voltage assignment technique. Basic framework is identical to that in [1]. First, cells are placed without considering the positional constraints on the supply voltage. VDDH cells and VDDL cells are mixed in any rows. Average sum of cell area within a row, A_{AVE} is computed at this stage. Next, the sum of the area of VDDL cells within the top row is computed. In [1], the authors determine that the top row should be a VDDL row if the sum reaches A_{AVE}. If not, they move to the second row and accumulate the area of VDDL cells. They keep scanning rows and accumulating the area of VDDL cells until the accumulated result reaches A_{AVE}. If it reaches A_{AVE} at the *m*-th row, they determine that the *m*-th row should be a VDDL row and the rows up to (*m*-1)th be VDDH ones. The same procedure is repeated again from the (*m*+1)th row. Thus the voltage of every row is determined. After this procedure, VDDL cells are moved to the VDDL rows because in the initial placement both VDDL and VDDH cells are mixed in any rows. VDDH cells are moved as well. This voltage assignment algorithm is simple and fast. However, area penalty was an issue. Our analysis showed the moving distances of the cells were fairly large. This indicates the final positions of the cells deviated from the optimum, resulting in area increase.

In an algorithm we propose, we assign the voltage to rows such that the moving distances of cells be minimized. The procedure of scanning rows and accumulating the area of VDDL cells is the same as that described above. The difference is that we compute the moving distances of the cells when the accumulated result of the area reaches A_{AVE} at the *m*-th row. In other words, we compute the sum of moving distances of cells assuming the *i*-th row is VDDL while changing *i* from 1 to *m*. Then we compare those sums and choose the voltage assignment with the minimum sum of moving distances.

5. OUTLINE OF MPEG4 CODEC CORE

We summarize features of an MPEG4 codec core to which we applied the design methodology described in Section 4. This MPEG4 video codec is the first design [4] which not only conforms to H.263 but implements essential functions in the MPEG4 Verification Model Version 7. The core consists of nine function blocks: a 16-bit RISC processor, DCT and IDCT blocks, a couple of motion estimation blocks, a motion compensation block, variable length encoding and decoding blocks, and a DMA controller. Most of them have local memories inside. Operation frequency is 30MHz. The core is fabricated in 0.3μm CMOS double-well triple-metal technology. Internal supply voltages of 2.5V and 1.75V are generated in a couple of on-chip DC-DC converters from an external 3.3V power supply.

6. RESULTS

We present results from applying our methodology to the core. Effectiveness is discussed from the viewpoint of power reduction, area overhead and design TAT.

6.1 Power Reduction by Dual-V_{DD} and VTCMOS Techniques

Power reductions in three designs are compared: 1) a conventional design, 2) a design to which only VTCMOS is applied, and 3) a design to which VTCMOS and Dual-V_{DD} techniques are both applied. Power dissipation was analyzed using a PowerMill simulation with real test vectors of the MPEG4 codec. Results are shown in Fig.4. VTCMOS using V_{DD}=2.5V and Vth=0.2V reduced power by 42% compared to the conventional design using V_{DD}=3.3V and Vth=0.55V. As shown in the figure, power is reduced in all of the components in the core such as logic gates, flip-flops, the clock tree and memories. This is because the supply voltage has been reduced from 3.3V to 2.5V while scaling Vth at every transistor in the core.

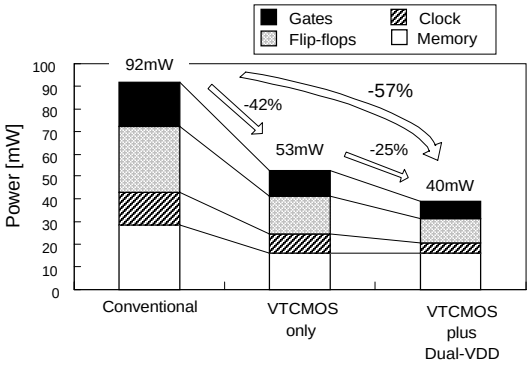


Figure 4 Power reduction resulting from VTCMOS and Dual-V_{DD} approach

Applying the Dual-V_{DD} approach in addition to VTCMOS reduces power further by 25%. The Dual-V_{DD} approach reduced power at logic gates by 30%, at flip-flops by 37%, and at the clock tree by 49%, while did not reduce power at memories. This is because we were obliged to use ASIC memory macros and those operating at VDDL were not supported. Therefore, we applied the Dual-V_{DD} approach to the portions except memories. At the portions to which we applied the Dual-V_{DD} approach, power was reduced by 35% even after applying VTCMOS.

Thus the power in the conventional design reduced by 57% while keeping the spec frequency 30MHz by combining both techniques together.

6.2 Effectiveness of Dual-V_{DD} Approach

We discuss effectiveness shown by the Dual-V_{DD} approach after applying VTCMOS in more detail. First we present power reduction at logic gates. Gates in non-critical paths are made to operate at VDDL in the Dual-V_{DD} approach, leading to reducing power. By applying the Dual-V_{DD} structure synthesis, 69% of the cells in the entire core have been made to operate at VDDL while meeting the timing constraints. In order to examine the share of non-critical paths in the total paths, we investigated 50 million paths in the entire core using a static path analyzer. Path-delay distribution is shown in Fig.5. The X-axis shows the path delay normalized by the cycle time, while the Y-axis shows the path count. In the design with only VTCMOS, the center of the distribution exists at the normalized path-delay of 0.5. By applying the Dual-V_{DD} technique to the design, the center of the distribution actually shifted toward the right, as shown in the figure. Thus, a lot of non-critical paths remain even in the design to which VTCMOS was applied. Excessive slack in those paths are effectively spent for power reduction in the Dual-V_{DD} approach.

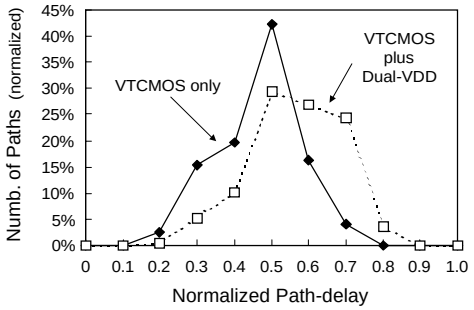


Figure 5 Path-delay distribution

Second, we present power reduction in flip-flops. In the design to which only VTCMOS is applied, conventional flip-flops are used. Meanwhile, in the design to which both the Dual-V_{DD} approach and VTCMOS are applied the FFLC's shown in Table 1 are used. The FFLC dissipates much less power than a conventional flip-flop, resulting in power reduction at flip-flops in the Dual-V_{DD} approach.

Third, we describe power reduction at the clock tree. The clock tree is made to operate at VDDH in the design with only VTCMOS, while the tree is made to operate at VDDL in the design with both the Dual-V_{DD} approach and VTCMOS. This leads to reducing power further at the clock tree in the design with the Dual-V_{DD} plus VTCMOS, compared to the design with only VTCMOS.

6.3 Power Overhead of DC-DC Converter

The VS scheme contains an on-chip DC-DC converter as a major component. The power consumption of a DC-DC converter was about 7.5mW. Two converters are needed to generate 2.5V and 1.75V from an external voltage of 3.3V, resulting in a power overhead of 15mW. The voltage scaling techniques presented in this paper can be applied to Intellectual Property (IP) modules such as an MPEG4 codec core, a PCI controller, and so on. When we apply the voltage scaling techniques to each IP module on a chip, the DC-DC converter is not needed for each module. Instead, the converter provided on a chip can be shared by all the IP modules. Compared to the total power savings resulting from the voltage scaling techniques with the DC-DC

converter, the power overhead described above is considered to be small enough.

6.4 Area Overhead

Figure 6 shows the photograph of the entire chip. Circuits for VTCMOS and VS schemes including DC-DC converters are placed at the corners of the chip. Since the corner of a chip is usually a dead space, area overhead of these circuits is almost negligible. The Dual-V_{DD} approach was applied to the random logic circuit located at the center of the chip. The area overhead was around 5%. This shows that our methodology has effectively reduced the area overhead in the Dual-V_{DD} design.

6.5 Design TAT (Turn-Around-Time)

The total design time from logic synthesis to the completion of the Dual-V_{DD} layout was only three weeks. Major factors that enabled us to shorten design time are as follows. One is to divide the design step into two sub-steps: the single-voltage design step and the Dual-V_{DD} design step. This approach was very effective to reduce the design complexity of logic synthesis in generating Dual-V_{DD} mapped netlist. Second is a flip-flop containing functionality of level conversion. The use of this flip-flop enabled us to save an additional floorplanning effort in the Dual-V_{DD} layout design, because the increase of the gate count or gate area became very small. Third is the improvement of the placement algorithm shortening the wire length in the Dual-V_{DD} layout. This resulted in eliminating redundant design loops between the layout step and Dual-V_{DD} structure synthesis step. Fourth is the fact that we extensively utilized not only standard cells but also memory macros in an ASIC cell-library.

6.6 Evaluation of Test Chip

We measured power dissipation of fabricated chips using a test vector with a function of MPEG4 video encoding. Results have shown that the core dissipates 45mW excluding the overhead of level converters. The power was 58% less than that of the conventional design.

7. FURURE WORK

By applying the methodology we have presented in this paper, power has been significantly reduced at logic gates, flip-flops and the clock tree. Consequently the power at memories has become to occupy large share. Power reduction techniques of memories will be required. Circuit technology of memories will be needed in which the performance is not degraded even at lower supply voltage. A structure synthesis technique such that memory macros in non-critical paths are made to operate at VDDL will be also required.

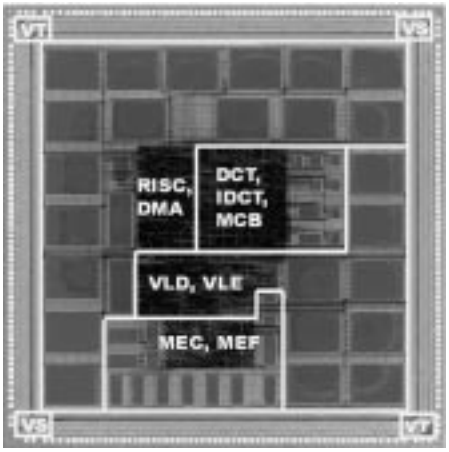


Figure 6 Chip micrograph of MPEG4 codec core

8. CONCLUSIONS

We have shown the potential advantages of our design methodology in which VTCMOS, a Dual-V_{DD} approach, and a VS scheme are combined together. These voltage scaling techniques work cooperatively at different design steps, and summation of each effectiveness becomes the total effectiveness. The VTCMOS and VS techniques reduce power at the entire circuit. The Dual-V_{DD} approach changes excessive slack in the circuit into power reduction.

We applied our design methodology to an MPEG4 codec core to examine the total power reduction and contribution of each technique. We have proved that our design methodology enables us to reduce power effectively with a small area overhead in a short design time.

9. REFERENCES

[1] Igarashi, M., et al. “A Low-power Design Method Using Multiple Supply Voltages”, *ACM/IEEE ISLPED-97*, pp.36-41, Aug. 1997.

[2] Kuroda, T., et al., “A High-Speed Low-Power 0.3um CMOS Gate-Array with Variable Threshold Voltage (VT) Scheme”, *IEEE CICC-96*, pp.53-56, May 1996.

[3] Suzuki, K., et al., ”A 300MIPS/W RISC Core Processor with Variable Supply-Voltage Scheme in Variable Threshold-Voltage CMOS”, *IEEE CICC-97*, pp.587-590, May 1997.

[4] Takahashi, M., et al, “A 60mW MPEG4 Video Codec Using Clustered Voltage Scaling with Variable Supply-Voltage Scheme”, *ISSCC Digest of Technical Papers*, pp.36-37, Feb. 1998.

[5] Usami, K., et al., ”Clustered Voltage Scaling Technique for Low-Power Design”, *ACM/IEEE ISLPD-95*, pp.3-8, April 1995.

[6] Usami, K., et al., “Automated Low-power Technique Exploiting Multiple Supply Voltages Applied to a Media Processor”, *IEEE CICC-97*, pp.131-134, May 1997.