BODY BIAS CONTROL FOR A COARSE GRAINED RECONFIGURABLE ACCELERATOR IMPLEMENTED WITH SILICON ON THIN BOX TECHNOLOGY

Honlian Su, Yu Fujita, Hideharu Amano

Dept. of ICS, Keio University, Yokohama Japan
email: leap@am.ics.keio.ac.jp

ABSTRACT

For low power yet high performance processing in battery driven devices, a coarse grained reconfigurable accelerator called Cool Mega Array (CMA)-SOTB is implemented by using Silicon on Thin BOX (SOTB), a new process technology developed by the Low-power Electronics Association & Project (LEAP). A real chip using a 65nm experimental process achieved a sustained performance of 192MOPS with a power supply of 0.4V and power consumption of 1.7mW. A clock frequency of 89MHz was achieved with a power supply of just 0.4V when a forward bias voltage was given. When using a reverse bias, the leakage current could be suppressed to less than 20μW in the stand-by mode. The key concept of CMA-SOTB is maintaining a balance between performance and leakage current by independently controlling the bias voltages of the PE array and the microcontroller. Evaluations of the operational frequency and power consumption of filter application programs shed light on how to find the combination of bias voltages that achieves the best energy efficiency for a required performance. The range of advantageous power supply voltage for a required performance considering the body bias was also found.

1. INTRODUCTION

One of the key components of providing the highly energy efficient operations required by sophisticated mobile devices is an accelerator. A coarse-grained reconfigurable processing array is a promising candidate for such accelerators since it can work with a low frequency clock without degrading performance by using highly parallel processing on a large processing element (PE) array. However, such architecture is prone to large power leakage, especially when extremely low power computation is required with a low voltage supply.

Silicon On Insulator (SOI) technology, in which transistors are formed on top of an insulator, has been developed for low voltage supply operation. It can work with a much lower supply voltage than that for bulk CMOS transistors. Also, the leakage current and delay of the transistors can be controlled by using the body bias. The Low-power Electronics Association & Project (LEAP)[1], a Japanese national project, recently developed a novel SOI CMOS technology called silicon on thin BOX (SOTB) that enables an extensive wide range of body bias control with extremely low supply voltage.

In the present study, we used the SOTB technology to develop a coarse-grained reconfigurable accelerator called Cool Mega Array (CMA)[2]. CMA provides a large processing element (PE) array consisting of combinatorial logic and a small microcontroller that manages data distribution and collection between data memory and the input/output of the PE array.

The contributions of this paper are threefold. (1) A novel SOTB technology is applied to a reconfigurable accelerator, and real chip evaluation results are reported. Aside from a simple microcontroller for automobile control[3], this is the first report in which practical large digital circuits have been applied. (2) A sustained performance of more than 192MOPS was achieved with a supply voltage of just 0.4V and power consumption of 1.7mW. In the stand-by mode, the leakage power can be suppressed to less than 20μW. (3) Bias control methods to achieve the best energy efficiency are shown. These methods find the best set of bias voltages for computational units and the memory access controller considering performance and power consumption. Although our previous trial[4] also controls the body bias to balance the performance and leakage current, only a limited effect was achieved as the target architecture used conventional bulk CMOS.

2. SOTB TECHNOLOGY

In this section, we give an overview of the silicon on thin BOX (SOTB) technology[5] developed by LEAP.

2.1. SOTB CMOSFET

The Silicon on Insulator (SOI) CMOS technology has attracted attention from the viewpoint of balancing the per-

This work was done in "Ultra-Low Voltage Device Project" of LEAP funded and supported by METI and NEDO.

formance and leakage power with the low voltage supply. Unlike in conventional bulk CMOS, in SOI, transistors are formed on top of the insulator(typically SiO_2). Surrounding the transistor with insulating material means that the electrical interference does not need to be considered, and the electric characteristics therefore become sharp.[5] There are two types of SOI: fully depleted (FD)-SOI and partially depleted (PD)-SOI. Although the process of fabricating FD-SOI is relatively difficult, the benefits of SOI become fully available. The SOTB utilized in this study (Figure 1) is classified into FD-SOI, but the transistors are formed on thin BOX (Buried Oxide) layer.

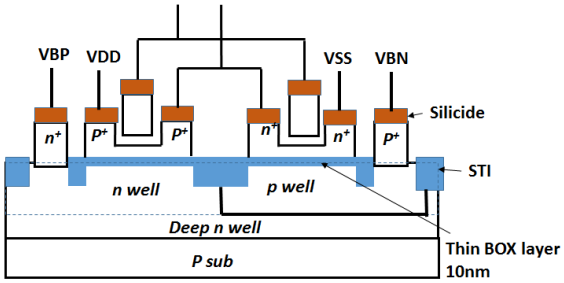


Fig. 1. Cross-sectional view of the SOTB Device

By using an ultra-thin FD-SOI layer and the BOX layer, we can suppress the detrimental short channel effect (SCE) in the SOTB. Since impurity doping (halo implant) to the channel is not necessary, the variation of the threshold voltage by the RDF can be reduced. Multi-threshold voltage design is easily available by doping an impurity into the substrate directly under the thin BOX layer. Thus, we can extensively control the range of body (back-gate) bias and optimize the performance and power consumption after fabrication.

One of the drawbacks of the FD-SOI structure is that the electrostatic discharge sensitivity is poor resulting in a small breakdown voltage. This can be solved easily by using bulk technology in the same wafer. The manufacturing process can easily incorporate bulk CMOSFET by removing the thin SOI layer and the BOX layer. Also the transistor and the independence of the back gate are guaranteed by the shallow trench isolation (STI). Although STI is quite similar to the technique used in the bulk device, it is sufficient to use for etching the three layers of the SOI, BOX and substrate. The well area under the BOX layer behaves as a ground plane and a back gate, and is connected to the back-gate contact. We use a triple-well structure to avoid leakage when applying the back-gate bias.

The characteristics of SOTB are summarized as follows: (1) The junction capacitance of the SOI is about 1/10 that of the bulk, thus making high-speed operation possible. The capacitance of the circuit becomes large as the junction capacitance increases and the operation speed becomes slow. The benefits of the SOI become more pronounced the lower the voltage operation. (2) One problem with CMOS devices is the possibility of latch-up which can destroy the device if it occurs. The latch-up is caused by a parasitic thyristor formed by adjacent transistors in bulk CMOS. However, these are not formed in SOI. (3) Anti-radiation tolerance is high. The part of the substrate of that generates charge by incident radiation is blocked by the insulation layer, and does not affect the operation of the circuit. (4) Noise propagation (cross-talk) is small because of the insulation.

3. CMA ARCHITECTURE

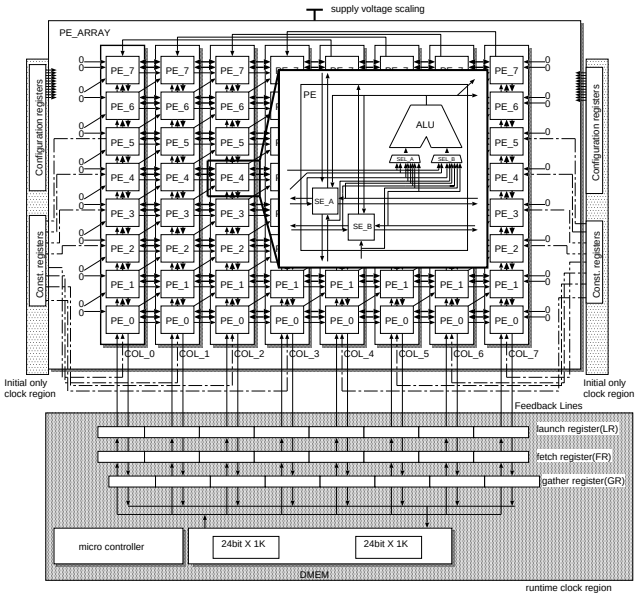
3.1. The concept of CMA architecture

The primary objective of CMA design is to design an architecture for executing a fixed number of computations in the required time with the minimum amount of energy. A key concept of CMA architecture is reducing any energy usage other than that required for computation. The PE array is built with combinatorial circuits to eliminate the power needed to store the intermediate results in registers and to clock distribution to each PE. The dataflow of the application is statically mapped on the PE array. Registers are only provided at the inputs/outputs of the PE array, and computation starts when all data are set up in the input register and the outputs of PE array are stored into the output registers with a certain delay time.

A microcontroller reads the data from the data memory (MEM) and distributes to the register attached to the input of the PE array. It also collects the results from the register attached to the output of the PE array, and writes them back it to the data memory. It flexibly manages the data transfer between the memory and registers by using mapping registers and vector operations. The above structure enables the implementation of various application programs without power hungry dynamic reconfiguration in the PE array.

Since the computation in the PE array and the data management by the microcontroller are performed in a pipelined manner, their execution speeds must be balanced. The original concept of CMA uses voltage scale control to keep this balance. If the computation delay is shorter than the data management delay, the voltage supplied to the PE array can be reduced. The total power required for computation can thus be reduced without degrading computing performance. On the other hand, if the data management delay is shorter than the computation delay, wave pipelining in the PE array can also be used[2]. The delay time for achieving wave pipelining can be also controlled by changing the voltage supplied to the PE array.

Figure 2 shows the architecture of our first prototype,



which we call CMA-1. The PE array network of the CMA-1 is a combination of two-channel island-style interconnection and direct links that connect to the north-east and east of the PE. Switching elements transfer the input data from the PE in the south, west and east of the PE and the output data of the ALU to the PE in the appropriate direction according to the configuration data, which is given from the outside, for switches.

3.2. Related work

The original intention of CMA is to balance the performance of for data distribution/correction by the microcontroller and computation in the PE array by using the supply voltage control. However, for lower power control, the supply voltage must be lowered close to the threshold level. In such a situation, the performance drastically decreases down with just a small reduction of supply voltage. This makes precise performance control quite difficult.

Although balancing the performance of the computational module and memory access is a common challenge in accelerators, methods other than voltage scale control have not been applied except in our previous trial[4]. Balancing leakage power and performance, on the other hand, has been well researched. [6][7][8]. Power-gating is one candidate to suppress the leakage power, and a few coarse-grain[9] or fine-grain reconfigurable architectures[10] have tried it. However, since power gating lowers the range of signal voltage, it is difficult to be used with supply voltage scaling.

FLEX Power FPGA[11] controls the back-gate bias by the configuration data in order to optimize the delay and leakage power. By using reverse bias only for the configurable blocks that are not on the critical paths of the design, a large amount of the power leakage is reduced without degrading performance. However, there are considerable overhead to control back-gate bias for fine grained reconfigurable elements.

Table 1. Specification of CMA-SOTB

Chip	Process	LEAP 65nm SOTB 7-metal
	Size	5mm $\times$ 5mm
	I/O	208pins
Tools	Design	Verilog HDL
	Synthesis	Synopsys Design Compiler 2011.09-SP2
	P&R	Synopsys IC Compiler 2010.12-SP5

4. CMA-SOTB

4.1. The Back-gate Bias

As shown in Figure 1, an SOTB transistor has a back-gate bias contact provided to its well. For an NMOS transistor VBN is given to its p-well. Here, zero bias (VBN=0) means the transistor works with its normal threshold. When reverse bias (VBN of negative value) is given, the threshold is increased and so the leakage current is reduced but the delay is stretched. In contrast, with the forward bias (VBN of positive value), the threshold is decreased and so the leakage current is increased but the operational speed is enhanced.

For a PMOS transistor, VBP is given to its n-well and zero bias corresponds to the supply voltage, that is, $VBP = VDD$. Reverse bias v means VBP of larger than VDD, and forward bias represents the case when VBP of smaller than VDD is given.

4.2. Design of CMA-SOTB

The architecture of CMA-SOTB is almost the same as that of CMA-1. Specifically, an 8×8 processing array with almost exactly the same interconnection network and PE structure is adopted. Since the memory macro with SOTB is under development, we used registers as the data memory. For this reason, memory size is limited to 256 words. Although only small application programs can be ported, it is sufficient to evaluate basic performance and energy consumption.

The specification of CMA-SOTB and the tools used for the design are listed in Table 1.

A photograph of CMA-SOTB is shown in Figure3. There are two separate cores: the microcontroller on the left and the PE array on the right. The area of the PE array is larger than that of the microcontroller but the difference is not large as expected. It is mainly due to the DMEM implemented by

registers and layout-ed with other logics enlarges the area for the microcontroller. Each core is controlled by an independent bias voltage. Here, bias voltages for the PE array are $VPNC$ and $VPBC$, and those for the microcontroller are $VPNM$ and $VPPM$.

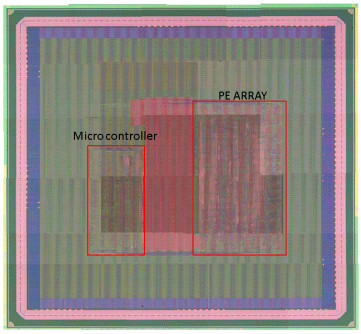


Fig. 2. Layout of CMA-SOTB

Note that both parts share the same VDD that is, we cannot separate the power consumed in the PE array and the microcontroller.

5. REAL CHIP EVALUATION

5.1. Leakage Current vs. Body bias voltage

The relationship between the leakage power and body bias voltage is shown in Figure 4. Here, the same bias voltage is given to both p-mos and n-mos, that is, $VBPC = VDD - VBNC$ and $VBPM = VDD - VBNM$. Note that VBN or VBP take negative values when reverse bias is given. Although the power for the two parts cannot be completely separated, the leakage power for one part can be measured by giving a strong enough reverse bias to suppress the power consumption of the other part.

Figure 4 shows that the leakage power is exponential to the bias voltage $VBNC$ and $VBNM$. Since the PE array has a larger area than the microcontroller, $VBNC$ is larger than $VBNM$. However, the slope of them is almost the same. Cases with three supply voltages ($VDD = 0.3, 0.4$, and $0.5V$) are shown in the graph. The leakage power with the strong reverse bias becomes slightly large when VDD of 0.3V is given. However, since the slope of larger VDD is steep, the order of leakage power is the same as the order of supply voltage with VBN of larger than -0.2V.

When $VBNC = VBNM = -1.3V$, the leakage current is 48μA, 59μA, and 67μA with VDD = 0.3V, 0.4V and 0.5V, respectively. This means that the stand-by power is extremely low: 14.4μW, 23.6μW and 33.5μW respectively. In contrast, when $VBNC$ or $VBNM$ is more than 0V, the total leakage current is larger than 5mA. As discussed later, it occupies more than half of the total operational power in most application programs.

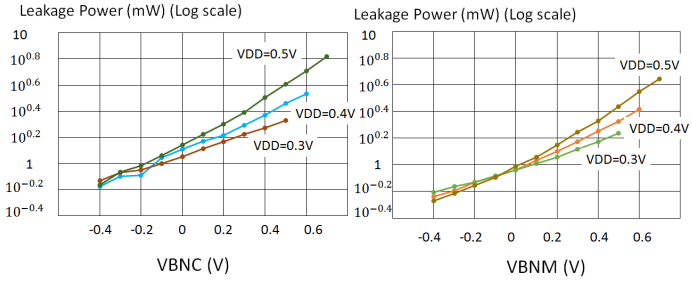


Fig. 3. Leakage Power vs. Bias Voltage

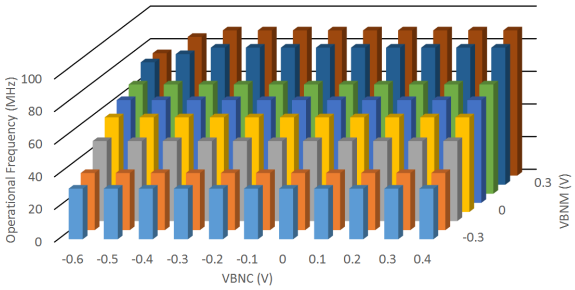


Fig. 4. Operational Freq. vs. Bias Voltage (*alpha*)

5.2. Maximum operational frequency

The maximum operational frequency of CMA is restricted either by the computation time in the PE array or the data management performance with the microcontroller. Both are influenced by the arithmetic intensity of the application programs. If the program requires a small amount of input data but has a large computational dataflow graph, the maximum operational frequency is bottle-necked with the delay time in the PE array. Otherwise, the time for reading input data/writing results from/to DMEM will limit the performance. Several application programs have been implemented and evaluated on CMA-SOTB. Here, because of the page limitation, we focus on the results of two image processing applications, *alpha* and *af*. *alpha*, 8-bit alpha bender of two input images is a typical light-weight program using just 16 PEs. The data management by the microcontroller bottlenecks the total job unless strong reverse bias is given to the PE array. In contrast, *af* is an alpha blender for 24bit RGB data. Since three pixcels are packed into a 24bit input data, the total job becomes computation centric. 48PEs are utilized in the data path of *af*.

In CMA-SOTB, the computation time of the PE array and the memory management performance of the microcontroller can be controlled by bias voltage $VBNC$ and $VBNM$. If the application is memory management dominant, the operational frequency is improved only by increas-

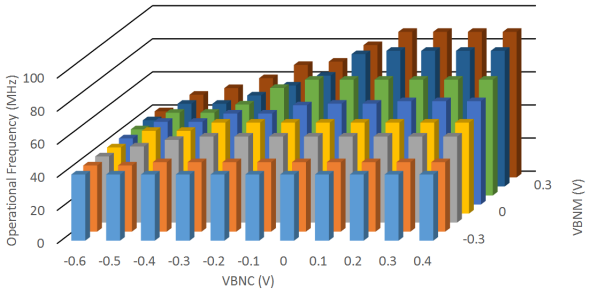


Fig. 5. Operational Freq. vs. Bias Voltage (*af*)

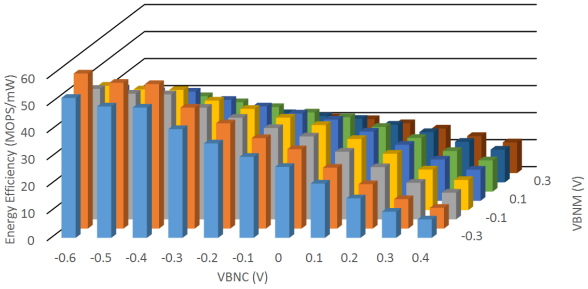


Fig. 6. Energy Efficiency vs. Bias Voltage (*alpha*)

ing $VBNM$. In this case, a large $VBNC$ is just a waste of the leakage power. In the opposite case, the operational frequency can be controlled by $VBNC$ not $VBNM$.

Figure 5 shows the operational frequency of *alpha*, a memory management dominant application when $VDD = 0.4V$. X-axis and y-axis are $VDDC$ and $VDDM$, respectively. A relatively high operational frequency of 89MHz is achieved with a low supply voltage of 0.4V, by making the best use of SOTB technology. Since the dataflow graph of *alpha* is light-weight, the operational frequency is mostly controlled by $VBNM$ from 31 to 89MHz in this range. When $VBNM$ is more than 0.3V, the operational frequency is degraded with a strong reverse bias for the PE array ($VBNC = -0.5V$).

Figure 6 shows the case of *af*, which uses 48PEs for computation. The range of operational frequency is almost the same as the case of *alpha*; from 40 to 88MHz. In this application, both $VBNC$ and $VBNM$ influence the operational frequency. When $VBNM$ is less than -0.1V, the operational frequency is mainly fixed by $VBNM$ as in the case of *alpha*. However, in the other area, if $VBNC$ is less than -0.1V, the operational frequency is also controlled by $VBNC$. That is, the delay of computation in the PE bottlenecks the system in this area. When the operation frequency is limited by the body bias for one part, increasing the bias for the other part just wastes the leakage power. Here, the minimum $VBNC$ or $VBNM$ with maximum operational frequency can be defined. For example, a maximum operational frequency of 50MHz is achieved when $VBNC = -0.5V$ and $VBNM > 0V$. We refer to the above case as $VBNM_{min}(VBNC = -0.5V) = 0V$ In contrast, the $VBNC_{min}(VBNM = 0.4V) = 0.1V$. This minimum bias voltage is important because the maximum energy efficiency can be achieved for the maximum operational frequency that can be achieved with the body bias voltage.

5.3. Energy efficiency

Although the operational frequency is improved by increasing $VBNC$ and $VBNM$, the exponential increasing leakage power will dominate the total power and severely degrade the energy efficiency.

Figs. 7 and 8 show the energy efficiency (MOPS/mW) when *alpha* and *af* are executed with various $VBNM$ and $VBNC$. As in Figs. 5 and 6, the x-axis and y-axis are $VDDC$ and $VDDM$, respectively. Million Operations Per Second (MOPS) is dependent on each application program and proportional to the operating frequency. Application programs with high arithmetic intensity (e.g. , *af*) achieve a large MOPS, while only a small MOPS is achieved by *alpha* using a small number of PEs. Since dynamic power is also proportional to the operational frequency, energy efficiency is improved with lower $VDDC$ and $VDDM$ by reducing the leakage power. In both Figs. 7 and 8, the maximum energy efficiency (57MOPS/mW in *alpha* and 192MOPS/mW in *af*) was achieved with the smallest $VDDC$ and $VDDM$.

This demonstrates that the best energy efficiency for CMA-SOTB can be achieved by decreasing $VBNC$, and $VBNM$ as much as possible with extremely low operational clock frequency. However, this setting is not always practical because a certain performance is required for each application program. $VBNC$ and $VBNM$ must be selected from minimum values so that the required performance can be achieved. In the case of CMA-SOTB, if the same required operational clock is to be achieved with different combinations of $VBNC$ and $VBNM$, they must be selected in accordance with the following guidelines. (1) If the acheived clock frequency is the same, select the lower bias voltage. (2) Avoid bias voltages of more than 0.1V if possible. Because of the exponential growth of the leakage current, selecting this range will degrade energy efficiency. (3) If multiple candidates satisfy the above conditions, use the lowest $VBNC$. This is because the leakage current of the PE array is larger than that of the microcontroller in CMA-SOTB.

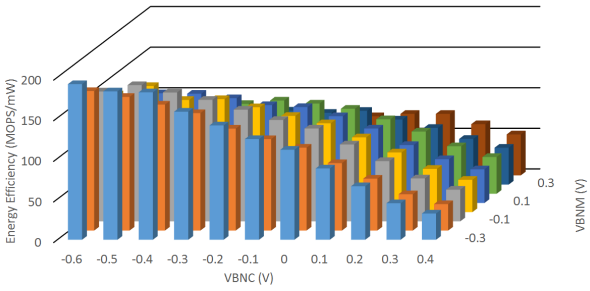


Fig. 7. Energy Efficiency. vs. Bias Voltage (*af*)

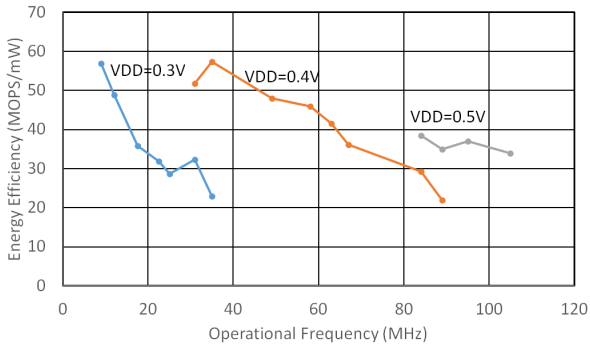


Fig. 8. Energy Efficiency. vs. Operational Frequency

5.4. Supply voltage vs. body bias voltage

In order to accelerate the performance, increasing the supply voltage is sometimes a more practical solution than increasing the body bias voltage. Increasing the supply voltage increases both dynamic and leakage power with squire order, while body bias affects only leakage power and with exponential order.

Figure 9 shows the energy efficiency vs. operational frequency when *alpha* is executed with a supply voltage of 0.3, 0.4 and 0.5V. $VDDC_{min}$ is utilized, since *alpha* is a memory management dominant application. The optimal supply voltage range is fixed according to the required performance. For example, if the target operational frequency is more than 80MHz, we should use Vdd of 0.5V rather than 0.4V. Even if the performance can be achieved with a lower supply voltage by increasing the body bias voltage, the energy efficiency would be significantly degraded, so the supply voltage must be fixed according to the required performance first and the body bias voltage then selected so as to maximize the energy efficiency.

6. CONCLUSION

We implemented a coarse grained reconfigurable accelerator CMA-SOTB by using SOTB technology developed by LEAP. A real chip using a 65nm experimental process achieved a sustained performance of 192MOPS with a power supply of 0.4V and power consumption of 1.7mW. A clock frequency of 89MHz was achieved with a power supply of just 0.4V when the forward bias voltage is given. By using a reverse bias, the leakage current could be suppressed to less than 20μW in the stand-by mode.

Further evaluation and comparison with previous chips, CMA-1 and CMA-bb are our future work. Also, dynamic control of body bias is another attractive research subject.

7. REFERENCES

[1] Low Power Electronics Association & Project, "*http://www.leap.or.jp/*."

[2] N.Ozaki et al., "Cool Mega Arrays: Ultra-low-Power Reconfigurable Accelerator Chips ," *IEEE Micro, vol.31,No.6*, pp. 6–11, 2011.

[3] K.Ishibashi, et al., "Soft-Error-Immune 0.22V-Vmin 13.6pJ/cycle 32bit CPU with Back-Bias dependent Logic and Memory in 65nm Hybrid SOTB Technology," *Proc. of CoolChips XVII*, April, 2014.

[4] H. Su, W. Wang, K. Kitamori, and H. Amano, "A Low power Reconfigurable Accelerator using a Back-gate Bias Control Technique," *Proc. of ICFPT*, 2014.

[5] Takashi Ishigaki, et al., "Ultralow-power LSI Technology with Silicon on Thin Buried Oxide (SOTB) CMOSFET," *Solid State Circuits Technologies, Jacobus W. Swart (Ed.), ISBN: 978-953-307-045-2, InTech*, pp. 146–156, 2010.

[6] F. Li, Y.Lin, L.He, and J.Cong, "Low-Power FPGA Using Pre-defined Dual-Vdd/Dual-Vt Fabrics," in *Proc. of the 2004 ACM/SIGDA 12th FPGA*, Feb. 2004, pp. 42–50.

[7] C.Q.Tran, H.Kawaguchi, and T.Sakurai, "Low-Power Low-Leakage FPGA Design using Zigzag power-gating, dual-VTH/VDD and micro-VDD-hopping," in *IEICE Trans. Electron, Vol.E89-C, No.3*, March 2006, pp. 280–286.

[8] K.Hirai, et al., "Leakage Power Reduction for Coarse-Grained Dynamically Reconfigurable Processing Arrays Using Dual VT Cells," *Proc. of ICFPT*, Dec. 2009.

[9] Y.Saito, et. al., "Leakage Power Reduction for Coarse Grain Dynamically Reconfigura ble Processor Arrays With Fine-Grained Power Gating Technique," in *Proc. of the ICFPT*, Dec. 2008, pp. 329–332.

[10] Assem A.M. Bsoul and Steven J.E.Wilton, "An FPGA with Power-Gated Switch Blocks," in *Proc. of ICFPT*, 2012, pp. 87–94.

[11] M.Hioki, T.Sekigawa, T.Nakagawa, H.Koike, Y.Matsumoto, T.Kawanami, T.Tsutsumi, "Fully-Functional FPGA Prototyple with Fine-Grain Programmable Body Biasing," *Proc. of 21st ACM/SIGDA International Symposium on FPGA* , Feb. 2013.