An Average-Performance-Oriented Subthreshold Processor Self-Timed by Memory Read Completion

Hiroshi Fuketa, *Member, IEEE*, Dan Kuroda, *Student Member, IEEE*, Masanori Hashimoto, *Member, IEEE*, and Takao Onoye, *Senior Member, IEEE*

***Abstract*—A self-timed subthreshold processor was developed in 65-nm complimentary metal–oxide–semiconductor process. This four-stage reduced instruction set computer processor synchronously operates with the memory read completion signal produced in 8.5-kb instruction and 2-kb data memories of subthreshold 10T static random-access memory. Measurement results show that the processor correctly functions from 0.56 to 0.36 V with a self-timed clock and achieves minimum energy per cycle of 3.47 pJ/cycle at 0.46-V supply voltage with 1.76-MHz average frequency. Compared with conventional synchronous operation with guardbanding, the proposed self-timed operation reduces the execution time of SHA-1 by 82% at 0.4-V supply voltage and saves energy by 40% to attain 1-MHz operation.**

***Index Terms*—Low power VLSI, self-timed processor, subthreshold circuit.**

I. Introduction

SUBTHRESHOLD circuits are promising for energy-constrained applications, such as processors for sensor networks [1], [2]. Since a sensor-node processor architecture is quite simple due to the limited functions required for sensing applications [1], access time to memory is a key factor that determines the clock cycle of the processor. For subthreshold processors, a memory cell (MC) with additional transistors dedicated to a read port, such as an 8T MC [3], has conventionally been used. Fig. 1 shows a preliminary evaluation of the fluctuation in read time of 8T static random-access memory (SRAM) due to within-die threshold voltage $V_{\rm th}$ variation. Since the ON currents of two transistors for the read operation have a direct impact on the read time, the read time is sensitive to within-die $V_{\rm th}$ variation as the supply voltage is lowered. Furthermore, the read time has a strong dependence on die-to-die $V_{\rm th}$ variation and environmental variability, such as supply voltage and temperature fluctuation. Therefore, the conventional synchronous operation considering the worst-case conditions of such variations is extremely pessimistic and inefficient due to the large delay margin for guardbanding.

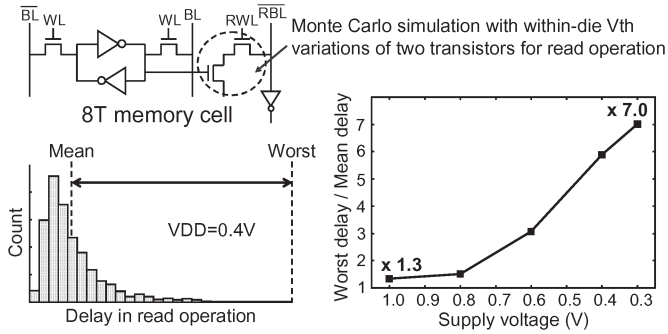

Fig. 1. Preliminary evaluation of the variation in read time of 8T SRAM due to within-die $V_{\rm th}$ variation in a 65-nm CMOS process. The lower left graph indicates the distribution of the delay in the read operation, obtained by Monte Carlo simulation (1000 trials). The delay in the read operation is defined as the time interval between the timings when read WL (RWL) is asserted and when the voltage of RBL becomes 0.5 $V_{\rm DD}$. The right graph illustrates the ratio of worst delay to mean delay as a function of supply voltage.

To reduce the large delay margins involved in conventional synchronous subthreshold circuits, Chang *et al.* proposed an asynchronous design approach using a critical-path replica [4]. While this approach can cope with global variation, including die-to-die $V_{\rm th}$ variation and temperature fluctuation, it is difficult to remove the delay margin for local variation, such as within-die $V_{\rm th}$ variation. Sjogren and Myers proposed an asynchronous operation self-timed by completion of pipeline stages [5]. However, it tackles data-dependent execution time variation.

This brief presents a novel subthreshold processor that synchronously operates with a signal that acknowledges memory read completion. This self-timed operation not only reduces the margin for die-to-die manufacturing variability and environmental variation but also mitigates the margin for within-die $V_{\rm th}$ variation, which reduces the execution time and the energy dissipation. In a conventional sensor node, a processor processes data from a sensor module stored in the memory and writes the processed data into the memory for sending to another module, such as a radio-frequency transmitter. Thus, asynchronous processor operation is compatible with other modules in sensor nodes.

II. Subthreshold Self-Timed Processor

A. Architecture Overview

Fig. 2 shows the architecture of the proposed self-timed processor. The processor has a reduced instruction set computer

Manuscript received December 6, 2010; accepted March 6, 2011. Date of publication May 23, 2011; date of current version June 8, 2011. This work was supported in part by the New Energy and Industrial Technology Development Organization (NEDO) of Japan. This work was recommended by Associate Editor G. B. Hwee.

H. Fuketa is with the Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan (e-mail: fuketa@iis.u-tokyo.ac.jp).

D. Kuroda is with the Department of Information Systems Engineering, Osaka University, Osaka 565-0871, Japan (e-mail: kuroda.dan@ist.osaka-u.ac.jp).

M. Hashimoto and T. Onoye are with the Department of Information Systems Engineering, Osaka University, Osaka 565-0871, Japan, and also with the Japan Science and Technology (JST), Core Research for Evolutionary Science and Technology (CREST), Tokyo 102-0075, Japan (e-mail: hasimoto@ist.osaka-u.ac.jp; onoye@ist.osaka-u.ac.jp).

Digital Object Identifier 10.1109/TCSII.2011.2149110

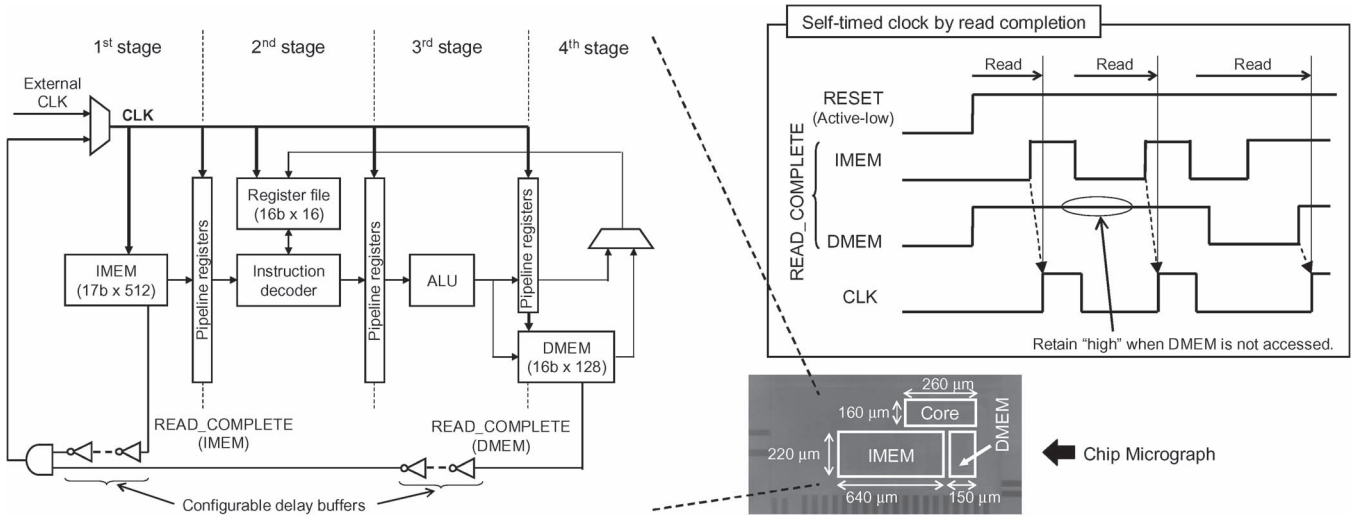

Fig. 2. Diagram and chip micrograph of the self-timed processor synchronously operating with the memory read completion signal.

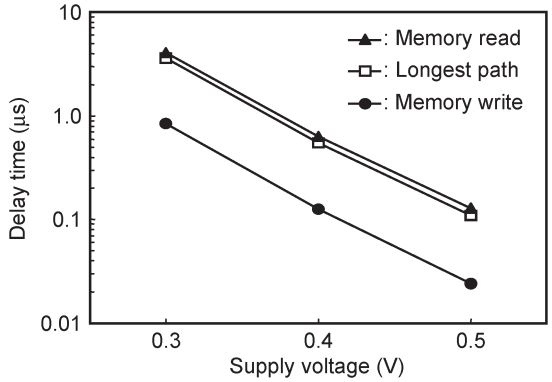

Fig. 3. Simulation result of the delay time required for memory read, memory write, and the longest path in the second and third pipeline stages.

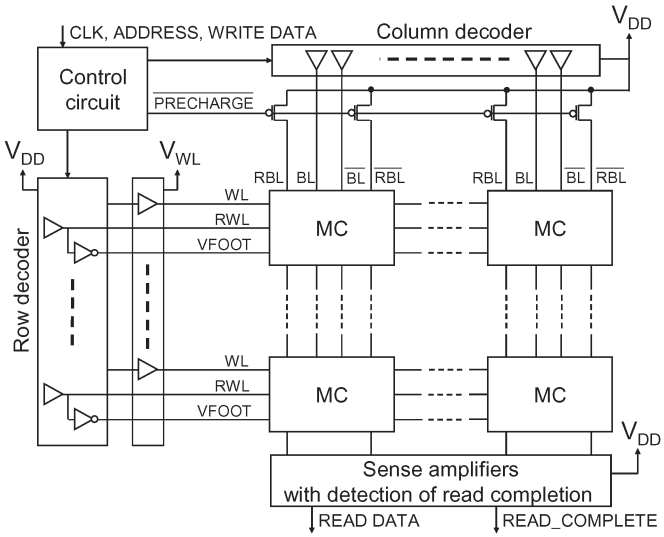

Fig. 4. Circuit structure of subthreshold SRAM with the detection of read completion. The structure of the MC is shown in Fig. 5. SAs with detection of read completion are explained in Fig. 6. V_{WL} is the supply voltage of WL drivers.

core consisting of four pipeline stages with a 16-bit data word and a 17-bit instruction word. The microarchitecture and the instruction set architecture were determined in preliminary work [6]. Twenty basic instructions required for sensor-node applications are implemented. The operation in each stage is completed in a cycle for all the instructions.

The main feature of the proposed processor is that the instruction memory (IMEM) and the data memory (DMEM) can detect the completion of the read operation, and the read completion signal is used as a clock. Fig. 3 shows the delay time required for memory read, memory write, and the longest path in the second and third pipeline stages. This figure indicates that 1) the delay time required for the read operation is longer than that for the write operation and 2) the pipeline stages that include the memory read operation are the slowest. Therefore, the read completion signal can be applied as a clock for this processor. In addition, the processor can also operate synchronously with an external clock for evaluation purpose.

When the RESET signal is asserted, the first instruction code is retrieved from IMEM. The completion signal of the read operation becomes the first clock of the processor. After that, the clock is generated from the read completion signals of IMEM and DMEM, and the instruction codes are read in synchronization with the clock signal. When DMEM is not accessed, the read completion signal of DMEM retains high. Note that access to IMEM occurs in every cycle for the instruction fetch, and hence, the clock signal is continuously generated. Configurable delay buffers are added to the read completion signals to imitate the delay from the output of the memory to the input of the pipeline registers or register files.

The self-timed processor was fabricated in a 65-nm CMOS process. Fig. 2 shows a micrograph of the chip. The processor core and the memory occupy 260 μm $\times$ 160 μm and 790 μm $\times$ 220 μm.

B. Memory Structure

Fig. 4 shows the circuit structure of subthreshold SRAM that detects read completion. Column-mux is not implemented because the memory size is small. Fig. 5 depicts the structure of a 10T MC [7] implemented in the proposed processor, which is based on the subthreshold 8T MC proposed in [3] and is

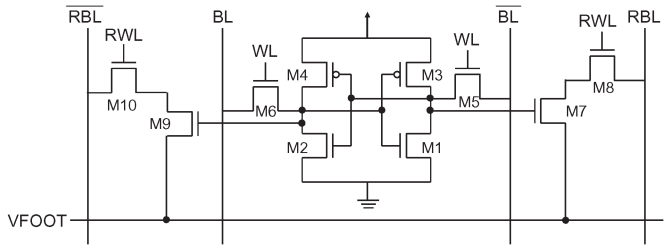

Fig. 5. 10T MC structure. VFOOT is used to ensure the correct operation in the subthreshold region. VFOOT of the accessed word is set to low, and those of the nonaccessed words remain high [3].

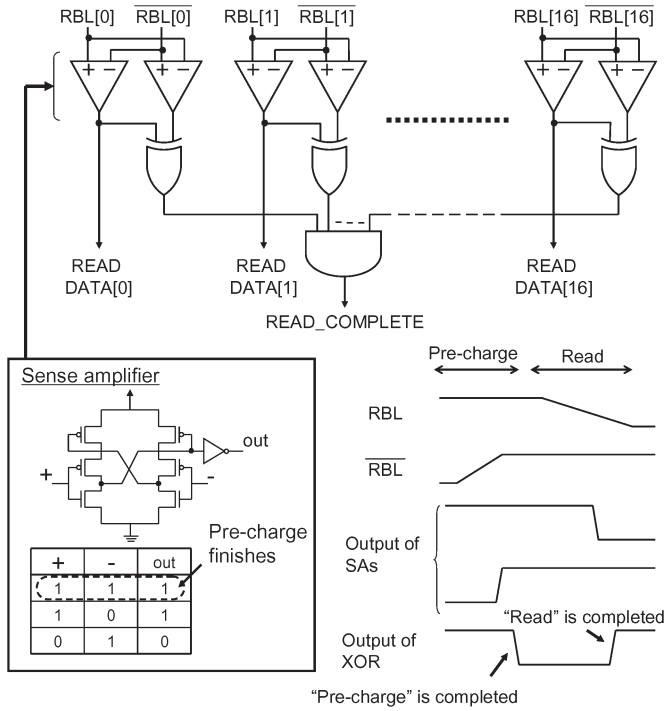

Fig. 6. Mechanism of detecting the completion of the read operation. Two SAs with an XOR gate are used for every bit.

improved such that a differential read operation can be achieved by adding two NMOSs (M9 and M10).

V_{WL} in Fig. 4 is the supply voltage of word line (WL) drivers. The access transistors (M5 and M6 in Fig. 5) must be weak such that the data of unaccessed cells are not corrupted due to leakage currents through the access transistors from bit lines (BLs) in the subthreshold region. We exploit the so-called "boosted word line" technique [3], [8] to ensure the write operation. In this brief, V_{WL} is set to $V_{\mathrm{DD}} + 0.1$ V. In addition, this technique makes the delay time for the write operation much shorter than that for the read operation, as shown in Fig. 3.

C. Detection Mechanism of Read Completion

Fig. 6 shows the mechanism of detecting the completion of the read operation. This mechanism is based on dual-rail coding [9]. In this brief, two sense amplifiers (SAs) with an XOR gate in each bit are used to detect the read completion. The SA output signals are high when both input signals are high, that is, when read BLs (RBL and $\overline{\mathrm{RBL}}$) have been precharged. When the read operation is completed, the output signals of the SAs become

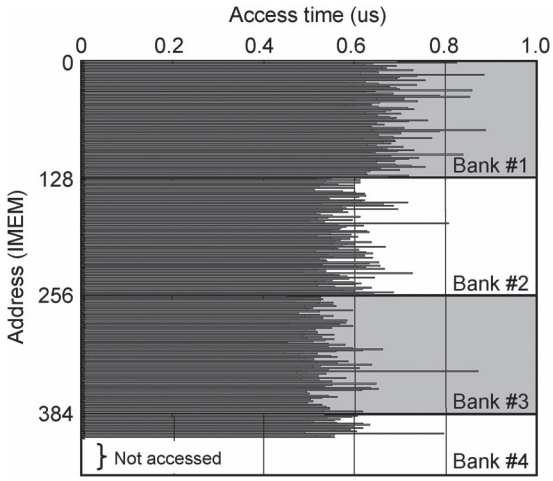

Fig. 7. Measured access time of each address in IMEM once SHA-1 is executed ($V_{\mathrm{DD}} = 0.45$ V).

different, which makes it possible to sense the completion of the read operation.

In our design, the width of the SA is larger than half the width of the MC (the MC width is equal to the column width). Thus, the two SAs are arranged in two rows for each column. As for the IMEM, the overheads of the two SAs on area, active power of the read operation, and leakage power (at 2 MHz, 0.5 V V_{DD}) are 1.5%, 4.4%, and 0.2%, respectively.

III. Measurement Results

A. Average Frequency and Energy of Self-Timed Operation

We measured the execution time and the energy dissipation required to perform SHA-1 ten times for 512-bit data when the processor is operated with a self-timed clock. Fig. 7 shows the access time of each address in IMEM. In the proposed processor, IMEM consists of four memory banks, and each memory bank shares the row/column decoders and the SAs. While the access time depends on the memory bank, the variation in access time of each address is much larger. This means that the delay variation of the read operation for every MC dominates the variation in access time of IMEM.

Fig. 8 shows the measured distribution of the clock cycle time. The total number of cycles to complete the execution is 178 K cycles. Fig. 8 shows that the decrease in supply voltage enlarges the difference between the mean and worst cycle times. This is because the circuit delay becomes more sensitive to V_{th} variation in IMEM and DMEM as the supply voltage is lowered. The proposed self-timed processor operates at the mean clock cycle time on average. Therefore, the processor can operate 2.06 times faster than the conventional synchronous processor with the worst clock cycle time, and the execution time is reduced by 51% when V_{DD} is 0.36 V.

Fig. 9 shows the average frequency, which is the inverse of the mean clock cycle time when the processor operates with a self-timed clock, as a function of the supply voltage in three fabricated chips (chips A–C). The processor correctly operated from 0.36- to 0.56-V supply voltage. Fig. 9 also depicts the energy per cycle of the proposed self-timed processor. The energies of the core and memories are included. When the supply voltage was 0.46–0.48 V, the energy was minimized, and

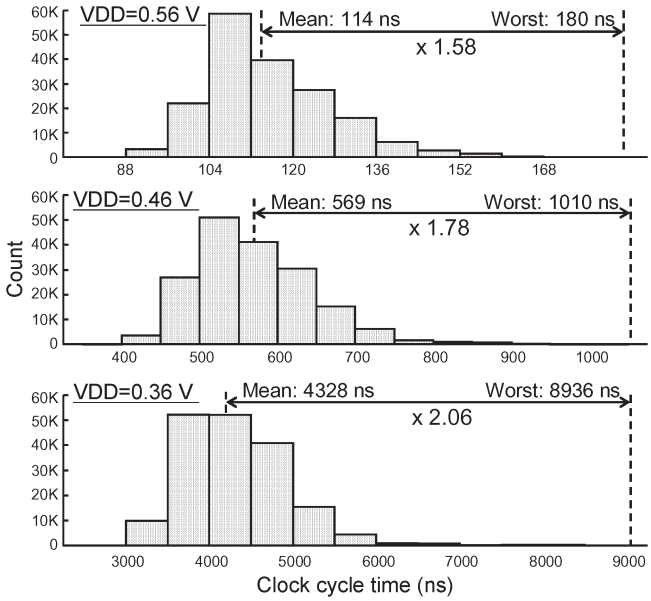

Fig. 8. Measured distribution of the clock cycle time of the self-timed operation when SHA-1 is executed ten times.

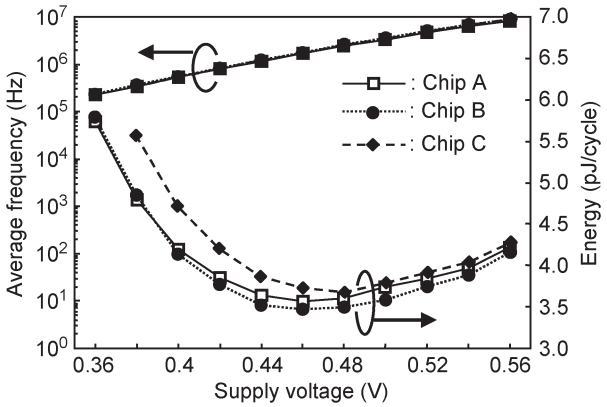

Fig. 9. Average frequency and energy per cycle of the self-timed operation as a function of the supply voltage. The average frequency is the inverse of the mean clock cycle time.

3.47-pJ/cycle operation (the average frequency was 1.76 MHz) was achieved.

The dependences of the average frequency of the self-timed operation on temperature are illustrated in Fig. 10. The supply voltage droops and the decrease in temperature worsened the circuit delay, which requires the margin in the conventional synchronous operation with guardbanding. On the other hand, Figs. 9 and 10 indicate that the self-timed processor can continue to work even if the supply voltage and temperature fluctuate because the clock is adaptively generated in accordance with the supply voltage and temperature.

B. Comparison With Synchronous Operation With Guardbanding

This section compares the proposed self-timed operation with the conventional synchronous operation with guardbanding considering the worst-case process and temperature (PT)

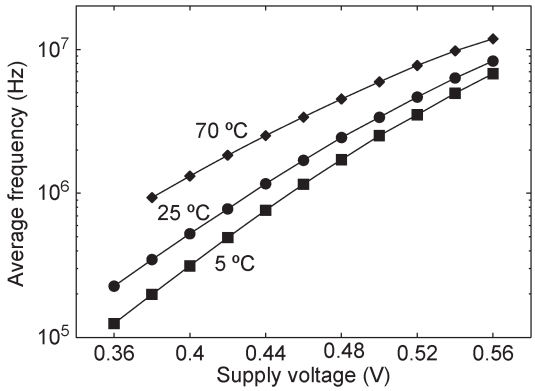

Fig. 10. Dependence of the average frequency of the self-timed operation on temperature.

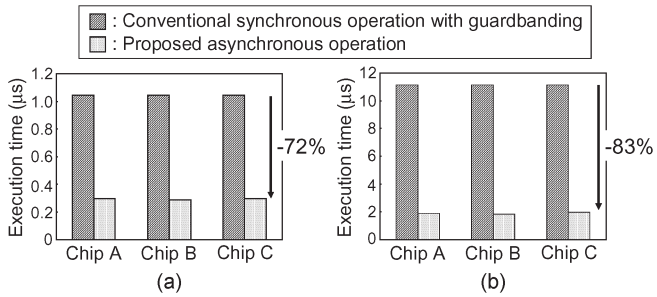

Fig. 11. Execution time of the proposed self-timed operation compared with the conventional synchronous operation with guardbanding at 25 °C when SHA-1 is executed ten times. (a) VDD $= 0.5$ V. (b) VDD $= 0.4$ V

conditions. In this brief, we defined the worst-case clock cycle for the conventional synchronous operation as follows:

1) The slowest chip among the measured 15 chips was chosen.
2) The temperature was set to 5 °C. We determined the minimum clock cycle time (maximum frequency) that sustains correct processor operation.
3) Step 2 was performed at each supply voltage.

Fig. 11 is a comparison with conventional synchronous operation with the worst-case clock when the supply voltage is 0.4 and 0.5 V. The self-timed operation mitigates the delay margin of global variation, such as die-to-die $V_{\rm th}$ variation and temperature fluctuation, in addition to the margin of within-die $V_{\rm th}$ variation, as shown in Fig. 8. Therefore, the execution time at 25 °C is reduced by 72% at $V_{\rm DD} = 0.5$ V and by 83% at $V_{\rm DD} = 0.4$ V compared with the conventional synchronous operation with the worst-case clock. The guardband margin for die-to-die delay variation and temperature fluctuation of the conventional synchronous operation is 58% of the clock cycle time.

Fig. 12 shows the energy per cycle as a function of frequency when the temperature is 70 °C. The proposed self-timed operation is more energy efficient than the conventional synchronous operation. The energy of the self-timed operation needed to attain 7.7-MHz operation, at which the energy is minimized, is 19% smaller than that of the synchronous operation with guardbanding, and the energy needed to achieve 1-MHz operation is 40% less. This is because the conventional synchronous operation requires the higher supply voltage due to the large delay margin, which enlarges the leakage energy.

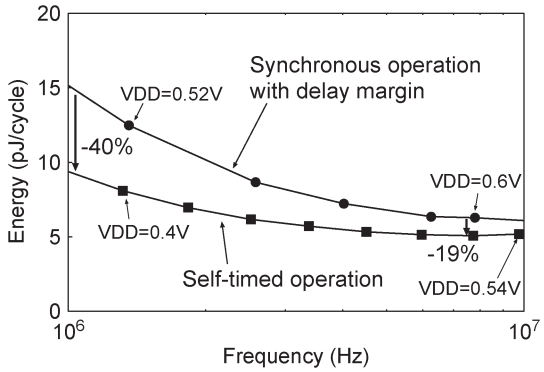

Fig. 12. Energy per cycle at 70 °C of the proposed asynchronous operation with self-timed clock and the conventional synchronous operation with guardbanding considering the worst-case PT conditions. The frequency of asynchronous operation with the self-timed clock is the average frequency.

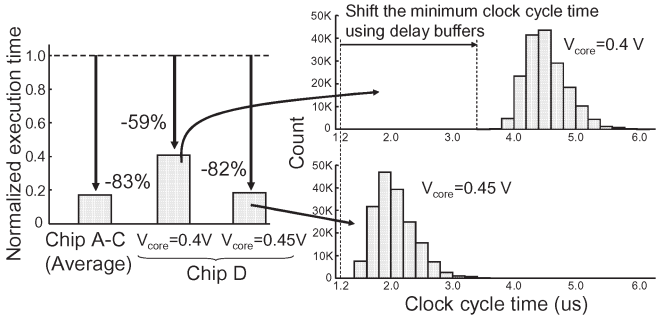

Fig. 13. Execution time of the asynchronous operation with self-timed clock time when $V_{\rm DD}$ is 0.4 V. The execution time is normalized by that of the conventional synchronous operation.

C. Discussion of the Delay Time of the Stages in Which the Memory is Not Accessed

Finally, we discuss the case when the pipeline stages, including the memory read operation (first and fourth stages in Fig. 2), are not the slowest. Fig. 13 shows the execution time of the self-timed operation for a certain chip (chip D) when $V_{\rm DD}$ is 0.4 V. The reduction in the execution time of chip D is 59%, which is much smaller than that of the other chips, because a large buffer delay is inserted into the read completion signal of IMEM to compensate for the slow operation of other pipeline stages that are irrelevant to memory access. In fact, when the supply voltage of the processor core $V_{\rm core}$ is set to 0.45 V, keeping the supply voltage of the memory at 0.4 V, the longer buffer delay is not required, because the higher $V_{\rm core}$ reduces the delay time of the stages in which the memories are not accessed.

One possible solution to deal with this situation is to exploit a replica delay of the critical path in the pipeline stages not including memory access. The clock is generated by the replica delay in addition to the read completion signals, as shown in Fig. 14(a). For example, when the replica delay is 1.5 times larger than the minimum memory access time, the distributions of the clock cycle time are depicted in Fig. 14(b). The clock

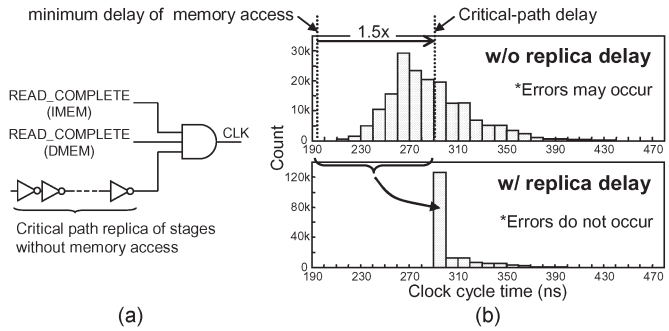

Fig. 14. Asynchronous operation with the clock generated by the critical path delay in addition to the read completion signals. (a) Clock generation with replica delay. (b) Distributions of clock cycle time.

generation with the replica delay removes the clock cycles in which the clock period is less than the replica delay, and consequently, the operation without errors is achieved. In this case, the overheads in execution time and the energy per cycle are 5% and 1%. This indicates that the asynchronous operation with the replica delay is still effective because the variation in the memory access time is much larger.

IV. Conclusion

In this brief, we have presented a subthreshold processor self-timed by memory read completion. The proposed self-timed operation mitigates the delay margin of global and local process, voltage, and temperature variations, which enabled us to reduce the execution time and the energy needed compared with the conventional synchronous operation with guardbanding.

References

[1] M. Seok, S. Hanson, Y. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "The phoenix processor: A 30 pW platform for sensor applications," in *Proc. Symp. VLSI Circuits Dig. Tech. Paper*, 2008, pp. 188–189.

[2] N. Ickes, D. Finchelstein, and A. Chandrakasan, "A 10-pJ/instruction, 4-MIPS micropower DSP for sensor applications," in *Proc. Asian Solid-State Circuits Conf.*, 2008, pp. 289–292.

[3] N. Verma and A. P. Chandrakasan, "A 65 nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *Proc. Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2007, pp. 328–329.

[4] I. J. Chang, S. P. Park, and K. Roy, "Exploring asynchronous design techniques for process-tolerant and energy-efficient subthreshold operation," *IEEE J. Solid-State Circuits*, vol. 45, no. 2, pp. 401–410, Feb. 2010.

[5] A. Sjogren and C. J. Myers, "Interfacing synchronous and asynchronous modules within a high-speed pipeline," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 5, pp. 573–583, Oct. 2000.

[6] D. Kuroda, H. Fuketa, M. Hashimoto, and T. Onoye, "A 16-bit RISC processor with 4.18 pJ/cycle at 0.5 V operation," in *Proc. Symp. Low-Power High-Speed Chips (COOL Chips)*, 2010, p. 190.

[7] H. Fuketa, Y. Mitsuyama, M. Hashimoto, and T. Onoye, "Alpha-particle-induced soft errors and multiple cell upsets in 65-nm 10T subthreshold SRAM," in *Proc. Int. Reliab. Phys. Symp.*, 2010, pp. 213–217.

[8] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T subthreshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, Feb. 2009.

[9] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits,* 2nd ed. Upper Saddle River, NJ: Pearson Educ., 2003.