Low Power Multiple Access Port Register File Design in 100 nm CMOS Technology

Chung-Hsien Hua and Wei Hwang
Institute of Electronics, National Chaio-Tung University, HsinChu, Taiwan

Abstract

A low power multiple access port register file suitable for parallel processing processor is proposed in this paper. New register file cell, read/write port architecture and low power circuit design techniques are used in register file design. Static noise margin under the constraint of multiple access ports is discussed and method to maintain static noise margin is proposed. All the results are simulated in TSMC 100nm CMOS technology. A maximum 5X leakage reduction is achieved by using Dual-Vt transistors in the register file cells and 2X energy saving by adjusting the size of the strong inverter compared to a normal register file cell design. An optimum sizing ratio is found to trade off between energy consumption and transistor size.

1. Introduction

Register files are situated very close to CPU that dissipates lots of power and causes high power density around CPU core. Nano-scale MOSFETs suffer from sub-threshold leakage current. Fig. 1 shows that the leakage current of a register file cell working at 100℃ is 10X larger than 25℃. Therefore, power consumption due to leakage current plays an important role in the nano-scale VLSI design.

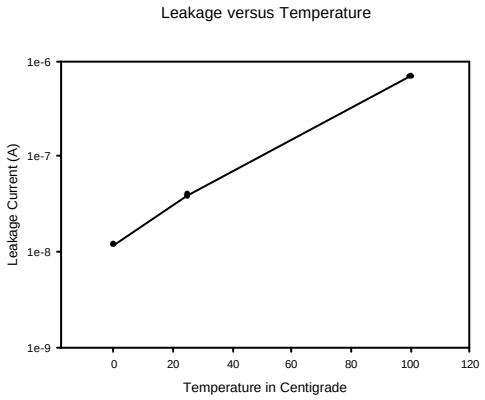


Fig. 1 Leakage Current at different temperatures in 100 nm MOSFET

Trade off between active power and read/write time is important due to energy efficiency issues but another key factor in register file is standby power consumption. As the technology advances, the ability to integrate millions of bits of memory onto a chip becomes a reality. Leakage power in on chip memory would be an important issue in overall power consumption.[1][2][3]

2. Multiple Port Register File Design Considerations

A. Active Power Consumption

Power consumption by read and write operations are simulated. Simulation shows that a transient short between V_{DD} and GND during the write operation consumes a large portion of the total power consumption in the memory cell. Therefore, a new register file cell is proposed and shows great improvement in power delay product.

B. Read/Write Time

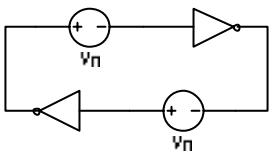
Performance constraint is still an important consideration in low power circuit design. Certain performance must be met while lowering the overall power consumption.

C. Stability/Static Noise Margin

The stability of register file cells determines its soft-error rate and its sensitivity to process tolerances and operating conditions [4][5]. In many cases, the stability of the cell is a critical factor to obtain a desired yield and to lower the cost of the chip. Many different tests and methods exist that try to capture different aspects of the cell's stability.

Static noises are DC disturbances such as offsets and mismatches due to process variations and changes in operating conditions [4][5]. The SNM of a register file cell is the maximum value of DC disturbances that can be tolerated before the cell's storage value is flipped. Fig. 2 shows the storage elements of the register file cell where the static noise sources, Vn are included explicitly.

The SNM of register file cells can be determined graphically by drawing and mirroring the inverter voltage characteristics and finding the maximum possible square between them as shown in Fig. 3 [1]. (V_R and V_L in Fig. 3 are the input voltages of the bottom and top inverters in Fig. 2 respectively.) The SNM can be thought as the noise voltage necessary at each of the cell storage nodes to shift the curve of the two cell inverters vertically or horizontally so that they intersect at only one point, where the cell no longer can reliably store a value. The SNM is measured during the read, since the cell is most vulnerable when the pass-transistors are conducting.



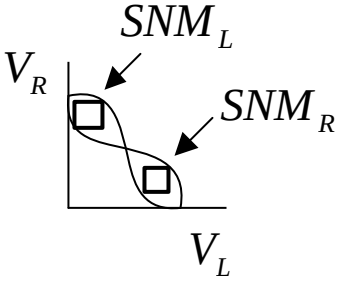


Fig. 3 Static Noise Margin

In order to investigate the impact of port numbers on static noise margin, a modified static noise margin formula shown in (1) takes port numbers into consideration.

The same methodology in [5] is applied to Multi-port SRAM and the SNM are derived.

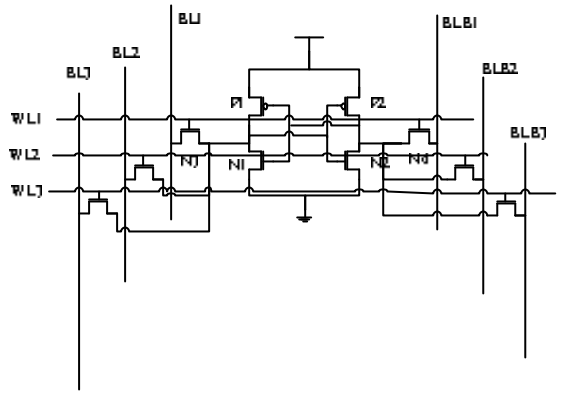


Fig. 4 Multiple port SRAM

$$SNM(V_n) = V_{th} - \frac{1}{(k+1)}\left[\frac{V_{dd} - \frac{2g'+1}{g'+1}V_{th}}{1+\frac{g'}{k(g'+1)}} - \frac{V_{dd}-2V_{th}}{1+k\frac{g'}{q'}+\sqrt{\frac{g'}{q'}\left(1+2k+k^2\frac{g'}{q'}\right)}}\right] \quad (1)$$

where

$$g' = \frac{\frac{W_{N1}}{L_{N1}} m_{effn} C_{ox}}{N_{port}\frac{W_{N4}}{L_{N4}} m_{effn} C_{ox}} \quad (2)$$

$$q' = \frac{\frac{W_{P1}}{L_{P1}} m_{effp} C_{oxp}}{N_{port}\frac{W_{N4}}{L_{N4}} m_{effn} C_{oxn}} \quad (3)$$

$$k = \frac{g'}{g'+1}\left(\sqrt{\frac{g'+1}{g'+1-\frac{V_s^2}{V_r^2}}} - 1\right) \quad (4)$$

$$V_r = V_s - \frac{g'}{g'+1}V_{th} \quad (5)$$

$$V_s = V_{dd} - V_{th} \quad (6)$$

and N_{port} is the number of the access ports.

A normalized static noise margin versus port numbers are shown in Fig. 5. From this figure we know that new read port architecture must be adopted in architectures with many read ports. The static noise margin degrades almost linearly with the increase of access ports due to the read current from the bitline.

3. Low Power Register File Cell Design

A. Dual-Vt Register File Cell

Four types of register file cells are compared in leakage current, speed to write this cell and power-delay product as shown in Fig. 6. Type 1 is the normal type targeted at high performance. Type 2 and Type 3 are modified version to trade performance for lower leakage current. Type 4 is targeted at minimum leakage current. As shown in Fig. 7, Type 3 can save 5X leakage power compared to type 1 with 2% of performance degradation. From Fig. 8 and Fig. 9, Type 3 and Type 4 are featured in minimum power-delay-product which saves energy per operation.

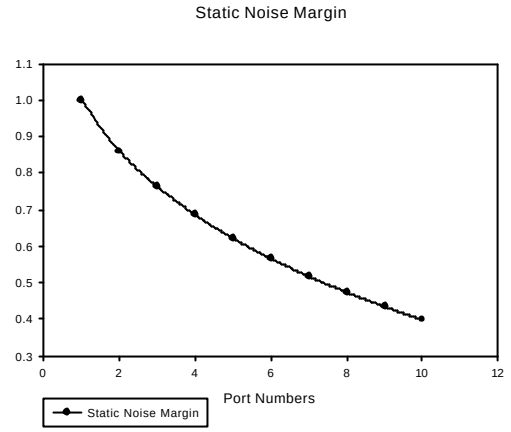


Fig. 5 Static Noise Margin versus Port Numbers

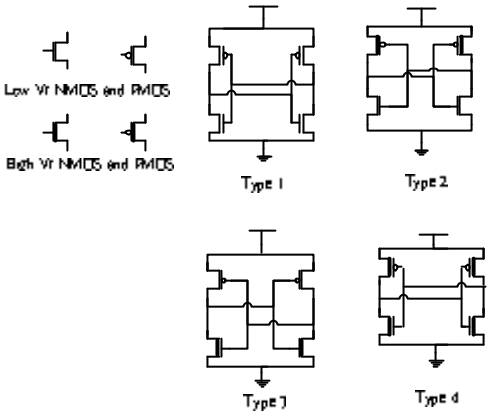


Fig. 6 Four types of MTCMOS memory cells

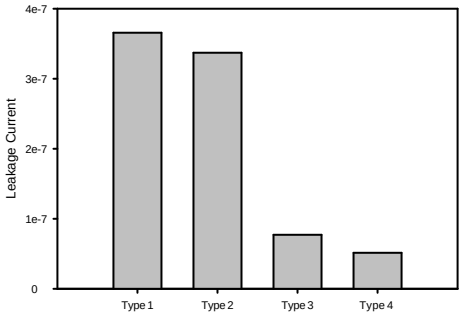


Fig. 7 Leakage Current of the four types of dual Vt register file cell

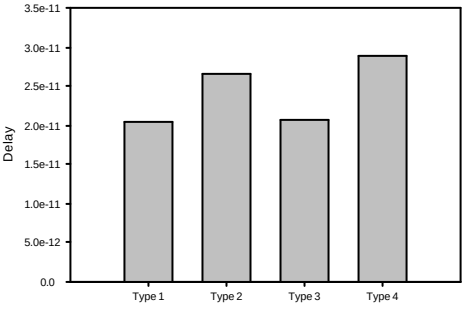


Fig. 8 Delay of the four types of dual Vt register file cell

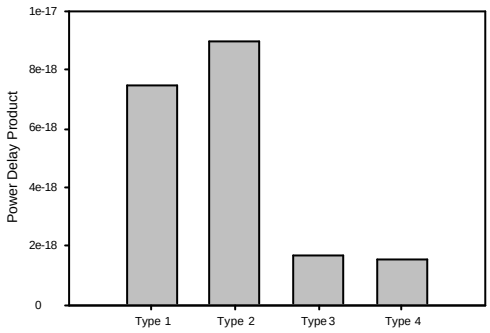


Fig. 9 Power-Delay-Product of the four types of dual Vt register file cell

4. Low Power Access Port Design

A. Low Power Read Port Design

In register files, multiple accesses to the same register file cell are possible. In the traditional SRAM, read port is composed of pass transistor. During the read operation large sink current will pass through n-transistor in the SRAM cell and pull the drain voltage of transistor higher which degrades the static noise margin. Therefore, isolation between sink current and the register file cell is necessary in multiple read port register files as shown in Fig. 10 and Fig. 11. In Fig. 12, static noise margin is maintained by using new read port. Static noise margin of the traditional pass transistor read port is attached in the same figure which is much smaller that the SNM of the new read port. Another feature of register file is its heavy metal routing around register file cell. Therefore, single-ended access port is desired throughout the register file design to alleviate the routing efforts.

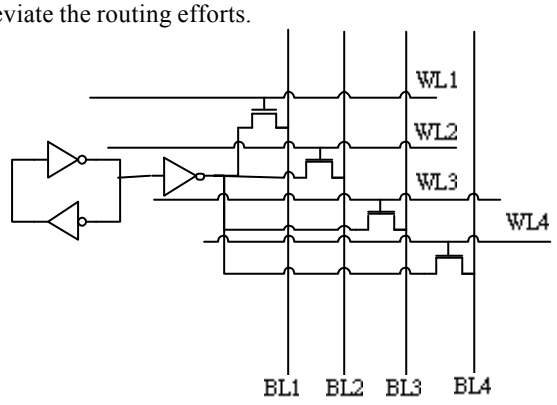


Fig. 10 Using inverter to isolate memory cells form read current

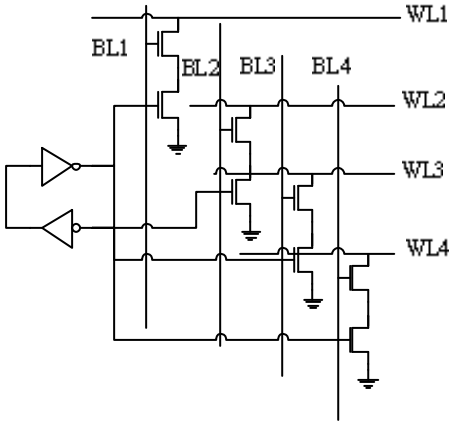


Fig. 11 Using logic circuits to sink read current

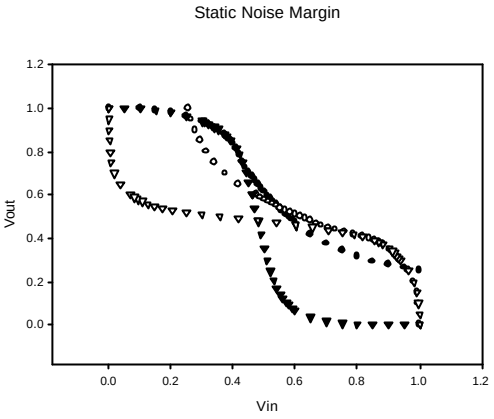


Fig. 12 Static noise margin comparison between traditional port architecture and new port architecture

B. Low Power Write Port Design

The basic components of a register file cell are two inverters connected to form a latch to hold data. Because of such feedback mechanism, writing ‘1’ to a register file cell storing ‘0’ or writing ‘0’ to a memory cell storing ‘1’ (Fig. 13) causes an instantaneous short current from Vdd to ground. In order to minimize or suppress such power consumption, strong inverter is proposed to reduce energy consumption while maintaining performance.

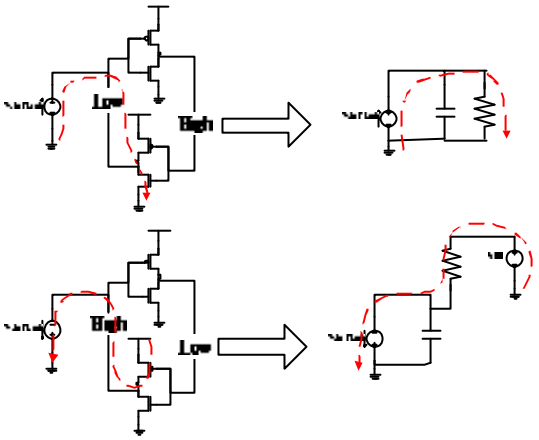


Fig. 13 Instantaneous short current in write operation

Strong inverter technique speeds up the transition process which will definitely reduce the duration of short current. It is good to use asymmetric inverters in single ended write scheme because multiple read ports requires sufficient driving ability to drive all the gate capacitances and the weak inverter. The weak inverter acts as feedback path to hold data as shown in Fig. 14. Due to single-ended access, transmission gates are used as write port to guarantee the value written into the register file cell.

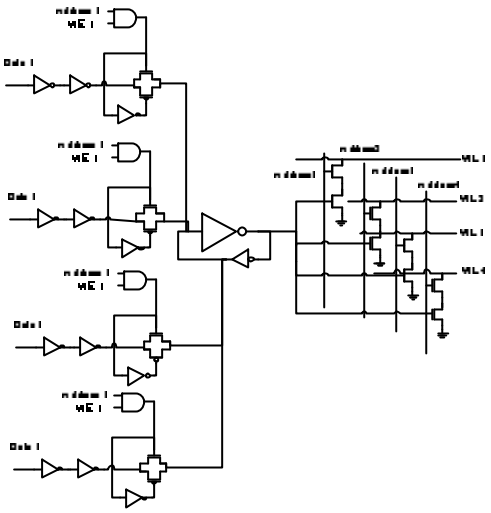


Fig. 14 Strong inverter

Due to the fast transition of the inverters, short current is suppressed significantly as shown in Fig. 15 while maintaining the speed. The power delay product reduces 50% by proper sizing of the strong inverter. Different port numbers that cause different loading at the output of strong inverter will have a different optimum size of the strong inverter.

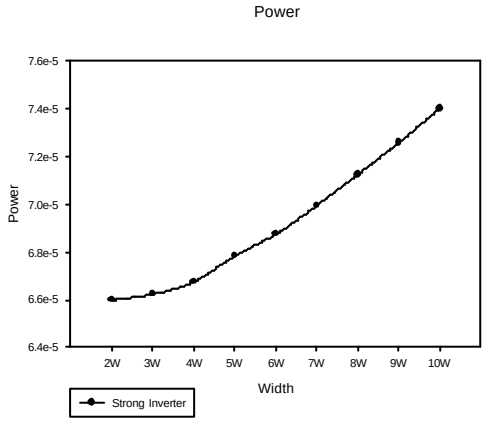


Fig. 15 Power consumption versus different size of strong inverter

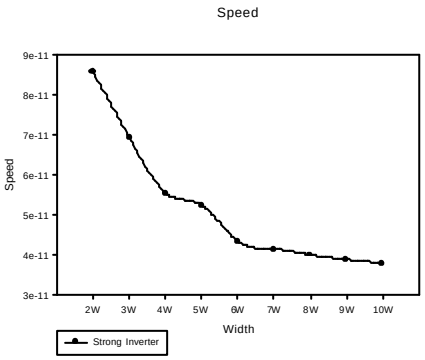


Fig. 16 Access Time versus different size of strong inverter

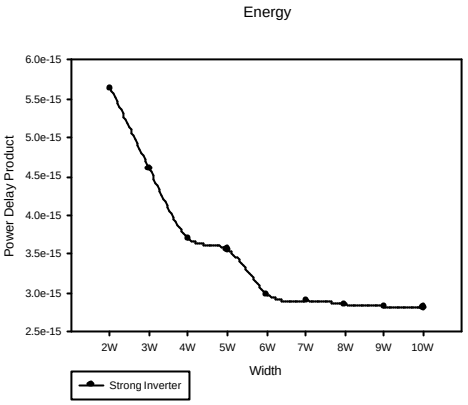


Fig. 17 Power delay product

5. Conclusions

A 4W8R 16word 32 bit register file which occupies 1x1 mm^2 silicon area is implemented. Fig. 18 shows the layout of the register file cell. The register file consumes 2mW when 32 bit data are written into the register file in the critical case and clock runs at 2GHz. The load store operation can be completed with one clock cycle. As can be seen from Fig. 18, metal routing and access ports occupy the major part of the overall layout. Low power register file cell and single-ended Read/Write ports are used to reduce power consumption and is especially useful in nano-scale CMOS technology.

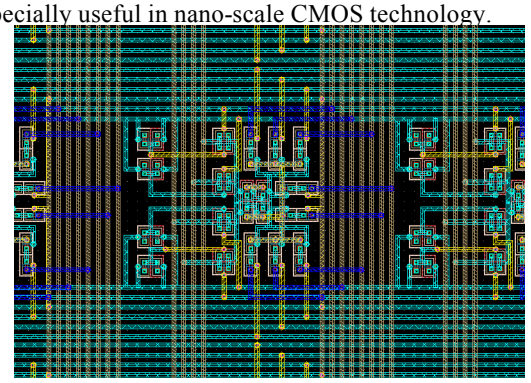


Fig. 18 Layout of Multiple Port Register File Cell

6. Acknowledgement

The authors thank TSMC for providing 100nm CMOS models.

References

[1] Fetzer, E.S.; Gibson, M.; Klein, A.; Calick, N.; Chengyu Zhu; Busta, E.; Mohammad, B.;, "A fully bypassed six-issue integer datapath and register file on the Itanium-2 microprocessor", Solid-State Circuits, IEEE Journal of , Volume: 37 Issue: 11 , Nov 2002, Page(s): 1433 -1440

[2] Tzartzanis, N.; Walker, W.W.; Nguyen, H.; Inoue, A.;, "A 34word x 64b 10R/6W write-through self-timed dual-supply-voltage register file", Solid-State Circuits Conference, 2002. Digest of Technical Papers. ISSCC. 2002 IEEE International , Volume: 2 , 2002, Page(s): 338 -537

[3] Wei Hwang; Joshi, R.V.; Henkels, W.H.;, "A 500-MHz, 32-word×64-bit, eight-port self-resetting CMOS register file", Solid-State Circuits, IEEE Journal of , Volume: 34 Issue: 1 , Jan 1999, Page(s): 56 -67

[4] Seevinck, E.; List, F.J.; Lohstroh, J., "Static-noise margin analysis of MOS SRAM cells", Solid-State Circuits, IEEE Journal of , Volume: 22 Issue: 5 , Oct 1987, Page(s): 748 -754

[5] Bhavnagarwala, A.J.; Xinghai Tang; Meindl, J.D., The impact of intrinsic device fluctuations on CMOS SRAM cell stability, Solid-State Circuits, IEEE Journal of , Volume: 36 Issue: 4 , Apr 2001, Page(s): 658 -665

International Symposium on Nanoelectronic Circuits and Giga-scale Systems (ISNCGS 2004)

A Configurable Scheme for On-Chip Up/Down Voltage Converters in 100nm CMOS Technology

Tung-Shuan Cheng and Wei Hwang

Dept. of Electronics Engineering, National Chiao-Tung University, HsinChu, Taiwan, 300

Abstract

In this paper a configurable scheme for multiple-level voltage generators is presented. Base on this scheme and the use of charge pumps, various voltage levels those are either higher than supply voltage (V_{DD}) or lower than ground (GND) can be produced. As being a back-bias (V_{BB}) generator for reversely biasing the substrate of transistors, it generates voltages between V_{DD} and 2 V_{DD} for PMOS, and voltages between GND and $-V_{DD}$ for NMOS. Moreover, this scheme exhibits the configurability and the output voltage is adjustable. The configurability has been demonstrated and the circuits are simulated by using TSMC 100nm CMOS technology. According to the input settings, various voltage levels can be achieved and the accuracy is higher than 90% without current loading. Moreover, the scheme works well even the supply voltage is lowered to 0.5V.

1. Introduction

In the past years, on-chip voltage generators have been widely used in commercial memory chips such as DRAMs and Flash memory [1]-[3]. For example, DRAM chips need various kinds of power-supply voltages, which have been generated internally by using single external power supply. Not only the advantages in memory designs, voltage generators are beneficial in other digital ICs as well. In the future of low-voltage CMOS ASIC and SoC (System-on-Chip) designs, internally generated voltages will be indispensable to reduce the subthreshold current that exponentially increases with a decreasing threshold voltage. The subthreshold leakage is given as [4]

$$I_{leakage} = I_0 \cdot 10^{(VGS-Vt)/n\frac{kT}{q}\ln 10} \quad (1)$$

where I_0 is a constant, k is the Boltzmann's constant, T is the absolute temperature, and n is the subthreshold swing coefficient constant. It's obvious from (1) that subthreshold leakage is inversely proportional to the threshold voltage, V_t. Many techniques have been applied to raise V_t and reduce the subthreshold leakage. The most popular and efficient way is to apply reverse back bias (RBB). The circuit techniques such as VTCMOS (variable threshold-voltage CMOS) [5] and Vth-hopping [6] have been realized to dynamically vary the threshold voltage to reduce active and standby power dissipation. Since the subthreshold leakage is the dominant component of the whole leakage currents, it can significantly reduce the standby leakage by applying RBB. However, applying large RBB will increase some other leakage currents in the system. For example, band-to-band tunneling (BTBT) currents can deteriorate to the level over the subthreshold leakage and dominant the drain leakage [7]. Moreover, many researches have demonstrated that there is an optimal reversed bias point and is unique to any specific technology generation [8-9]. There are some kinds of back-bias voltage generators and they work well, but the common disadvantage among them is that they generate single one voltage level [10-11]. Usually, that generated voltage is not the optimal value and several kinds of voltage generators are needed in a large design. In this paper we propose a novel architecture that can generate various voltage levels by configuring the input settings. In Section 2, the architecture of the proposed configurable multiple-level voltage generator will be described. Moreover, two back-bias generators based on the architecture are presented in Section 3. Section 4 shows the simulation results of both the circuits in Section 3. Finally, we make some conclusions in Section 5.

2. Configurable Scheme

Figure 1 shows the proposed scheme of configurable multiple-level voltage generator. It's mainly composed of ring oscillator, code converter, D/A converter, initial control, charge pump, and recovery circuit. The ring oscillator supplies the required pumping signals internally. It's a basic inverter chain with odd number of inverters and an enable control. The D/A converter is the most crucial part of the whole architecture. The D/A converter is used to generate clocking signals whose swings are equal or less than V_{DD}, according to the binary inputs. Due to the better linearity, a unary-weighted D/A converter is preferred. With the use of unary-weighted D/A converter and binary inputs, a code converter that transforms binary codes to thermometer codes is necessary. Table 1 shows the example of 3-bit binary-to-thermometer transformation. The initial control initializes the D/A converter and maintains the linearity.

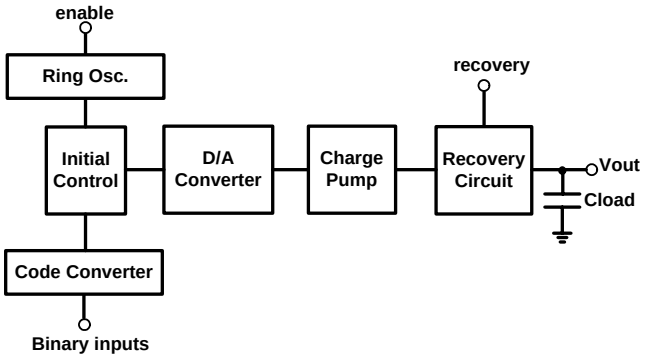


Figure 1. The proposed architecture of configurable multiple-level voltage generator.

Table 1. The 3-bit binary-to-thermometer transformation.

binary inputs	thermometer outputs
D2 D1 D0	T6 T5 T4 T3 T2 T1 T0
0 0 0	0 0 0 0 0 0 0
0 0 1	0 0 0 0 0 0 1
0 1 0	0 0 0 0 0 1 1
0 1 1	0 0 0 0 1 1 1
1 0 0	0 0 0 1 1 1 1
1 0 1	0 0 1 1 1 1 1
1 1 0	0 1 1 1 1 1 1
1 1 1	1 1 1 1 1 1 1

From the description above, clocking signals with various swing voltages can be obtained and the swings are configurable. These clocking signals can be fed into a charge pump to generate voltages either higher than V_{DD} or lower than GND. The charge pump can be either pump-up or pump-down charge pump. In the next section, two voltage generators based on this architecture will be presented to demonstrate the configurability and flexibility of this architecture. If we treat the circuits as back-bias generators, which supply reversed back bias voltage in the standby modes, they should have the ability to return to the nominal voltage values. Therefore, a recovery circuit as shown in Figure 1 is necessary to execute the recovery operation. In the following we show two back-bias generators based on the architecture and one generates negative voltages and the other generates positive voltages higher than V_{DD}.

3. Back-Bias (V_{BB}) Generators

In this section, two back-bias generators are constructed and then the configurability and flexibility of the proposed architecture will be demonstrated.

3.1 V_{BB} Generator For NMOS

Since almost all the NMOS transistors connect their Sources to the GND, the applied back-bias voltages must be lower than zero. It means that charge pumps generate negative-value voltages are required. Conventional pump-down charge pumps are composed of diode-connected NMOS and kicking capacitors [12]. The simple structure has a drawback, however, it suffers from body effect and the output voltage is shallower. Some pump-down charge pumps without V_t-loss have been developed, as shown in Figure 2(b) [10] and (c) [11]. Both of the two charge pumps suffer no V_t-loss and produce voltage that is almost equal to $-V_{DD}$, where V_{DD} is the voltage swing of the pumping signals. Therefore, charge pumps without V_t-loss are preferred here due to the higher flexibility. Generally speaking, if the swings of the pumping signals are V1, the charge pumps produce voltage equals to $-V1$. That's why the D/A converter is applied to generate various clocking signals. By feeding various clocking signals into the charge pumps, various voltage levels are obtained. The output voltage expression is given as

$$Vout = -V_{DD} * \left(\frac{input\ \ value}{2^{(no.\ \ of\ \ inputs)}-1}\right) \qquad (2)$$

where the term "input" above means the binary inputs. As for the recovery circuit, it should return output voltage to the level of GND. A proposed circuit and the operation table are shown in Figure 3. When the signal recovery is high, the V_{BB} generator operates normally and pumps down to certain voltage levels. If recovery goes low, the PMOS turns ON and thus the NMOS at the bottom turns ON as well. Therefore, Vout is pulled to GND.

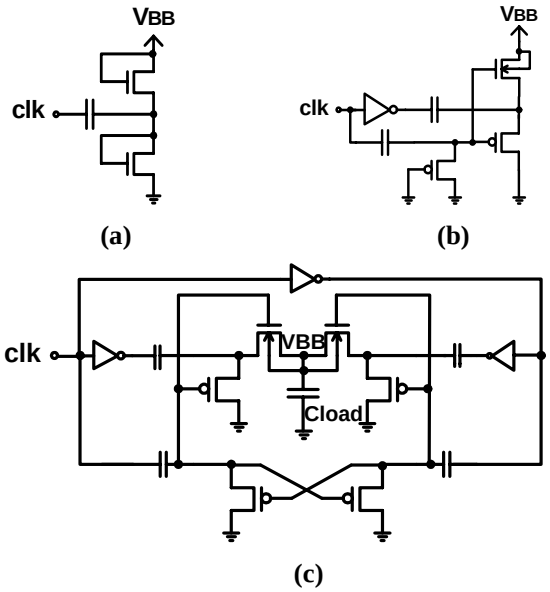
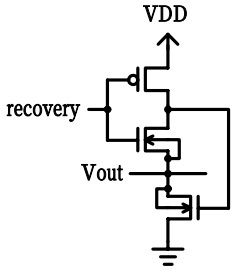


Figure 2. (a) Conventional pump-down circuit. (b) Pump-down circuit without Vt-loss in [10]. (c) Pump-down circuit without Vt-loss in [11].



recovery	Vout
0	GND (recovery)
1	Pump down

Figure 3. The recovery circuit and the operation table.

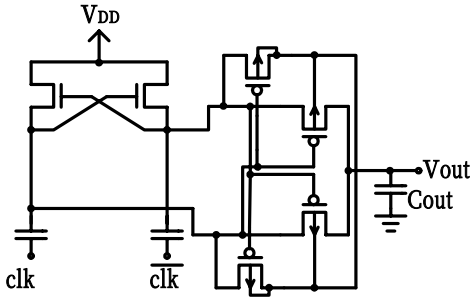
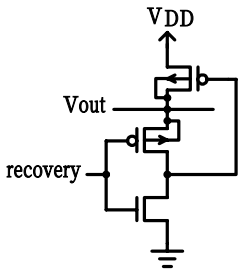


Figure 4. A voltage doubler proposed in [15].



recovery	Vout
0	Pump up
1	VDD (recovery)

Figure 5. A useful recovery circuit [13].

3.2 V_{BB} Generator For PMOS

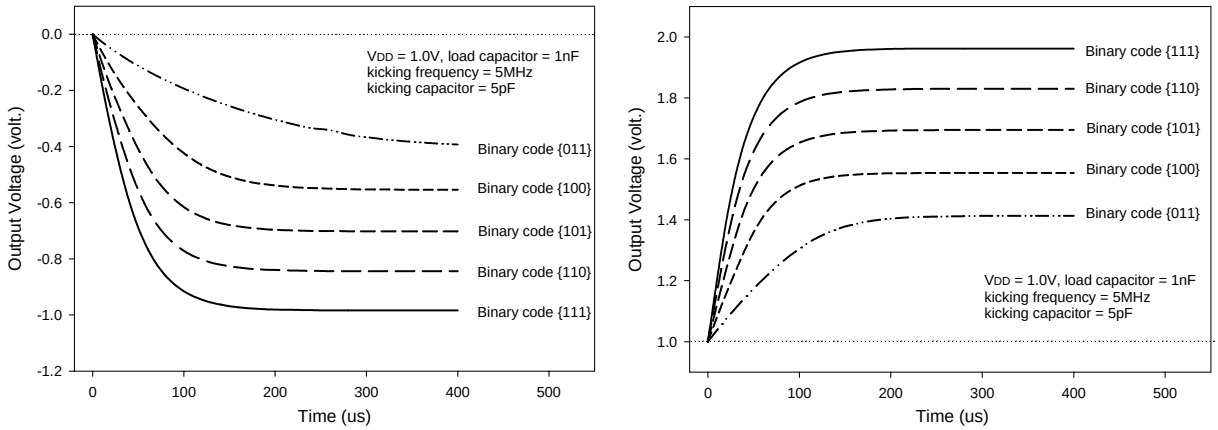
In contrast to V_{BB} generator for NMOS, as described previously, voltages that are higher than V_{DD} are required to reversely bias the substrate of PMOS. However, the voltages cannot be too large due to the excess BTBT currents, and therefore output voltages that are between V_{DD} and $2V_{DD}$ are preferred. Most pump-up charge pumps are based on the circuit proposed by Dickson [14], and the circuit is called "Dickson charge pump". The circuit is similar to Figure 2(a) and composed of diode-connected MOSFET's and kicking capacitors. A voltage doubler can generate twice the magnitude of the input voltage, and it's quite suitable for the application here. Figure 4 shows a high-performance voltage doubler proposed in [15], and it's composed of a cross-coupled structure and series switches. In general cases, the swing of pumping signal is V_{DD} and the output voltage is twice the magnitude of V_{DD}. In the application here, however, the swing of clk is not always equals V_{DD} and it doesn't behave as a voltage doubler. The output voltage equation is given as

$$V_{out} = V_{DD} + V_{DD} * \left(\frac{input\ \ value}{2^{(no.\ of\ \ inputs)}-1}\right) \qquad (3)$$

where the term "input" above means the binary inputs. The recovery circuit used here must have the ability to pull output voltage to the level of V_{DD}, the nominal substrate voltage of PMOS. As in Figure 5 [13], it has no influence on output when recovery is low, and forces output to V_{DD} when recovery goes high. When recovery is low, the PMOS in the middle turns ON and the topmost PMOS is OFF. When recovery goes high, the bottommost NMOS turns ON and causes the topmost PMOS to turn ON, thus the output voltage is forced to be V_{DD}.

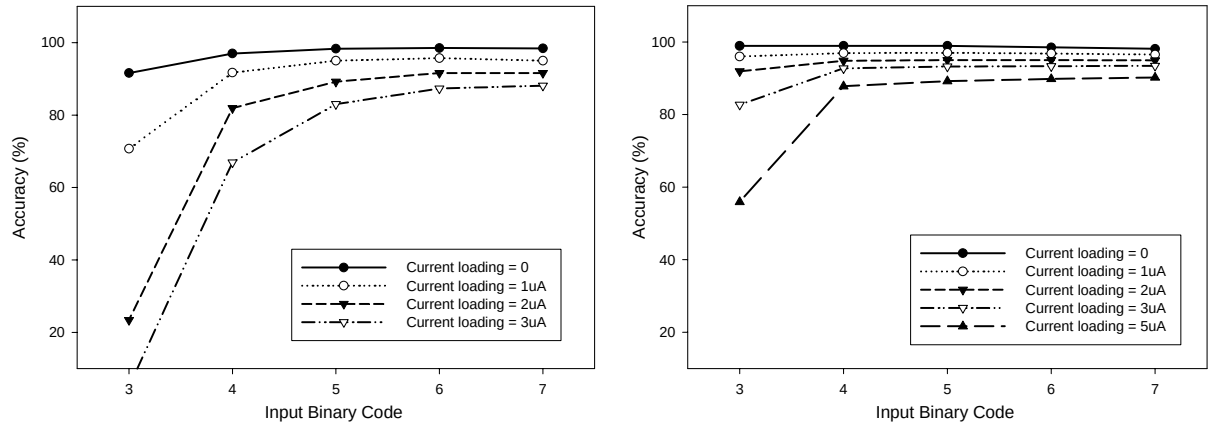
4. Simulation Results

In this section, we construct two V_{BB} generators based on the proposed architecture, as described in the previous sections, and the simulation results will be shown. The simulation is done by HSPICE simulation with the spice parameter of TSMC 100nm CMOS technology. Figure 6 shows the output transient waveforms of the V_{BB} generators with different values of the binary inputs. Note that the number of bits of the binary inputs is 3. Figure 6(a) is the V_{BB} generator for NMOS and Figure 6(b) is for PMOS. Both the figures not only illustrate the flexibility of the proposed architecture, but also demonstrate the feature of configurability. Various voltage levels can be achieved according to the configurable binary inputs. Some waveforms with smaller binary codes are not shown because their pumping speeds are much slower in comparison with the larger binary codes. Figure 7 illustrates the accuracy with and without the influence of current loading. The accuracy is defined as the ratio of simulated output voltage and ideal output voltage. The output accuracy is higher than 90% without current loading. The applied current loading is to be served as the substrate current flows in the substrate of a chip, and it degrades the pumping efficiency and accuracy. In both of the situations in Figure 7, the accuracy starts to degrade severely at the binary input code {011}, and the accuracy curves maintain flat with input codes larger than {011}. The output voltage levels are settled by the charge pumping current and the substrate current loading. The absolute value of output voltage is larger with larger pumping current, which is proportional to the voltage swing of the pumping signals. Figure 8 shows the output voltage versus different operating supply voltages. Figure 8 (a) and (b) demonstrate that the proposed scheme works well even when the supply voltage is lowered to 0.5V.



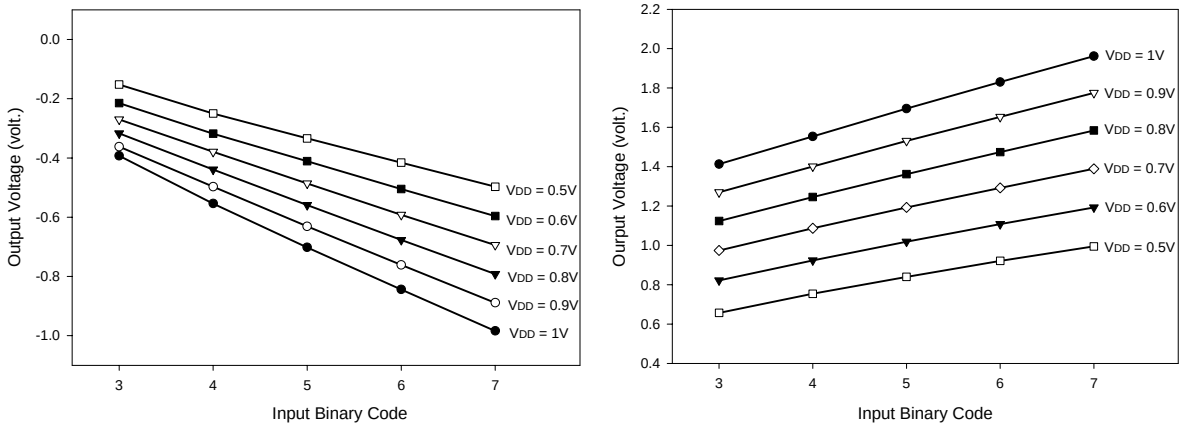
(a) **(b)**

Figure 6. Output transient waveforms of the V_{BB} generators (a) for NMOS and (b) for PMOS.



(a) **(b)**

Figure 7. Accuracy versus current loading of the V_{BB} generators (a) for NMOS and (b) for PMOS.



(a) **(b)**

Figure 8. Output voltage versus different supply voltages. (a) V_{BB} generators for NMOS and (b) for PMOS.

5. Conclusion

The architecture of configurable multiple-level voltage generators is proposed and various voltage levels can be achieved through the configurable inputs. Base on this architecture, two paradigmatic V_{BB} generators are constructed and they can reversely bias the substrate of transistors. From the simulation results, these circuits indeed produce various voltage levels with different input settings. As mentioned above, an optimal reversed bias point exists and is unique to any specific technology. Whatever the optimal bias point is, it can be achieved by using this architecture with proper input setting. If using simple charge pumps as in Figure 2, they cannot always generate the optimal voltages required. Besides, several different charge pumps are needed when various voltage levels are desired in a system. It's bothersome to deal with so many kinds of charge pumps. Therefore, the configurability and the reusability of this architecture are quite beneficial in SoC designs. Moreover, the accuracy of output voltage without current loading is almost higher than 90%, and the scheme works well even the supply voltage is lowered to 0.5V in 100nm CMOS technology.

6. References

[1] [1] Kiyoo Itoh, Katsuro Sasaki, and Yoshinobu Nakagome, "Trends in Low-Power RAM Circuit Technologies," Proc. of the IEEE, vol. 83, No. 4, April 1995.

[2] Kiyoo Itoh, Yoshinobu Nakagome, Shin'ichiro Kimura, and Takao Watanabe, "Limitations and Challenges of Multigigabit DRAM Chip Design," *IEEE J. Solid-State Circuits*, vol. 32, No. 5, pp. 624-634, May 1997.

[3] Hiroo Masuda, Ryoichi Hori, Yoshiaki Kamigaki, Kiyoo Itoh, Hiroshi Kawamoto, and Hisao Katto, "A 5 V-Only 64K Dynamic RAM Based on High S/N Design," *IEEE J. Solid-State Circuits*, vol. SC-15, No. 5, pp. 846-854, Oct. 1980.

[4] A. Keshavarzi, K. Roy, and C. Hawkins, "Intrinsic Leakage in Low Power Deep Submicron IC's," *Proc. Int. Test Conf.*, pp. 146–155, Nov. 1997.

[5] Tadahiro Kuroda, Tetsuya Fujita, Shinji Mita, Tetsu Nagamatsu, Shinchi Yoshioka, Kojiro Suzuki, Fumihiko Sano, Masayuki Norishima, Masayuki Murota, Makoto Kako, Masaaki Kinugawa, Masakazu Kakumu, and Takayasu Sakurai, "A 0.9V, 150-MHz, 10-mW, 4mm2, 2-D Discrete Cosine Transform Core Processor With Variable Threshold-Voltage (VT) Scheme," *IEEE J. Solid-State Circuits*, vol. 31, pp. 1770-1779, Nov. 1996.

[6] Koichi Nose, Masayuki Hirabayashi, Hiroshi Kawaguchi, Seongsoo Lee, and Takayasu Sakurai, "Vth-Hopping Scheme to Reduce Subthreshold Leakage for Low-Power Processors," *IEEE J. Solid-State Circuits*, vol. 37, pp. 413-419, March 2002.

[7] Ming-Jer Chen, Huan-Tsung Huang, Chin-Shan Hou, and Kuo-Nan Yang, "Back-Gate Bias Enhanced Band-to-Band Tunneling Leakage in Scaled MOSFET's," *IEEE Electron Device Lett.*, vol. 19, pp. 134-136, 1998.

[8] Cassondra Neau, and Kaushik Roy, "Optimal Body Bias Selection for Leakage Improvement and Process Compensation Over Different Technology Generations," *Proc. of ISLPED*, pp. 116-121, 2003.

[9] Yo-Sheng Liu, Chung-Cheng Wu, Chih-Sheng Chang, Rong-Ping Yang, Wei-Ming Chen, Jhon-Jhy Liaw, and Carlos H. Diaz, "Leakage Scaling in Deep Submicron CMOS for SoC", *IEEE Trans. on Electron Devices*, vol. 49, pp. 1034-1041, June 2002.

[10] Y. Tsukikawa, T. Kajimoto, Y. Okasaka, Y. Morooka, K. Furntani, H. Miyamoto, and H. Ozaki, "An Efficient Back-Bias Generator With Hybrid Pumping Circuit for 1.5 V DRAM's," *IEEE J. Solid-State Circuits*, vol. 29, pp. 534–538, Apr. 1994.

[11] Kyeong-Sik Min, and Jin-Yong Chung, "A Fast Pump-Down V_{BB} Generator for Sub-1.5V DRAMs," *IEEE J. Solid-State Circuits,* vol. 36, pp. 1154–1157, July, 2001.

[12] Katsuyuki Sato, Hiroshi Kawamoto, Kazumasa Yanagisawa, Tetsuro Matsumoto, and Shinji Shimizu, "A 20ns Static Column 1Mb DRAM in CMOS Technology," *ISSCC Dig. Tech. Papers*, pp. 254–255, Feb. 1985.

[13] Seong-Ik Cho, Jung-Hwan Lee, Hong-June Park, Gyu-Ho Lim, and Young-Hee Kim, "Two-Phase Boosted Voltage Generator for Low-Voltage DRAMs," *IEEE J. Solid-State Circuits,* vol. 38, pp. 1726–1729, Oct., 2003.

[14] J. F. Dickson, "On-Chip High-Voltage Generation in NMOS Integrated Circuits Using an Improved Voltage Multiplier Technique," *IEEE J. Solid-State Circuits*, vol. 11, pp. 374–378, June 1976.

[15] Pierre Favrat, Philippe Deval, and Michael J. Declercq, "A High-Efficiency CMOS Voltage Doubler," *IEEE J. Solid-State Circuits,* vol. 33, pp. 410–416, Apr. 1998.

THE MICROARCHITECTURE OF A LOW POWER CLUSTERED REGISTER FILE FOR PARALLEL PROCESSORS

Chung-Hsien Hua and Wei Hwang

Institute of Electronics, National Chiao-Tung University, Hsin-Chu 300, Taiwan
cshua.ee90g@nctu.edu.tw & hwang@eic.nctu.edu.tw

ABSTRACT

The access time, power consumption and silicon area of the register file microarchitecture are critical to the overall performance in multiple issue microprocessors. These terms grow super-linearly with the number of read and write ports. A clustered register file with global registers is presented in this work. This leads to the result of reducing silicon area, power consumption and increase speed at the same time. The flexibility to access results from other function units is improved in this global register architecture. Based on the simulation results suing TSMC 180nm CMOS technology, the proposed clustered register file with global registers exhibits up to 70% reduction in silicon area, 20% increase in operation frequency, 12% active power consumption reduction and 28% reduction in power delay product compared to the central register file architecture. The clustered register file with global registers has 128words x 32bits in size with eight read ports and four write ports. It has been implemented in 180nm CMOS technology.

1. INTRODUCTION

The future processors for computing multimedia data, communication modulations and encryption or decryption algorithms require tens to hundreds of GOPS (Giga-Operations-Per-Second) computation capabilities. With such high demand in computation capabilities, digital signal processors or parallel processors are becoming a vital element in the future application systems to handle the whole calculations.

As the number of function units in the processors increase to these levels, the microarchitecture of register file and communication between function units become critical factors dominating the area, cycle time, and power consumption of the processor. This is because such register files needed to be large enough to support multiple in-flight instructions and multiple-port to avoid stalling the multiple issues. These factors also mean they contribute significantly to the processor's power consumption. For N function units, the area of the register file grows as N^3, the delay as $N^{3/2}$, and the power dissipation as N^3.

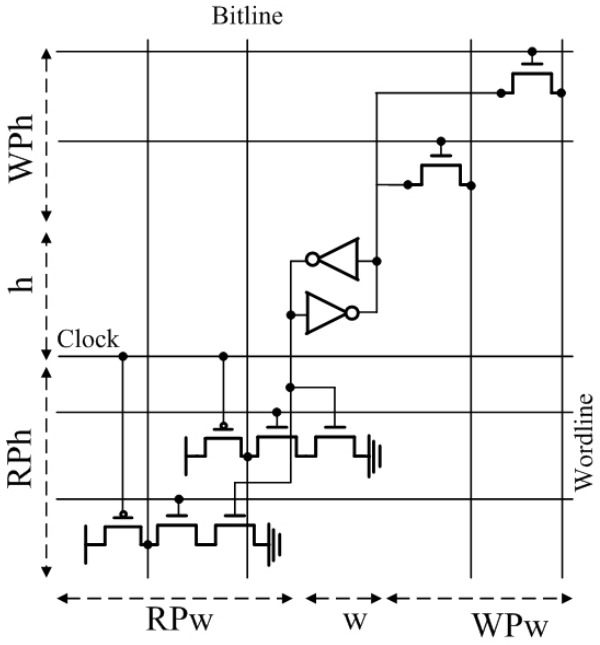


Fig. 1 Schematic view of single register file cell with two read ports and two write ports

The area of register file array is the product of the number of registers, R, the number of bits per register, b, and the size of a register file cell. The schematic of a register file cell[1] is given in Fig. 1. It shows that each cell is w+WPw+RFw wide and h+WPh+RPh high. WPw and WPh are the width and height of the whole write ports, respectively. RPw and RPh are the width and height of the whole read ports.

The delay of a register file access is composed of two parts: the wire propagation delay and fan-in/fan-out delay. The wire propagation delay grows linearly with distance under optimally spaced repeater. The fan-in/fan-out delay is the minimum drive delay of a lumped capacitive load using a buffer chain, which grows logarithmically with the lumped capacitance. In deep-submicron-meter or nano-meter technology, the wire propagation delay dominates the access delay. Long wires such as bitline or wordline make the wire propagation delay even larger. A generalized register file access is shown in Fig. 2 and a detailed view of the hierarchical bitline and wordline connections is shown

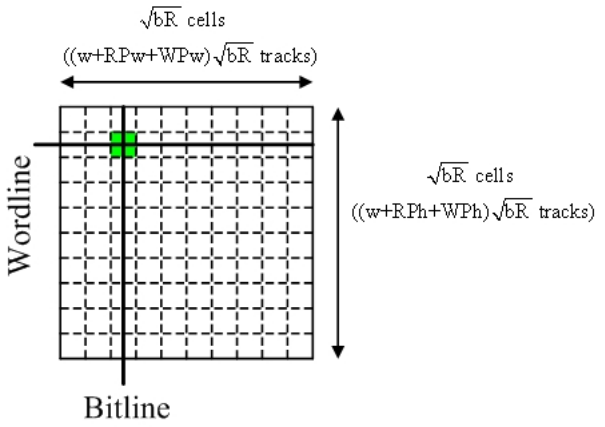


Fig. 2 Access and area estimation for basic block of register file array

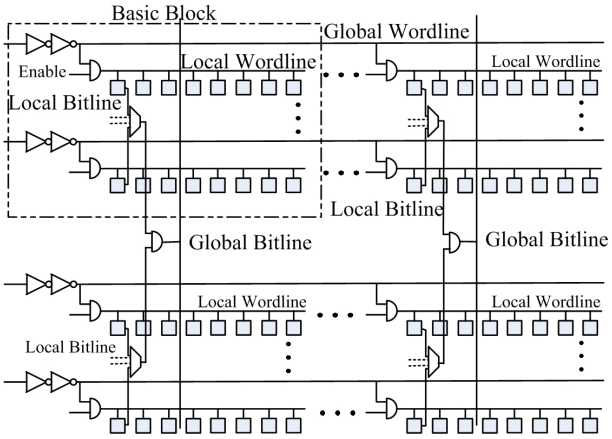


Fig. 3 Detailed view of the hierarchical bitline and wordline connections

in Fig. 3 The signal must traverse a wordline of length $(RPh + h + WPh)\sqrt{bR}$ and then a bitline of length $(RPw + w + WPw)\sqrt{bR}$, resulting in a delay that is proportional to the number of total access ports and the number of register file cells.

The energy dissipated in a register file is proportional to the capacitance that must be switched for each access. Since every bitline and only a single wordline must be switched for every register file access, the power dissipation is dominated by the bitlines' capacitance. For a register file with a large number of ports, this capacitance is mostly wire capacitance. As shown in Fig. 2, each port has $\sqrt{bR}$ bitlines that connect $\sqrt{bR}$ register cells, resulting in a wire capacitance proportional to $bR(h + RPh + WPh)C_w$. The number of ports in central register files increases with the number of ALUs. As can be seen in Fig. 1, increasing the number of ports in a register file will increase the length of wires in each port, while the number of transistors on those wires remains the same. Therefore, for register files with a large number of ports, the power dissipation of total p ports is dominated by wire capacitance and grows as p^2R. Register file architectural classification and

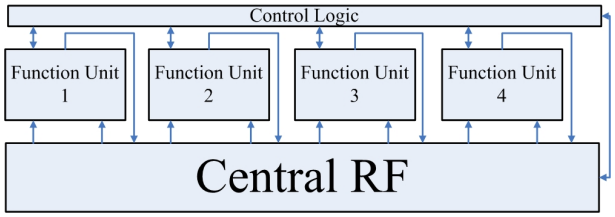


Fig. 4 Central register file architecture

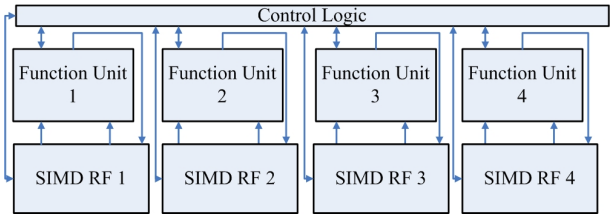


Fig. 5 SIMD register file architecture

modeling are also shown in reference [2], [3], [4], [5] and [6].

The analysis shown above can serve as a baseline comparison of different register file architectures such as the basic block shown in Fig. 3. Different register file architectures may have different analytical formulation. However, the basic idea shown here applies conceptually to all kinds of register files.

2. CENTRAL AND SIMD REGISTER FILE

The conventional register file architecture in SISD processors is a unified register file with two read ports and one write port. These two read ports provide the required data to function units and the write port writes the data from function units to register file. The area of centralized register file is proportional to $(\text{number of cells})(\text{number of ports})^2$. In a four issue centralized register file as shown in Fig. 4, each function unit can read and write every individual register file cell in the central register file. To achieve this goal, each register file cell is connected to eight read ports and four write ports. Large decoding circuits are needed to allocate the required register file cell and a large portion of silicon area is occupied by routing metals which is a dominant factor in speed and power consumption in deep-submicron-meter and nano-meter technologies.

The central register file of 128 words X 32 bits with 8 read ports and 4 write ports occupies around 1550um X 1550um silicon area in 180nm CMOS technology. Detailed performance comparisons are shown in Table 1, Table 2 and Table 3.

The cost of the register file can be reduced for data-parallel media applications by partitioning the registers across groups of function units. When a Fig. 6 data-parallel loop is unrolled k times there is little or no communication between the iterations. Thus we can partition the register file into k identical clusters by moving the expensive, but little-used, communication paths between clusters. This type of register file architecture is widely used in SIMD and vector processors.

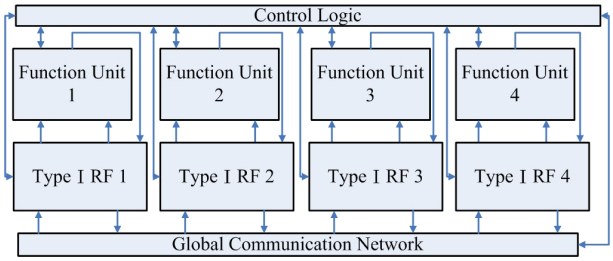


Fig. 6 Type Ⅰ modified clustered register file architecture with **read sharing**

The SIMD register file architecture is shown in Fig. 5. It divides the whole register file into smaller sets of registers and dedicates the subset of registers to a specific group of function units. Therefore, the numbers of read and write ports are reduced from 8R/4W to 2R/1W. From Section 1, we know that power consumption is definitely reduced due to less wire capacitance and silicon area and access time are also reduced due to the reduction of register file cell size. However, this architecture suffers from the inflexibility of accessing every registers when function units require data from other clusters of registers. The overall CPU performance will degrade due to such data dependence in SIMD architecture.

From Table 1, Table 2 and Table 3, we know that SIMD architecture is an efficient implementation of register file from circuit designer's standpoint. From a architectural designer's standpoint, IPC (instruction-per-cycle) loss due to data dependence between clusters are inevitable in SIMD architecture. Therefore, we proposed two modified register file architectures to reduce circuit design difficulties while maintaining IPC. The detailed description of the modified clustered register file will be presented in the next section.

3. MODIFIED CLUSTERED REGISTER FILE

We found that a large portion of silicon area is occupied by access port and metal routing in a conventional central 128 words x 32 bits 8R/4W 4-issue register file. However, the interconnect delay of deep-submicron-meter and nano-meter technologies are higher than intrinsic gate delay. To design a high performance and low power consumption parallel processor with large register file cells and access ports, we must modify the central register file architecture to architectures with less routing complexity.

In order to reduce the routing complexity while maintaining the flexibility to access every registers, we can add extra communication networks to the SIMD register file. Shown in Fig. 6 is the clustered register file with communication network. In this architecture, the function units can read every register file cells from every register file clusters. However, the results generated from the specified group of function units can only write the data back to its own register file cluster. This architecture reduces routing complexity of write ports and provides all the function units virtually a unified register file. A Type

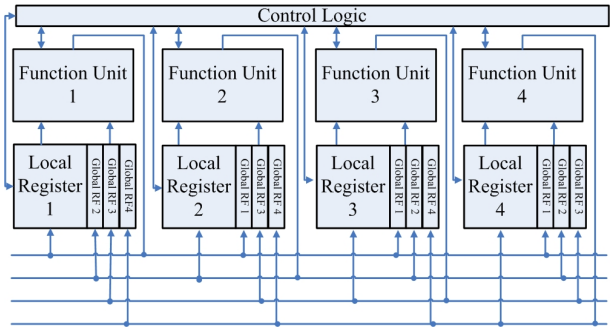


Fig. 7 Type Ⅱ modified clustered register file architecture with **write broadcasting with global registers**

Ⅰ 128 words x 32 bits register file can operates at a clock rate of over 500MHz in 180nm CMOS technology and consume 1.9W. 17% reduction in power delay product and 25% reduction in silicon area compared to centralized register file are achieved.

By clustering the whole register file, register file cells require eight read ports and one write port within the cluster to work with the specified group of function units.

Because the access ports per register file cell are reduced from 8R/4W to 8R/1W, the silicon area is reduced and the active power consumption is also reduced due to the reduction of loading capacitance. Therefore, clustering the whole register file in parallel processor shows great advantages over the centralized counterparts.

Another modified clustered register file is by using write broadcasting with global registers in each cluster. Fig. 7 is the block diagram of Type Ⅱ register file architecture. The whole register file is divided into clusters, which is 4 in this case, to support parallel processing. Each cluster of registers is further divided into local registers and global registers. Local register is used to store and write data from the same cluster of function units. The global registers are used to store the data broadcasted from other clusters of function units. In our example, the first cluster of function units can read all the registers from the first cluster of registers. Its data is written back to the local registers of the first cluster and global registers of the other clusters. Data dependence problems between clusters are solved through the broadcasting of write data and the presence global registers. The circuits of global register cells are the same as the local registers of the same cluster. They can easily be accessed just like local registers except that only a specific cluster of function units can modify the values stored in global registers. A Type Ⅱ 128 words x 32 bits register file can operate at a clock rate of over 500MHz in 180nm CMOS technology and consume 1.7W. 27% reduction in power delay product and 69% reduction in silicon area are achieved compared to centralized register file are achieved.

4. DISCUSSIONS

Four different 128 words x 32 bits 8R/4W register files are implemented in this paper. When the clock rate runs

at 400MHz, 7% active power reduction is shown in Type Ⅰ register file and 19% active power reduction is shown in Type Ⅱ register file.

Conventional central register file suffers from large silicon area and routing complexity when the number of register file cells and number of access ports increase. The need of multiple accesses to the same register file further complicates the register file design in parallel processors. Another register file architecture used in parallel processing is called SIMD. This architecture dedicates a subset of the whole register file to a group of function units. This will simplify the register file design complexity in sacrificing the flexibility of accessing every individual register file cells.Two modified clustered register file are proposed and implemented in this paper. These register files are clustered to reduce the heavy routing complexity and access ports to reduce power consumption, access time and silicon area. In order to reduce the impact of data dependence in SIMD architecture, extra communication networks are adhered to the clustered register file cells.

Type Ⅰ clustered register file uses read sharing which means that al the function units can read data from every register cell via the help of global communication network as shown in Fig. 6 but write data to its own cluster.

Type Ⅱ clustered register file uses write sharing with global registers which means that all the function units can read the data from its own cluster but the data from all function units are broadcasts to all clusters and stores in a specified global register. This type of register file reduces more silicon area and power consumption than Type Ⅰ and is suitable for parallel processing.

5. CONCLUSIONS

By comparing these four register file architecture in 180nm CMOS technology, we observe that the microarchitecture of Type Ⅱ register file has low power advantages over other microarchitectures used in parallel processor.

Type Ⅱ clustered register file with global registers exhibits up to 70% reduction in silicon area, 20% increase in operation frequency, 12% active power consumption reduction and 28% reduction in power delay product at full speed compared to the central register file architecture. It also exhibits up to 19% active power consumption reduction and 28% reduction in power delay product at 400MHz compared to the central register file architecture.

6. ACKNOWLEDGEMENTS

The research is supported by TSMC grant and NSC 92-2220-E-009-011.

7. REFERENCES

[1] Wei Hwang, R. V. Joshi, and W. H. Henkels, “A 500-MHz, 32-wordX64-bit, eight-port self-resetting CMOS register file,” In IEEE journal of solid-state circuits, volume: 34, Issue: 1, Jan. 1999, pp. 56-67

[2] N. S. Kim, and T. Mudge, “ The microarchitecture of a low power register file,” In proceedings of the 2003 international symposium on low power electronics and design, 25-27 Aug. 2003, pp. 384-389

[3] K. M. B. Ahin, P. Patra, and F. N. Najm, “ESTIMA: an architectural-level power estimator for multi-ported pipelined register files,” In proceedings of the 2003 international symposium on low power electronics and design, 25-27 Aug. 2003, pp. 294-299

[4] A. Seznec, E. Toullec, and O. Rochecouste, “Register write specialization register read specialization: a path to complexity-effective wide-issue superscalar processors,” In proceedings of 35th annual IEEE/ACM international symposium on microarchitecture, 18-22 Nov. 2002, pp. 383-394

[5] S. Rixner, W. J. Dally, B. Khailany, P. Mattson, U. J. Kapasi, and J. D. Owens, “Register organization for media processing,” In proceedings of 6th international symposium on high-performance computer architecture, 8-12 Jan. 2000, pp. 375-386

[6] J. Zalamea, J. Llosa, E. Avguade, and M. Valero, “ Hierarchical clustered register file organization for VLIW processors,” In proceedings of international parallel and distributed processing symposium 2003, 22-26 April 2003, 10 pp.

@Full Speed	Normalized Area	Max. Delay (ns)	Active Power (W)	Leak-age Power (uW)	PDP (nJ)
Central	1.000	2.17	1.958	242	4.249
SIMD	0.291	1.83	1.686	60	3.085
Type Ⅰ	0.749	1.86	1.902	196	3.538
Type Ⅱ	0.306	1.81	1.707	68	3.089

Table 1 Performance comparison of different register file architecture operating at full speed

Normalized @Full Speed	Area	Max. Freq.	Active Power	Leak-age power	PDP
Central	100%	100%	100%	100%	100%
SIMD	29%	119%	86%	25%	73%
Type Ⅰ	75%	117%	97%	81%	83%
Type Ⅱ	31%	120%	87%	28%	73%

Table 2 Normalized performance comparison of different register file architecture operating at full speed

Normalized @ 400MHz	Area	Max. Delay (ns)	Active Power (W)	Active Power (W)
Central	1.000	2.17	1.5259	100%
SIMD	0.291	1.83	1.2339	81%
Type Ⅰ	0.749	1.86	1.4152	93%
Type Ⅱ	0.306	1.81	1.2357	81%

Table 3 Performance comparison of different register file architecture operating at 400MHz