Adaptive Voltage Scaling System Based on Indirect Timing Monitoring

Hao Wang[1,2], Sheng Liu[2,*], Yulong Qiao[1]

[1]School of Information and Communication Engineering, Harbin Engineering University, Harbin, China
[2]School of Computer Science, National University of Defense Technology, Changsha, China

Abstract

With the rapid development of semiconductor manufacturing technology, the number of transistors integrated in the same area has increased by an order of magnitude. This has led to higher power consumption per unit area of the chip, which in turn imposes additional requirements for low-power design technology. In order to reduce the power consumption of the chip, this paper constructs an adjustable delay chain that indirectly monitors the timing information of the chip's critical path and regulates the scaling of the power supply voltage, thereby reducing the chip's power consumption. In addition, the AXI-based adaptive voltage scaling bus (AVSBus) is implemented to achieve real-time communication between the chip system and the power system. Experimental simulation shows that the chip's operating frequency changes from 1.3GHz to 800MHz at 12nm and 125℃, saving power consumption by 10.6~53.4% compared with fixed voltage.

Keywords

Adaptive voltage scaling, Low-power design, Timing monitoring, AVSBus

1. Background and current situation of AVS research

After the integrated circuit entered the system on chip era, semiconductor manufacturing technology has developed rapidly in a manner that has not only improved chip integration, but also led to a sharp increase in chip power consumption. The latter has gradually become a limiting factor in the development of semiconductor devices. At the same time, power consumption has become an important indicator for use in measuring the performance of a chip. Therefore, researchers who focus on integrated circuits have conducted extensive research into methods that reduce power consumption under the premise of ensuring the chip's basic performance. Among them, adaptive voltage scaling (AVS) technology, a low-power method that can monitor the timing of the chip system and regulate the power supply voltage in real time, has become a research hotspot due to its remarkable ability to reduce chip power consumption.

In 2003, ARM and the National Semiconductor (NS) corporation of the United States jointly proposed AVS technology in [1]. Under AVS, the timing information of the chip system is detected in real time through the hardware performance monitor and, transmitted to the power control unit for judgment; the voltage regulation information is then output, after which the energy management unit is used to regulate the chip power supply voltage. In the same year, the classical Razor structure was proposed in [2]. This structure adds a shadow latch on the basis of the main trigger at the end of the critical path, and obtains the timing alarm information (via XOR) of the output results of the two. Several works [3–5] have conducted structural optimization on the basis of Razor; however, these works are still based on the principle of obtaining critical path timing information by means of the double sampling method. In [6], timing information is obtained by inserting timing monitoring units in the middle of the critical path, which effectively reduces the number of monitoring units. The

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control
EMAIL: *liusheng83@nudt.edu.cn (Sheng Liu)

method of indirectly monitoring timing information by copying the critical path is introduced in [7–10].

2. AVS principle

Power consumption in the chip can be broadly divided into dynamic power consumption and static power consumption. In turn, the former can be subdivided into switching power consumption and short-circuit power consumption. Here, switching power consumption is the power consumed by charging and discharging the load capacitor when the circuit signal is overturned, while short-circuit power consumption is generated by the short-circuit current due to the intermediate level of the input signal during the rising or falling period, which makes the NMOS tube and PMOS tube in the logic gate circuit conduct at the same time. The switching power consumption can be obtained by integrating the instantaneous power consumption in the corresponding period with equation (1) in [11].

$$P_{switch} = f \cdot N_{sw} \cdot \int_0^T p(t)dt = \frac{1}{2} C \cdot V_{DD}^2 \cdot f \cdot N_{sw} \tag{1}$$

Here, N_{sw} is the number of transistors flipped in a single clock cycle, f is the working clock frequency of the system, C is the load capacitance, and V_{DD} is the supply voltage. The static power consumption occurs primarily due to current leakage in the transistor, which is mainly related to the chip manufacturing process and cannot be effectively reduced in the chip design stage. The main component of dynamic power consumption is switching power consumption, which is proportional to the square of the supply voltage and the operating frequency according to formula (1); thus, the chip power consumption can be effectively reduced by reducing the voltage.

There is a correlation between chip power supply voltage and path timing. A higher power supply voltage can improve the charging and discharging speed of the load capacitor, thus reducing the delay of the logic gate circuit. Equation (2) in [12] reflects the relationship between CMOS logic gate delay and power supply voltage.

$$t_{gate} \propto \frac{V_{DD}}{\beta (V_{DD} - V_T)^{\alpha}} \tag{2}$$

Here, V_{DD} is the supply voltage, V_T is the effective threshold voltage, and α and β are the fitting parameters of the actual delay of the logic gate. In the design phase, the chip reserves a certain amount of voltage slack after considering the impact of process, voltage, and temperature (PVT) along with the actual working situation, which result in a certain degree of power wastage when the chip is operating. AVS technology can adjust the power supply voltage to the lowest possible level to ensure the normal operation of the chip by monitoring the chip's timing information in real time, which resolves whether to scale or maintain the voltage. This approach effectively compresses the voltage slack reserved at the design stage and reduces the power consumption of the chip.

3. AVS system structure design

3.1. Overall structure of AVS

As shown in Figure 1, the AVS system designed in this paper is mainly composed of a delay chain, timing monitoring unit, AVS control unit, and AVSBus. Among them, the delay chain tracks the timing change of the critical path in the chip system through concatenate adjustable-number inverters. The timing monitoring unit obtains the timing information by using multiple triggers to sample the status of signal flipping in the delay chain. The AVS control unit converts the sampling results into a signal to increase, decrease, or maintain voltage. The voltage regulation signal is then transmitted to the Point of Load (POL) power supply control unit through the AVSBus to regulate the supply voltage of the digital load.

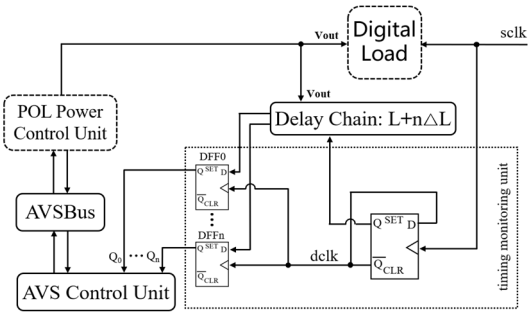


Figure 1 Overall structure diagram of AVS system

3.2. Construction of delay chain

The most important element of the AVS system is obtaining the timing information of the chip's critical path in real time. The length of the critical path in the chip system is often affected by the PVT and other conditions. This requires that the delay chain constructed in the indirect monitoring method should have a good ability to track the timing changes of the critical path, to ensure that the timing information obtained by monitoring the delay chain can accurately reflect the timing changes of the critical path.

In this paper, the 64-bit multiplication and accumulation operation unit (MAC) in a DSP chip is used as the digital load, and the delay chain is constructed through concatenate inverters to indirectly monitor the timing change of the critical path in the digital load. First, we conduct static timing analysis on the digital load and delay chain respectively on the basis of layout generation. After determining the length of the critical path in the digital load, the delay chain (L) length (number of inverters) of the copied critical path in Figure 2 is determined under the constraints of the following two conditions:

(1) The delay of the inverter delay chain shall not be less than the delay of the critical path in the digital load;

(2) The number of inverters in the delay chain shall be even.

In more detail, the first condition is that when the delay of the inverter delay chain is less than the critical path delay, the timing monitoring unit detects that the delay chain has a larger timing slack than the critical path, this causes the AVS control unit to generate an incorrect voltage reduction signal, making the supply voltage lower than that required for normal operation of the digital load, which can in turn cause serious errors such as chip timing violations. The second condition is implemented to ensure that the signal flipping direction of the trigger sampling position in the timing monitoring unit is consistent with that of the delay chain input.

In Figure 2, in order to monitor the timing slack of the digital load critical path in real time, some adjustment chains (ΔL) are added on the basis of the delay chain (L) for copying the critical path. In this paper, the scaling step value of the voltage is 50mv. However, because the digital load has different timing changes in the critical path (caused by increasing or decreasing a voltage step value under different voltage levels), a dynamic simulation is carried out for the digital load to obtain the maximum value of the timing change in the critical path caused by scaling a voltage step value under different voltage levels. Moreover, the maximum value of the timing change is mapped to the delay chain (ΔL) reflected by the step voltage value in Figure 2, which also needs to meet the condition that the number of inverters shall be even.

3.3. Timing monitoring unit

The timing monitoring unit developed in this paper consists of multiple triggers, which monitor the timing information of the inverter delay chain in real time. As shown in Figure 2, one of the triggers divides the system clock (sclk) frequency by two to generate a monitoring pulse signal, and the inverse phase output of the divider is used as the monitoring clock (dclk) for the timing monitoring unit. Other triggers sample the signal overturning at a specific position in the delay chain, then output the

sampling results to the AVS control unit, which outputs a signal to adjust the voltage according to the sampling results. The number of sampling triggers depends on the regulation range of the power supply voltage. In this paper, the digital load operates between 0.7V and 1.2V, and the voltage scaling step value is 50mv. Therefore, there are a total of 10 delay regulation chains (ΔL). In addition, there is a delay chain (L) for replicating the critical path derived from the static timing simulation. The output signal flipping status for all of these 11 inverter delay chains needs to be sampled. Therefore, the timing monitoring unit is composed of 12 triggers, 11 of which are used to sample the signal flipping status in the delay path.

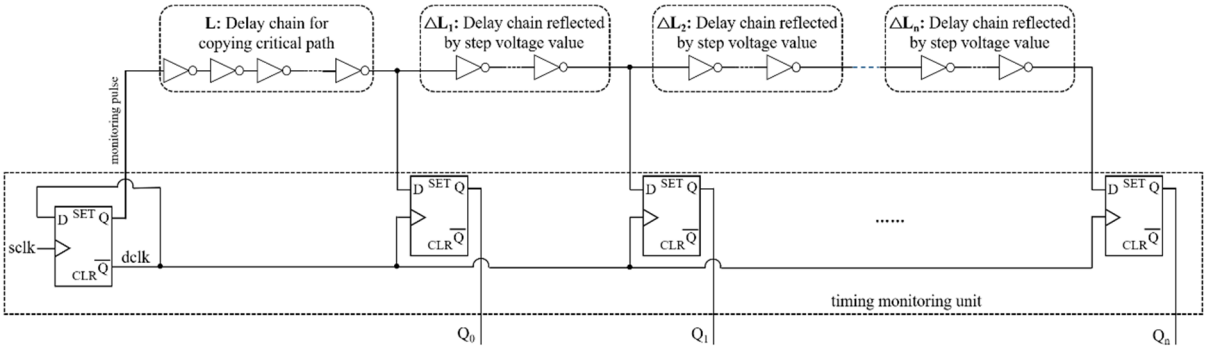


Figure 2 Structure diagram of inverter delay chain and timing monitoring unit

The output timing relationship of the trigger in the timing monitoring unit is illustrated in Figure 3. When the rising edge of the monitoring clock arrives, the output signal of the K-segment adjustment chain did not flip, while the output of the K+1 and all subsequent adjustment chains has flipped, indicating that there is timing slack in the critical path of the digital load. Moreover, this timing slack is the delay from the first to the K-th segment adjustment chain. The value of the K indicates the timing slack of the critical path in the digital load: the larger the K value, the greater the timing slack of the critical path in the digital load.

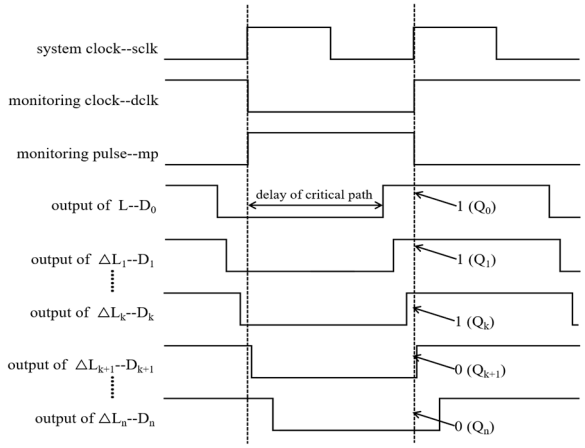


Figure 3 Timing relation of timing monitoring unit output

The timing monitoring unit implemented in this paper maps the K value to the number of 1 in an11-bit binary number through trigger sampling. The AVS control unit will then output a signal to scale or maintain voltage according to the number of 1. First, the highest bit in this 11-bit binary number must be 1, because it is the sampling output of the delay chain that replicates the critical path; otherwise, timing violations will occur. Then, the number of 1 in this 11-bit binary number (we use N to represent this number) can be divided into three cases. The first is when N is less than 2, which indicates that the critical path of the digital load is time-tight and the power supply voltage needs to be increased. The second is when N is greater than 2, which means the timing of the critical path of the digital load is loose, and the AVS control unit will output the lower voltage signal. Moreover, when N is equal to 2, the AVS control unit will output a voltage hold signal.

3.4. AVSBus and its implementation

As an interface bus protocol, AVSBus is used to implement point-to-point communication between ASIC, FPGA or other logical storage, processor devices and POL power control devices on the system to achieve adaptive voltage scaling of the circuit system. In March 2014, AVSBus was released as the third part of the Power Management Bus (PMBus) version 1.3, an extension of the System Management Bus (SMBus), which was developed and maintained by the System Management Interface Forum (SMIF) in 2005.

The AVSBus implemented in this paper is a three-wire communication link. As Figure 4 shows, the three links are AVS_Clock, AVS_MData and AVS_SData. Here, AVS_MData is driven by the master device and sends data to the slave device, AVS_SData is driven by the slave device and sends data (response) to the master device, and AVS_ Clock is driven by the master device and provides AVS_MData and AVS_SData with the clock.

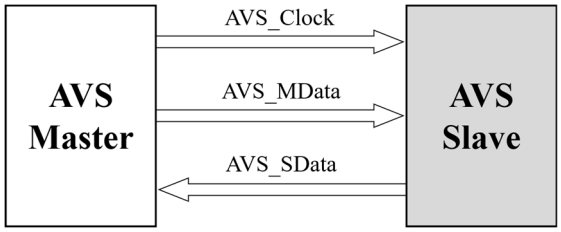
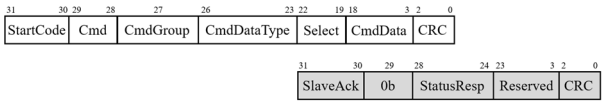


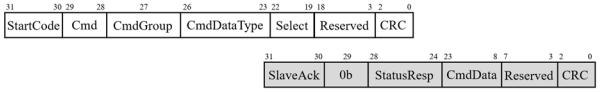
Figure 4 Communication mode of AVSBus

As shown in Figure 5, the AVSBus communication protocol consists of two frames, namely a write frame and read frame; the bit width of each frame is 64 bits. Each frame is in turn composed of two subframes: the main subframe is sent by AVS_MData, and the slave subframe is sent by AVS_SData. All subframes also support 3-bit CRC verification. The sender uses the CRC generation polynomial of formula (3) for the first 29 bits of the subframe, and the receiver obtains the entire subframe including the 3-bit CRC, using the same polynomial to verify the CRC to confirm the completeness of the data.

$$CRC(x) = x^3 + x^2 + x \qquad (3)$$



(a) The structure of write frame



(b) The structure of read frame

Figure 5 Structure of AVSBus read/write frames

The meaning of each field in the read/write frame is shown in Table 1. When the AVS control unit outputs a signal to decrease, increase, or maintain the voltage, the <CmdDataType> is 0110, 0111, and 1000 respectively in the AVSBus main subframe of the write frame.

Table 1 Meaning of AVSBus read/write frame field

Field Name	Width	Meanings
<StartCode>	2	01b as startup code
<Cmd>	2	Command code to determine the device needs: 11b: Read data 10b: Reserved 01b: Write data and keep it 00b: Write data and submit all pending voltage writes
<CmdGroup>	1	Qualifiers that distinguish between two sets of data types: 0b: AVSBus data type for full definition 1b: manufacturer-specific data types
<CmdDataType>	4	Data types applicable to <CmdData>: 0000b: target supply rail voltage 0001b: target power rail V_{out} conversion efficiency 0010b: power rail current (read-only) 0011b: power rail temperature (read-only) 0100b: reset the power rail voltage to the default value (write-only) 0101b: power supply mode for power rail 0110b: 50mv down command issued by AVS 0111b: 50mv up command issued by AVS 1000b: maintain voltage command from AVS 1001b~1101b: reserved command data type 1110b: AVSBus status 111lb: AVSBus version
<Select>	4	Selector field, used to distinguish instances of command data types on the device
<CmdData>	16	Transmitted data
<CRC>	3	CRC check bits
<Reserved_N>	N	Keep some bits to be developed. Reserved bits all 1 sent
<SlaveACK>	2	Response from slave device
<StatusResp>	5	State response from slave device

The AVSBus slave device will respond to the <SlaveAck> in the slave subframe to indicate that the specific operation requested by the AVSBus master device has not been executed due to a CRC check error, and that the AVSBus master device command that exceeds the voltage write execution range has also not been executed. In addition, when the AVS_MData maintains a high level, AVSBus will resynchronize its communication interface after the slave device receives 34 clock pulses continuously, after which it waits for the next <StartCode> to start another communication. This resynchronization mechanism enables AVSBus to avoid error states caused by line noise and other human factors.

4. Simulation and analysis of AVS system performance

4.1. Simulation of AVS system performance

Using the C language Application Program Interface (API) of the HSIM simulation software proposed in [13] to simulate and model the off-chip voltage regulation module, this paper achieves the simulation and evaluation of the voltage scaling function of the AVS system. On this basis, a hybrid simulation platform based on VCS-HSIM is built to simulate the performance of the AVS system in the 64-bit MAC unit of a DSP under 12 nm process, SS process angle and 125°C temperature conditions. As shown in Figure 6, compared with a fixed voltage of 1.2V, the AVS system adjusts the power supply voltage to the lowest voltage, which guarantees that the digital circuit system will work properly at different operating frequencies, resulting in a voltage drop of 4.2~41.7% and a power loss of 10.6~53.4%.

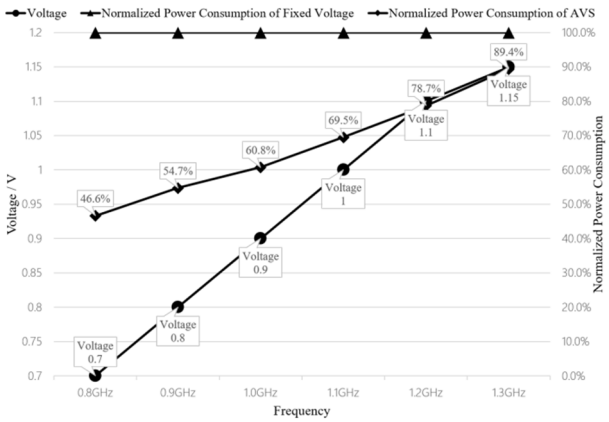


Figure 6 Performance of AVS system in terms of voltage and power consumption reduction

4.2. Analysis of simulation results

The AVS system built in this paper can obtain the timing slack of the critical path in the digital load in real time, quantify it as the number of inverters in the delay chain, and adaptively scale the power supply voltage according to the actual timing status of the digital load, thereby achieving the effect of reducing power consumption. Because the static power consumption produced by leakage current in a lower process (12nm in this paper) is larger than that in a higher process (55nm in [14]), it can be seen from Table 2 that the AVS system designed in this paper produces a somewhat smaller reduction in chip power consumption (primarily dynamic power consumption), 53.4% in this paper vs 64.7% in [14]. However, compared with the fixed power supply voltage, this is still a significant chip power consumption reduction effect over a wide range of voltages and high frequencies. In addition, the AVS system designed in this paper has less area cost and reaches competitive performance.

Table 2 Performance of the AVS system designed in this paper compared with published literature

	This Paper	Reference 7	Reference 9	Reference 14
Process/nm	12	0.18μm	22	55
Frequency/MHz	800~1300	3~123	250~800	300~700
Voltage/V	0.7~1.2	0.9~1.6	0.4~0.7	1.0~1.5
Energy Saving	53.4%	40%	33%	64.7%
Area/mm²	0.013	not report	3.38 (core and monitor)	0.016
Architecture	64 bits MAC	a MPU with 64Mb DRAM	a graphics execution core	FIR

5. Conclusion

The AVS system designed in this paper simulates the timing changes of the critical path in digital circuit systems by constructing an adjustable-length delay chain with inverters, using triggers to build the timing monitoring unit in order to detect the signal flipping of the inverter delay chain. This indirectly allows for the timing slack information of the actual critical path to be obtained. The AVS system will adjust the power supply voltage to the minimum required to support the normal operation of the digital circuit system, according to the timing slack. When the working frequency of the MAC unit decreases from 1.3GHz to 800MHz under 12 nm process, SS process corner, and 125℃ conditions, the AVS system achieves energy savings of 10.6~53.4% compared to the fixed 1.2V supply voltage. In addition, the AVS system proposed in this paper has completed point-to-point communication with the power system using AVSBus, which can be applied to the digital power management of digital circuit systems. In the future, the AVS technology can be integrated in Chiplet technology to reduce chip power consumption more flexibly.

6. Acknowledgement

This work was supported by Heterogeneous Multi-Core Digital Signal Processor project under agreement number 2009ZYHJ0007.

7. References

[1] Maksimovic D, Dhar S, Ambatipudi R, et al. Adaptive voltage scaling power supply for use in a digital processing component and method of operating the same: U.S. Patent 6,548,991[P]. 2003-4-15.

[2] Ernst D, Kim N S, Das S, et al. Razor: A low-power pipeline based on circuit-level timing speculation[C]//Proceedings. 36th Annual IEEE/ACM International Symposium on Microarchitecture, 2003. MICRO-36. IEEE Computer Society, 2003: 7-7.

[3] Calhoun B H, Chandrakasan A P. Standby power reduction using dynamic voltage scaling and canary flip-flop structures[J]. IEEE Journal of Solid-State Circuits, 2004, 39(9): 1504-1511.

[4] Fojtik M, Fick D, Kim Y, et al. Bubble razor: Eliminating timing margins in an ARM cortex-M3 processor in 45 nm CMOS using architecturally independent error detection and correction[J]. IEEE Journal of Solid-State Circuits, 2012, 48(1): 66-81.

[5] Reyserhove H, Dehaene W. Margin elimination through timing error detection in a near-threshold enabled 32-bit microcontroller in 40-nm CMOS[J]. IEEE Journal of Solid-State Circuits, 2018, 53(7): 2101-2113.

[6] Lai L, Chandra V, Aitken R, et al. Slackprobe: A low overhead in situ on-line timing slack monitoring methodology[C]//2013 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2013: 282-287.

[7] Nakai M, Akui S, Seno K, et al. Dynamic voltage and frequency management for a low-power embedded microprocessor[J]. IEEE journal of solid-state Circuits, 2005, 40(1): 28-35.

[8] Wilson R, Beigne E, Flatresse P, et al. A 460mhz at 397mv, 2.6 ghz at 1.3 v, 32b vliw dsp, embedding f max tracking[C]//2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014: 452-453.

[9] Cho M, Kim S T, Tokunaga C, et al. Postsilicon voltage guard-band reduction in a 22 nm graphics execution core using adaptive voltage scaling and dynamic power gating[J]. IEEE Journal of Solid-State Circuits, 2016, 52(1): 50-63.

[10] Gomez R G, Bano E, Cathelin A, et al. A Performance-Flexible Energy-Optimized Automotive-Grade Cortex-R4F SoC through Combined AVS/ABB/Bias-in-Memory-Array Closed-Loop Regulation in 28nm FD-SOI[C]//2020 IEEE Symposium on VLSI Circuits. IEEE, 2020: 1-2.

[11] Wei Guo, SoC design method and implementation [M]. Electronic Industry Press, 2017.

[12] Sakurai T, Newton A R. Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas[J]. IEEE Journal of solid-state circuits, 1990, 25(2): 584-594.

[13] Zhaoxuan Tian. Design and implementation of low power consumption SOC chip based on adaptive voltage regulation[D]. Southeast University, 2014.

[14] Shengnan Lin, Liping Liang. Adaptive voltage regulating system based on performance matching for SoC [J]. Electronic Design Engineering, 2022, 30 (6): 6.