Pipeline Stage Unification: A Low-Energy Consumption Technique for Future Mobile Processors

Hajime Shimada, Hideki Ando, Toshio Shimada
Department of Information Electronics, Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi-ken, Japan. 464-8603
{hajime,ando,shimada}@shimada.nuee.nagoya-u.ac.jp

ABSTRACT

Recent mobile processors are required to exhibit both low-energy consumption and high performance. To satisfy these requirements, dynamic voltage scaling (DVS) is currently employed. However, its effectiveness will be limited in the future because of shrinking the variable supply voltage range. As an alternative, we previously proposed pipeline stage unification (PSU), which unifies multiple pipeline stages without reducing the supply voltage at a power-saving mode. This paper compares effectiveness of PSU to DVS in current and future process generations. Our evaluation results show PSU will reduce energy consumption by 27-34% more than DVS after about 10 years.

Categories and Subject Descriptors: C.1.3 [Computer Systems Organization]: Processor Architectures, Other Architecture Styles

General Terms: Design, Performance

Keywords: low-power consumption, future process technology, dynamic voltage scaling, pipeline stage

1. INTRODUCTION

Recent mobile processors are required to exhibit low-energy consumption as well as high performance. To satisfy these requirements, a method called *dynamic voltage scaling* or DVS is currently employed. If the current workload is light, DVS decreases the clock frequency to reduce power consumption. In order to adjust the signal delay to the lengthened clock cycle time, DVS reduces the supply voltage. This saves energy consumption used for program execution.

Although DVS is currently an effective method for reducing energy consumption, this effectiveness will diminish as semiconductor technology advances. The reason for this is as follows. Firstly, if the threshold voltage is significantly reduced from its current level, a dramatic increase in subthreshold leakage will occur. Therefore, scaling of the threshold voltage will slow as technology advances. Since the minimum supply voltage for transistors to work properly is bounded by the threshold voltage, the variable range of the supply voltage (more precisely, the ratio of the variable range to the maximum supply voltage) in DVS will reduce, and thus the effectiveness of DVS will decrease. Secondly, DVS degrades the reliability of a processor due to the increase of transient faults probability. In future technology, transient faults will become a serious problem due to lower supply voltages. Since DVS reduces the supply voltage at the power-saving mode, it makes the problem more serious.

As an alternative, we presented a method called *pipeline stage unification* or PSU [11]. PSU dynamically scales the clock frequency to reduce energy consumption as with DVS, but unlike DVS, it unifies multiple pipeline stages by bypassing pipeline registers, instead of scaling down the supply voltage. PSU saves energy consumption in two ways. Firstly, PSU saves power consumption by reducing the total load capacitance of the clock driver. This is accomplished by stopping the clock signal to bypassed pipeline registers. Secondly, PSU reduces the clock cycle count of program execution by reducing the number of pipeline stages. Our evaluation results show that PSU is moderately more effective than DVS in the current process technology. However, more importantly, the effectiveness of DVS will significantly decrease as process generations advance, while the effectiveness of PSU will remain constant. As a result, PSU will become a more effective alternative to DVS for energy saving.

2. IMPLEMENTATION OF PIPELINE STAGE UNIFICATION

This section describes the implementation of PSU. For convenience, we will describe the implementation of two-stage unification as an example.

We prepare three signals, called *full-time clock*, *part-time clock*, and *unification signal*. The full-time clock is a clock signal which is always active regardless of unification, while the part-time clock is a clock signal which is deactivated when pipeline stages are unified; it is active when they are not unified. The unification signal indicates pipeline stage unification. Since the pipeline register among two adjacent combination logic circuit is inactive or bypassed, the two logic circuits operate together as a single stage. Note that the pipeline registers do not include decoupling memory elements (e.g. the instruction window) among the front end, the execution core, and the back end.

There are two ways to bypass a pipeline register. One way is to organize the pipeline register logic so that a signal can pass through it regardless of the clock signal when PSU is enabled. We can easily implement this if a transparent latch

ISLPED'03, August 25–27, 2003, Seoul, Korea.

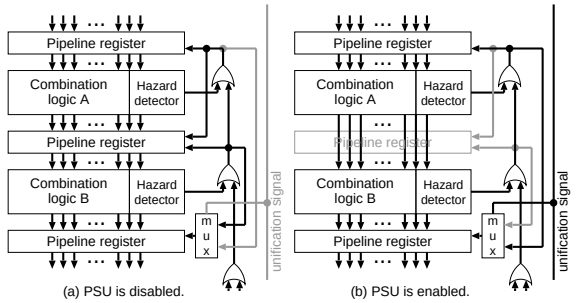

Figure 1: Pipeline interlock circuit.

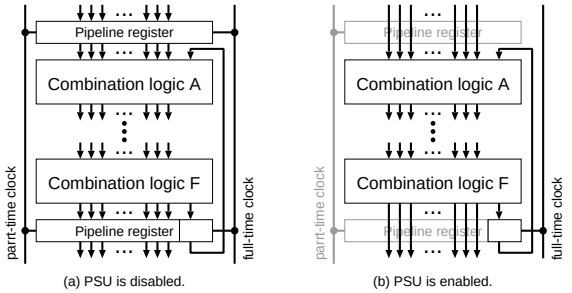

Figure 2: Looped signal path.

is used for the pipeline register. Another way is to place a multiplexer after the pipeline registers, which selects from the outputs of the pipeline register and the previous stage according to the unification signal. This method can be applied to any circuit for the pipeline register.

A pipeline interlock circuit must also be modified. Fig. 1 shows the modified interlock circuit. In general, a pipeline stage must stall if a hazard is detected at its own stage or its succeeding stage will stall. The pipeline register before the combination logic A is not necessary to be changed for PSU, but the pipeline register after the combination logic B must select stall signal for logic A, because hazard in logic A must be detected after unification. This requires a multiplexer.

Note that a pipeline register between unified stages is basically bypassed, however this is not always the case. If the signal is forming part of a looped signal path, we cannot bypass the pipeline register. Fig. 2 shows an example. Fig. 2(a) shows a pipeline with a looped signal path. When stage F and its following stage are unified, the pipeline register between these stages is basically bypassed, however the part which the looped signal goes through is not bypassed as shown in Fig. 2(b). By not bypassing the pipeline register, the looped signal path can maintain correct timing.

3. ENERGY REDUCTION

In general, energy E consumed for program execution can be expressed as follows:

$$E = P \times T_{ex} \tag{1}$$

where P is power consumption and T_{ex} is execution time. P and T_{ex} are given by:

$$P = f \times C \times {V_{DD}}^2 \tag{2}$$

$$T_{ex} = \frac{N}{IPC \times f} \tag{3}$$

where f is the clock frequency, C is the average capacitance of switching nodes, N is the number of executed instructions, and IPC is the average number of instructions per clock cycle.

3.1 Energy Reduction with DVS

Substituting eq. (2) and (3) into (1), energy consumption $E(f, V_{DD})$ is derived as follows:

$$E(f, V_{DD}) = \frac{N \times C \times {V_{DD}}^2}{IPC} \tag{4}$$

Now consider a DVS processor which runs at the maximum clock frequency f_{max} and the maximum supply voltage V_{DDmax} in the normal mode, and runs at f_{low} and V_{DDlow} in the power-saving mode. Energy consumption in the power-saving mode normalized by that in the normal mode is expressed as follows:

$$E_{DVS,n}(f_{low}, V_{DDlow}) = \frac{IPC_{max}}{IPC_{low}} \times \left(\frac{V_{DDlow}}{V_{DDmax}}\right)^2 \tag{5}$$

where IPC_{max} and IPC_{low} are IPCs in the normal and power-saving modes, respectively. Under a simple assumption that devices other than the processor (including memory) also decrease their speed in proportion to the processor clock frequency, IPC_{low} is identical to IPC_{max}. Thus, the following equation is derived:

$$E_{DVS,n}(f_{low}, V_{DDlow}) = \left(\frac{V_{DDlow}}{V_{DDmax}}\right)^2 \tag{6}$$

We find that the reduction of energy consumption is achieved only by reduction of the supply voltage, and it is a quadratic function of the supply voltage ratio. Being quadratic, this leads to a dramatic reduction. At the same time, eq. (6) implies that the effectiveness of DVS will rapidly diminish in future technology, where V_{DDlow}/V_{DDmax} cannot be as small as it is in current technology.

3.2 Energy Reduction with PSU

PSU saves power consumption by stopping the part-time clock. When a processor runs with U-stage unification or *unification-degree* U, power consumed by the clock drivers can be reduced by $1/U$, ideally. Also, as in normal processors, total power consumption is reduced by the clock frequency reduction rate. Thus, power consumption of a PSU processor which runs at clock frequency f_{low} with unification-degree U is expressed as follows:

$$P_{PSU}(f_{low}, U) = \left(P_{total} - P_{clock} + \frac{P_{clock}}{U}\right) \times \frac{f_{low}}{f_{max}} \tag{7}$$

where P_{total} and P_{clock} are the total power consumption of the processor and the power consumption of the clock drivers in the normal mode, respectively. Using eq. (1), energy consumption normalized by that in the normal mode is expressed as follows:

$$E_{PSU,n}(f_{low}, U) = \frac{P_{PSU}(f_{low}, U) \times T_{ex}(f_{low}, U)}{P_{total} \times T_{ex}(f_{max}, 1)} \tag{8}$$

where $T_{ex}(f, U)$ is execution time with clock frequency f and unification-degree U. Note that $T_{ex}(f_{max}, 1)$ is execution time in the normal mode. Substituting eq. (3) and (7) into (8), we obtain the following equation:

$$E_{PSU,n}(f_{low}, U) = \frac{IPC_{max}}{IPC_{low}} \times \left\{1 - k \times \left(1 - \frac{1}{U}\right)\right\} \tag{9}$$

where k is equal to P_{clock}/P_{total}.

As found from (9), energy consumption is reduced in inverse proportion to the ratio of IPC improvement (note that $IPC_{max} < IPC_{low}$ because of the shortened pipeline). Also, the reduction depends on how much the power consumed by the clock driver contributes to the total power consumption. Since this rate is significant in current high-speed processors

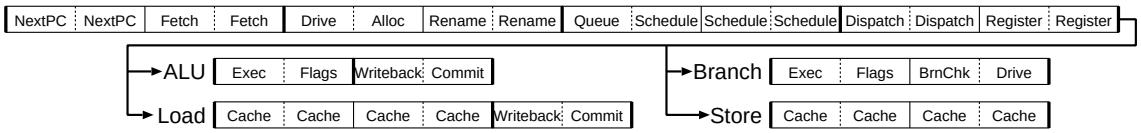

Figure 3: Assumed PSU pipeline.

Table 1: Processor configuration.

processor	8-way out-of-order issue, 128-entry RUU, 64-entry LSQ, 8 int ALU, 4 int mult/div, 8 fp ALU, 4 fp mult/div, 8 memory ports
branch prediction	8K-entry gshare, 6-bit history, 2K-entry BTB, 16-entry RAS
L1 I- and D-cache	64KB/32B line/direct map
L2 unified cache	2MB/64B line/4-way
memory	64 cycles first hit, 2 cycles burst interval
TLB	16-entry I-TLB, 32-entry D-TLB 128 cycles miss latency

Table 2: Assumptions of latencies and penalty.

# of stages unified	1	2	4
clock frequency rate f	100%	50%	25%
int Mult latency	3	2	1
fp ALU latency	2	1	1
fp Mult latency	4	2	1
branch misprediction penalty	20	10	5
L1 cache hit latency	4	2	1
L2 cache hit latency	16	8	4

and this trend will continue toward the future (for a deep pipeline and small clock skew, etc), we expect that PSU can considerably reduce energy consumption in the future.

4. EXPERIMENTAL ASSUMPTIONS

We measure IPC by varying the number of pipeline stages using an out-of-order execution simulator in the SimpleScalar tool set [3]. We use eight benchmark programs from SPECint95. Table 1 lists the processor configuration. We assume a deep pipeline similar to current processors. Fig. 3 shows the base, two-stage unified, and the four-stage unified pipeline. When two stages are unified, the pipeline registers represented by dotted lines are bypassed. When four stages are unified, the pipeline registers represented by thin lines are bypassed. We also assume that the memory degrades its speed in proportion to the processor clock frequency. Thus, memory access cycle count is constant independent of change of the processor clock frequency. Table 2 summarizes the instruction execution latencies, the branch misprediction penalty, and the cache hit latencies in these pipelines.

The power consumption rate of the clock driver k depends on the processor design. According to previous papers [2, 5, 7, 8], it ranges from 18% to 40%. Unless explicitly specified, we assume it to be 30% in our evaluation given in Section 5, which is an approximate median of these values. Also, we assume that the power consumption of the clock driver is proportional to the number of driving pipeline registers. Or more simply, we assume it is inversely proportional to the unification degree U. This assumption is rough, but we believe it is reasonable in our evaluation for the following reason. In general, a clock signal is delivered by a network. The power of the clock drivers is mostly consumed by the final-stage driver (for example, the final-stage driver consumes 88% of the total power consumed by the hierarchical clock network in the Intel Itanium 2 [2].). Also, the load capacitance of the final-stage driver is roughly proportional to its fanout or the number of pipeline registers.

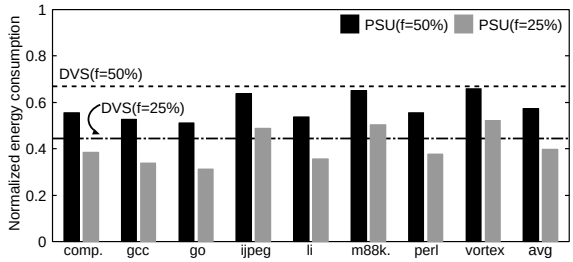

Figure 4: Normalized energy consumption for 50% and 25% frequencies in the 180nm technology.

5. EXPERIMENTAL RESULTS

5.1 Energy Consumption in the Current

In this subsection, we evaluate the energy consumption of PSU and DVS, and compare them. We assumed the supply voltages of DVS are 1.65V at the f=100%, 1.35V at the f=30%, and 1.10V at the f=25%. In contrast, we assumed supply voltage of PSU is constant value 1.65V in any frequencies. The assumptions of DVS are derived from corresponding data of Transmeta Crusoe TM5400 [9].

Fig. 4 compares the energy consumption of PSU to that of DVS. The vertical axis indicates energy consumption normalized by that in the normal mode. As can be clearly seen, in the case of f=50%, the energy consumption in PSU is smaller than that in DVS for any benchmark. PSU can reduce energy consumption by 14% more than DVS on average. In the case of f=25%, the advantages of PSU over DVS become small. In a few of the benchmarks (ijpeg, m88ksim, and vortex), PSU even consumes more energy than DVS. However, PSU can reduce energy consumption by 11% more than DVS on average.

5.2 Energy Consumption in the Future

In order to estimate the reduction of energy consumption with DVS in the future, we must investigate how supply voltages will change as process technology advances, and to what extent the supply voltage can be reduced depending on the lowered clock frequency. Regarding the supply voltage V_{DDmax} in the normal mode where the processor runs at the maximum frequency, we estimated it by referring to the trend [4] and data announced by TSMC and UMC [1, 6] with an approximate formula. We derived the following equation:

$$V_{DDmax} = 0.0381 \times Tech^{0.7171} \tag{10}$$

where $Tech$ [nm] is the feature size of the process technology.

On the other hand, regarding the supply voltages in the power-saving mode, there is not enough data available for estimation in the way we have done for V_{DDmax} above. However, we have derived it in the following way. In general, if the supply voltage V_{DD} and the threshold voltage V_{th} are given, the following relationship between the maximum clock frequency of gate operation f and those voltages exist.

$$f \propto \frac{(V_{DD} - V_{th})^2}{V_{DD}} \tag{11}$$

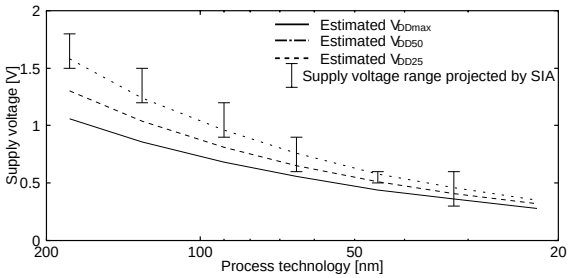

Figure 5: Estimated supply voltages in the normal and power-saving modes in future DVS processors.

Let f be the clock frequency in the power-saving mode normalized by that in the normal mode, and let a be a proportionality constant. The above equation can then be expressed as follows:

$$f = a \times \frac{(V_{DD} - V_{th})^2}{V_{DD}} \quad (12)$$

The constant a can be derived from the supply voltage V_{DDmax} (eq. (10)) at $f = 1$ and V_{th} as follows:

$$a = \frac{V_{DDmax}}{(V_{DDmax} - V_{th})^2} \quad (13)$$

We estimated V_{th} using the trends and data from [1, 4, 6]. The following equation was derived:

$$V_{th} = 0.0226 \times Tech^{0.5111} \quad (14)$$

Having obtained constant a, we can derive V_{DD} in the power-saving mode from eq. (12). Note that M is added to eq.. In practice, it is necessary to associate a certain margin with the value of V_{DD}. The margin should be larger for various reasons as V_{DD} becomes lower. For example, eq. (11) implies that the deviation of V_{th} caused by deviations in the LSI process affects the maximum clock frequency more as V_{DD} becomes lower; also noise has a greater influence on gate operation as V_{DD} becomes lower. Therefore, letting M be this margin, V_{DD} can be expressed as follows:

$$V_{DD} = M \times \frac{(2aV_{th} + f) + \sqrt{(2aV_{th} + f)^2 - 4a^2{V_{th}}^2}}{2a} \quad (15)$$

We assumed M=1 in the normal mode and $M > 1$ in the power-saving mode, and it is constant independent of the process generation. To calculate values for M for f=50% and 25%, which are our cases of interest, we used data for the supply voltages for each clock frequency from Crusoe TM5400 [9] (180nm process technology) and Crusoe TM5800 [12] (130nm process technology). Substituting these data, along with V_{th} from eq. (14), and a from eq. (13) into eq. (15), we can calculate M. We determined M=1.00 at the f=100%, M=1.24 at the f=50%, and M=1.40 at the f=25%.

Fig. 5 shows the estimation results of V_{DDmax}, V_{DD50}, and V_{DD25}, where V_{DD50} and V_{DD25} are the supply voltages at f=50% and 25%, respectively. The ranges of the supply voltages in the normal mode projected by SIA [10] are also shown. As the figure clearly shows, the ratios V_{DD50} / V_{DDmax} and V_{DD25} / V_{DDmax} increase as the process generation advances. This indicates that the effectiveness of DVS will decrease in the future.

Fig. 6 shows the energy consumption (an average of the benchmarks) in each process technology normalized by that in the normal mode. First, it can be seen that the effectiveness of DVS will steadily decrease as the process generations

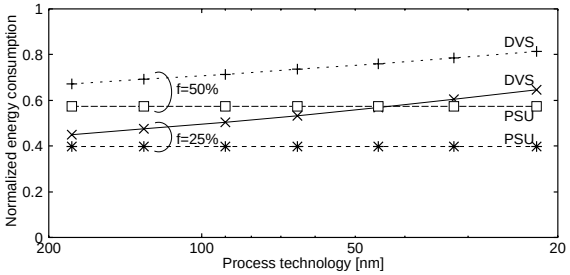

Figure 6: Normalized energy consumption at 50% and 25% frequencies in future technology.

advance. As a result, for example, when the process generation advances from the current 180nm process to 32nm after about 10 years, the energy consumption reduction rates at f=50% and 25% will decrease to 21% and 40%, respectively. In contrast, the energy consumption in PSU remains constant, independent of the advancement of the process generation. As a result, for example, PSU can reduce energy consumption by 27% and 34% more than DVS at f=50% and 25% respectively in the 32nm process technology.

6. CONCLUSIONS

In this paper, we evaluated an energy consumption reduction method called PSU in both current and future process technologies, and compared it to an existing method DVS. Our estimates show that currently PSU can reduce energy consumption by 14% and 11% more than DVS at f=50% and 25%, respectively. Although this improvement is moderate, the advantage of PSU over DVS will increase as process technology advances. For example, in about 10 years, the improvement rate at f=50% and 25% will increase to 27% and 34%, respectively. Consequently, PSU will become much more attractive as an energy consumption saving method in future mobile processors.

7. REFERENCES

[1] http://www.umc.com/

[2] F. E. Anderson et al. The Core Clock System on the Next-Generation Itanium Microprocessor. *ISSCC 2002 Visual Supplement to the Digest of Technical Papers*, pages 110–111, Feb. 2002.

[3] D. Burger et al. The SimpleScalar Tool Set, Version 2.0. Technical Report CS-TR-97-1342, Univ. of Wisconsin-Madison Computer Sciences Dept., July 1997.

[4] J. A. Butts et al. A Static Power Model for Architecture. *MICRO-33*, pages 191–201, Dec. 2000.

[5] L. T. Clark et al. An Embedded 32-b Microprocessor Core for Low-Power and High-Performance Applications. *IEEE JSSC*, 36(11):1599–1608, Nov. 2001.

[6] K. Diefendorff. TSMC Sets Sights on #1. *Microprocessor Report*, 14(6):17–21, June 2000.

[7] M. K. Gowan et al. Power Considerations in the Design of the Alpha 21264 Microprocessor. *DAC-35*, pages 726–731, June 1998.

[8] P. E. Gronowski et al. High-Performance Microprocessor Design. *IEEE JSSC*, 33(5):677–686, May 1998.

[9] D. Laird. *Crusoe Processor Products and Technology.* Transmeta Corporation, Jan. 2000.

[10] Semiconductor Industry Association. *International Technology Roadmap for Semiconductors 2000 Update Process Integration, Devices, Structures*, 2000.

[11] H. Shimada et al. Pipeline Stage Unification for Low-Power Consumption. *COOL Chips V*, pages 194–200, Apr. 2002.

[12] Transmeta Corporation. *Crusoe Processor Model TM5800 Product Brief*, July 2001.