A Hybrid Power Reduction Scheme Using Pipeline Stage Unification and Dynamic Voltage Scaling

Hajime Shimada[†], Hideki Ando[‡], and Toshio Shimada[‡]

† Graduate School of Kyoto University, ‡ Graduate School of Nagoya University
Engineering Building #10, Yoshida-honmachi, Sakyo-ku, Kyoto-shi 606-8501 Japan.
Phone:+81-75-753-5393, FAX:+81-753-5379, E-mail:shimada@kuis.kyoto-u.ac.jp

Abstract

Recent mobile processors have been required to provide both low-power consumption and high performance. To satisfy these requirements, we previously proposed pipeline stage unification (PSU), and showed that it can reduce energy consumption more than that of dynamic voltage scaling (DVS). This paper proposes a hybrid control scheme that combines PSU and DVS to enhance PSU. This scheme dynamically changes the number of pipeline stages, clock frequency, and supply voltage according to the required throughput. Our evaluation results in various throughput show that our scheme reduces power consumption by a maximum of 44% compared to DVS.

Keywords: low-power architecture, dynamic modification pipeline stage, dynamic voltage scaling

1 Introduction

Mobile processors must deliver both low-power consumption and high performance. *Dynamic voltage scaling* (DVS) is widely used in commercial processors (e.g. SpeedStep of Intel Pentium M) for this requirement. DVS changes the clock frequency and the supply voltage according to the processor workload and reduces power consumption accordingly. Thus, it is attractive for prolonging operating time for battery powered systems. Although DVS is currently an effective method for reducing power consumption, its effectiveness will diminish in the future. One major reason for this is scaling of the threshold voltage of transistors will slow due to leakage increase while scaling of the maximum supply voltage will be well as technology advances. This narrows the variable range of the supply voltage.

As an alternative, a method called *pipeline stage unification* (PSU) which dynamically unifies pipeline stages has been proposed [1, 2]. 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 power consumption in two ways. First, PSU saves power consumption by reducing the total load capacitance of the clock driver. Second, it attains higher IPC (instructions per clock cycle) because it reduces several latencies by reducing the number of pipeline stages.

We previously demonstrated that PSU can reduce energy consumption more than DVS at particular clock frequencies which we call *unification switching points* [1, 2]. A unification switching point is a clock frequency at which the unified numbers of pipeline stages are switched. Between the switching points, power is controlled only by the clock frequency. However, power consumption can be reduced more if the supply voltage is controlled at the same time. This idea motivates a hybrid scheme that combines PSU and DVS. Our hybrid scheme dynamically optimizes the number of unified stages, the supply voltage, and the clock frequency, adapting to the change of program phases, with satisfying given target throughput.

2 Hybrid Control Scheme Combining PSU and DVS

2.1 Scheme

Basically, our hybrid control scheme periodically adjusts a parameter set which consists of the *unification degree* U, the clock frequency f_U, and the supply voltage V_U, so that target throughput TP_{target} can be satisfied with minimum power consumption. Here, unification degree U is the number of unified stages and TP_{target} is assumed to be specified by the operating system. For this adaptation, execution is periodically sampled in a short period. We call this period a *sampling phase.*

In a sampling phase, all possible unification degrees are tried and IPC is measured. We call the period of each trial a *sampling sub-phase.* The scheme determines the parameter set (U, f_U, V_U) for each unification degree U (we will later describe how to derive f_U and V_U), and then estimates power consumption for each parameter set (see [2] about the estimation). Finally, the scheme chooses the best parameter set that is expected to exhibit lowest power consumption. The processor runs with the determined parameter set until the next sampling phase. We call this phase an *execution phase.* Fig. 1 shows an outline of the control scheme, which illustrates above-mentioned phases .

In each sub-sampling phase, f_U and V_U are derived from measured IPC as follows. First, The lowest clock frequency f_U that satisfies TP_{target} is calculated from the IPC. Then, the throughput from the beginning of the program execution to present is calculated, and the error from TP_{target} is derived. According to the obtained error, f_U is adjusted so that the error can shrink in the following execution phase. Finally, the minimum supply voltage V_U where the processor runs correctly in terms of timing with the obtained U and f_U is derived from a predetermined table (see Tab. 2 described later).

2.2 Implementation

The scheme is implemented as software. The software is stored in a dedicated ROM. As hardware, counters that count execution instructions and cycle count are prepared. The software is invoked by hardware when the counter reaches the end of an execution phase. It is executed as a user thread simultaneously with the program thread in an SMT (simultaneous multithreading) environment.

3 Evaluation Results

3.1 Simulation Environment

We built a superscalar simulator with our hybrid scheme based on the out-of-order execution simulator in the SimpleScalar tool set. Table 1 shows the processor configuration for our simulation. We use eight benchmark programs of SPECint95. We assume PSU with three unification degrees, U=1, U=2, and U=4. Fig. 2 shows the pipeline. The pipeline registers represented by dotted lines are bypassed in U=2. The pipeline registers represented by thin lines are further bypassed in U=4. We assume the clock driver consumes 30% of the total power, which is an approximate median of surveyed processors. Also, we assume that the clock frequency can be changed with 20 steps from 5% to 100% of the maximum frequency with equal intervals. The relationships between the supply voltages and the clock frequencies for U=1 are based on the data of the Intel Pentium M model 755 [3]. We assume that the supply voltage cannot be reduced under 0.988V which is the minimum supply voltage of the Pentium M.

We obtain the relationships between the supply voltages and the clock frequencies for $U \geq 2$ by the following equation:

$$V(U, f) = V(1, U \times f) \tag{1}$$

where $V(U, f)$ is the supply voltage of a unification degree U and a clock frequency f. Table 2 shows the relationships of the supply voltages and the clock frequencies in each of the unification degrees.

3.2 Power Consumption Reduction

We first measure throughput when the processor runs with the maximum clock frequency and U=1. We then evaluate the power consumption for 10 TP_{target} values, which vary from 10% to 100% of the maximum throughput.

We assume that the maximum clock frequency is 2GHz, the length of an execution phase is 2ms, and the length of a sampling sub-phase is 20μs. We do not include the overhead time for the algorithm execution in the evaluation since it is short enough compared to the time of an execution phase.

Fig. 3 shows the power consumptions and power reduction ratios of the hybrid control scheme and two DVS models. The horizontal axis denotes TP_{target} and the vertical axis denotes the power consumption normalized by that in the execution with the maximum throughput (lines) or power consumption reduction ratio (bars). Three lines represent the average power consumption of the processor with real DVS, ideal DVS, and the hybrid control scheme, respectively. The processor with the real DVS runs with U=1, and the clock frequency and the supply voltage are dynamically determined by the algorithm described in Section 2.1. The ideal DVS is *ideal* because it knows IPC when program execution finishes *a priori* and thus the best clock frequency for the minimum power consumption can be determined without adaption.

As found in Fig. 3, the hybrid scheme achieves lower power consumption compared to both the real and ideal DVSs at any point of TP_{target}. The maximum power reduction ratio to the real DVS is 44% at 20% TP_{target}, and even to the ideal DVS 40%.

As a general tendency, reduction ratio becomes larger as TP_{target} becomes smaller. There are two reasons for this. First, the DVS loses the power reduction ability at the clock frequency of 30% or less because of the bounded supply voltage. Meanwhile, the hybrid scheme can use PSU at such low clock frequencies for power reduction. Second, the hybrid scheme has more choices about the unification degrees when lower clock frequencies are chosen (see Table 2).

4 Conclusions

We proposed a hybrid power control scheme that combines PSU and DVS. Our scheme chooses best parameters of a unification degree, clock frequency, and supply voltage with satisfying given target throughput. Our evaluation results show that our scheme reduces power consumption for various target throughput more than DVS.

References

[1] H. Shimada, H. Ando, and T. Shimada, "Pipeline Stage Unification for Low-Power Consumption," CoolChips-V, pp.194-200, Apr. 2002.

[2] H. Shimada, H. Ando, and T. Shimada, "Pipeline Stage Unification: A Low-Energy Consumption Technique for Future Mobile Processors," ISLPED-2003, pp.326-329, Aug. 2003.

[3] Intel Corporation, "Intel Pentium M Processor Datasheet," Jun. 2003.

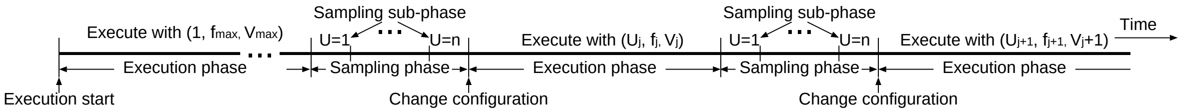

Figure 1: Outline of hybrid control scheme.

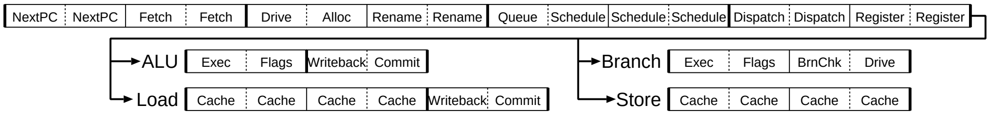

Figure 2: Assumed PSU pipeline.

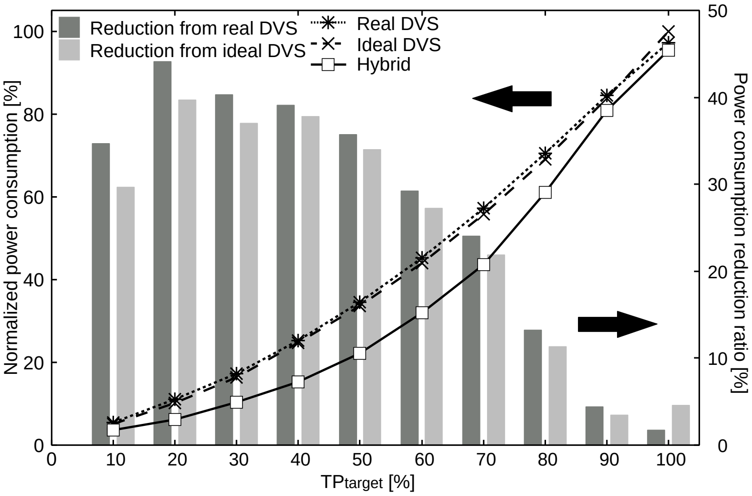

Figure 3: Power consumption of the hybrid control scheme and DVS.

Table 1: Processor configuration.

processor core	8-way of out-of-order issue, 64-entry RUU, 32-entry LSQ, 8 int ALU, 4 int mult/div, 8 fp ALU, 4 fp mult/div, 8 memory ports, 20 pipeline stages
branch prediction	8K-entry gshare, 6-bit history, 2K-entry BTB, 16-entry RAS
L1 I-cache	64KB/32B line/1-way
L1 D-cache	64KB/32B line/1-way
L2 unified cache	2MB/64B line/4-way
memory	64 cycles for first hit, 2 cycles for burst interval
I-TLB	16-entry, 128 cycles miss latency
D-TLB	32-entry, 128 cycles miss latency

Table 2: Relationships of the supply voltages and the clock frequencies.

clock frequency	U=1	U=2	U=4
100%	1.340V	—	—
95%	1.316V	—	—
90%	1.292V	—	—
85%	1.268V	—	—
80%	1.244V	—	—
75%	1.220V	—	—
70%	1.196V	—	—
65%	1.172V	—	—
60%	1.148V	—	—
55%	1.124V	—	—
50%	1.100V	1.340V	—
45%	1.076V	1.292V	—
40%	1.052V	1.244V	—
35%	1.020V	1.196V	—
30%	0.988V	1.148V	—
25%	0.988V	1.100V	1.340V
20%	0.988V	1.052V	1.244V
15%	0.988V	0.988V	1.148V
10%	0.988V	0.988V	1.052V
5%	0.988V	0.988V	0.988V