Saravanan *et al. Hum. Cent. Comput. Inf. Sci.* (2015) 5:28
DOI 10.1186/s13673-015-0046-x

Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

An energy-delay product study on chip multi-processors for variable stage pipelining

Vijayalakshmi Saravanan[1*], Alagan Anpalagan[1] and Isaac Woungang[2]

*Correspondence:
vsaravan@rnet.ryerson.ca
[1] WINCORE Lab, ELCE Department, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada
Full list of author information is available at the end of the article

Abstract

Power management is a major concern for computer architects and system designers. As reported by the International Technology Roadmap for Semiconductors (ITRS), energy consumption has become one of the most dominant issues for the semiconductor industry when the size of transistors scales down from 22 to 11 nm nodes. In this regard, current existing techniques such as dynamic voltage scaling, clock gating, and the Complementary metal-oxide semiconductor technology have shown their physical limits; therefore, scaling will no longer be a valid strategy for achieving power-performance improvement. To overcome this critical issue in energy-efficient processor design, there is a clear demand for alternative solution. In this paper, an approach that provides a promising solution for energy reduction is proposed, by using a micro-architectural technique referred to as variable stage pipelining, which can be further validated and extended to different application domains such as mobile and desktop. An analytical model for evaluating the relationship between the number of cores and the pipeline stage depth in a chip multi-processor is also proposed, based on which the optimal pipeline depth for various metrics are calculated.

Keywords: Chip multi-processors (CMP), Variable stage pipelining (VSP), Power-performance, Optimal pipeline

Background

In the recent years, there has been a growing demand for more efficient power management schemes for computing domains such as mobile, enterprise, cloud computing, to name a few; and energy efficiency has become one of the main design goals for such schemes. For mobile processors, high performance and low energy have been among the main design targets for computer architects and hardware designers. But, the design process of such processors is yet to be fully adjusted to fulfill the needed goals. Power efficiency refers to not only low leakage currents and small switched capacitance, but also to the efficiency of the power distribution network, the power conversion circuitry, and the heat removal. Therefore, there is a clear demand for a holistic approach to power optimization and management that considers all these factors; and a general shift away from CPU-centric design thinking is taking place. In mobile systems, the focus should be on display, radio, and sensors; whereas in enterprise systems, memory, storage and networks are becoming increasingly important in terms of their power usage.

Recently, in order to save their battery life while yielding high performance, handheld devices such as laptops and mobile processors have been required to exhibit low-power consumption. Even though the Complementary metal-oxide semiconductor (CMOS) and dynamic voltage scaling (DVS) technologies have potentials for achieving the power-performance, their effectiveness is expected to be greatly reduced as the process technology advances. With the shrinking CMOS minimum feature sizes, higher chip densities and lower operating voltages have led to the issue of voltage and temperature (PVT) variability, which is one of the main challenges in the design of power-efficient processors. With respect to power-aware processor design, the challenge is to develop an energy-delay optimization framework, which includes some methods for the energy-delay product (EDP). Other issues that stand out in power optimization and management design challenges are: (1) the statistical uncertainty about the workload, parameters and target system, and (2) a lack of benchmarks and evaluation techniques.

The important factors that have motivated our study of the multi-core power-performance trade-offs are: (1) how will the technology transition from 22 to 11 nm nodes affect the design of power efficiency integrated with multi-processors (2) what is the impact of power-performance on the power efficiency of the chip?, and (3) what are the optimal number of pipeline depth and cores desired to achieve balanced tradeoffs in terms of power efficiency versus computational speed?

It is expected that by 2020, the process technology size will be close to 7 nm (as depicted in Fig. 1). Therefore, existing techniques such as dynamic voltage and frequency scaling (DVFS), CMOS, memory and process technology, if used, will be performing under some limitations due to the shrinking process technology below 22 nm [1]. In this paper, a technique referred to as variable stage pipelining (VSP) is proposed as an alternative to the DVFS scheme for power saving purpose.

In an attempt to answer these questions discussed above, our previous work on power management in multi-core processors [3] has been improved in this paper, as follows:

- State-of-the-art pipelining techniques in chip multi-processors (CMP) design are utilized and simulated, and their power and performance results are presented. It should be noted that so far, these techniques have mostly been tested on simple and

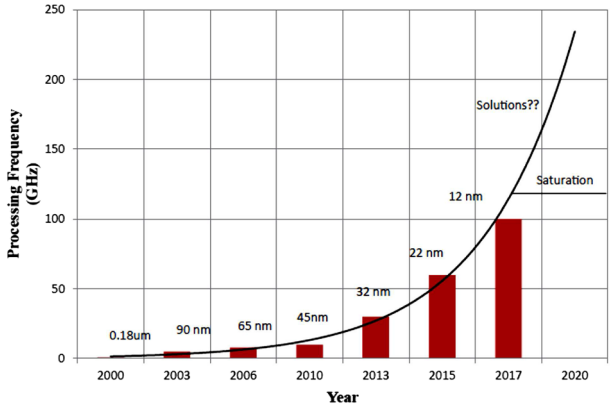

Fig. 1 Technology trends in transistor scaling at 12 nm [2]

well-known theoretical functions. To the best of our knowledge, earlier studies on power and performance on real time implementation are rare.
- A mathematical model for the evaluation of the proposed approach is introduced, and then solved using a software program execution of code.
- Different aspects of the power-performance efficiency are compared against each other using the existing techniques, and four performance metrics.

The rest of the paper is organized as follows. In "Related works", some related works on optimal pipelining are presented. In "Analytical modeling", an analytical modeling and description of our architectural framework is given. In "Experimental modeling", an experimental modeling and the simulation results are presented. In "Conclusions", we conclude our work.

Related works

In the recent years, the exponential growth of process technology arises the problem of energy consumption has become the major constraints in chip manufacturing industries. It becomes the major problem in performance improvements in desktop to high-end processors. In classic CMOS scaling, the increase in performance has been mostly achieved by increasing the instructions per cycle (IPC) and the clock frequency. These improvements arise from a substantial increase in the pipeline depth (so-called deeper pipelines). On the other hand, high-end data center designs have been driven by performance alone issues whereas power has become the main concern in microprocessor design. Therefore, both power and performance have to be taken into consideration even at the micro-architectural level, and appropriate power/performance metrics have to be devised for simulation purpose. In [4], Kunkel and Smith studied the optimum pipeline depth and defined a set of performance metrics. Recently, this work has been revisited using performance-only metrics [5–7, 11], and some pipelined processor power models have been formulated.

There is a growing demand of low power design, and the increasing demand of processor performance improvements leads to deeper pipelines in processor design points. In [8], authors studied the optimum metric for various workloads and proposed a theoretical approach to find the power and performance trade-off and dynamically changing pipeline depth during program execution as described in [9–11]. In [12], authors discussed about minimizing power consumption to get the optimal power-performance under throughput constraints.

A measure of performance increase with pipeline depth is the change in CPI (cycles per instruction) [13]. This can be attributed to the following main reasons (1) adding more pipelines by using higher clock speeds and lower supply voltages generally lead to shorter logic depth. Thus, the delay would be considerably minimized. But, the real measure of performance optimization with pipeline depth is to consider the ratio MIPS/BIPS, where MIPS stands for million instructions per second and BIPS stands for billion instructions per second, or to use the time per instruction (TPI) of the machine; (2) TPI is the product of CPI and the cycle time of the processor. As pipeline depth is increased, the cycle time goes down. This is attributed to the fact that the entire logic time is broken down into multiple numbers of intervals, but the total time taken to process an instruction is not increased. It should be noted that the results obtained in our previous

studies [3] are based on simulations using a particular system configuration, whereas in this paper, the micro-architecture is taken into consideration.

Nowadays, the voltage scaling techniques widely used technique for power savings. However, there is a huge concerns for future scaling due to advancement in process technologies. The energy reduction techniques are investigated and some studies on the variable stage pipelining and alternative for existing DVFS techniques have been conducted in [3, 14, 15]. However, these studies focused on fixed pipeline depth during the program execution. Using such pipeline depth may be efficient for certain programs, but may also lead to optimal pipeline depth for other programs with different behaviors. In [14], it is argued that deactivating the pipeline stages and using a shallow pipeline can help reducing the processor's power consumption. Most existing representative schemes for reducing the processor's power consumption are localized and often application specific. On the other hand, understanding and modeling properly the emerging applications and the mobile user behaviors, as well as developing the metrics for user's experience, are essential. The reason for doing so is that this will allow a system to decide how much power or energy should be allocated to a given computation [16].

In this paper, the approach used shows that the various intensities of workloads may cause the optimal pipeline depth and its corresponding cores. Our focus is on finding the relationship between pipeline depth and power consumption to the adopted variable stage pipelining. The pipeline stage unification and the number of cores of a chip multiprocessors are optimized simultaneously in order to derive the best possible power-performance ratio.

Analytical modeling

Attaining low power consumption with high performance is a desirable goal for a number of applications that involve sensors, hand-held devices to high-end computing. Such systems are typically wasting unnecessary energy consumption through switching activity. Thus, it is necessary to study the energy consumption of the logic circuits. This can be expressed as:

$$E_{Total} = \alpha \cdot (1/2 \cdot C_{Load} \cdot V_{cc}^2) + I_{Leak} \cdot V_{cc} \cdot \tau_{Switching}. \quad (1)$$

From Eq. (1), the easiest way to minimize energy consumption is to scale the supply voltage V_{cc}. However, in CMOS circuits scaling V_{cc} can lead V_{th} (threshold voltage) to operate below threshold level hence causing a delay in $\tau_{Switching}$ which in turn causes leakage energy component to become the dominant contributor to E_{Total}, especially for low activity (α) factors. There are many studies on minimizing energy consumption E_{Total} as suggested in [17–19].

The overall power consumption is expressed as follows:

$$P = A \times C \times V^2 \times f + VI_{leak}, \quad (2)$$

where A is the activity factor and C is the capacitance and the second term represents the static power consumption due to the leakage current, I_{leak}. From Eq. (2), it can be deduced that reducing the frequency has a good impact on the battery life of the PC at the expense of performance and also shows that I_{leak} has dependance on V_{th}. For the

every technology generation, V_{th} should be decreased according to V_{cc} scaling [20]. Thus the total power consumption will be dominated by leakage current I_{leak} as technology progresses. This is also suggested in Eq. (2) that reducing the supply voltage is the most effective way of reducing power consumption. But it is also understood that halving the voltage also reduces the processor's maximum operating frequency which in turn leads to performance loss. In order to compensate this performance loss, we can use either parallel or pipelined implementations. Also, DVFS and pipelining are among the most popular techniques to reduce the frequency of the processor, but their applications are limited only to single-core processors. In case of multi-core processors, these techniques are less effective.

But how the pipelining and parallel implementations are useful for reducing the power consumption? In research paper [20] illustrate that the leakage current is the dominant source of energy consumption in scaled transistors. Because, sub-threshold and leakage current both depend on the total gate count, transistors and gate width, a pipelined approach makes substantial contribution in reducing the leakage current. As noted, pipelining gives the low-power processor solution because it always runs at low voltage [21]. With this insight, we propose an alternate solution for power-efficient processor design using pipelining concept called variable stage pipelining (VSP).

Power-performance vs. pipeline stage unification degree

As described in [3], for a given technology, both the instructions per second (IPS) and the instructions per cycle (IPC) are considered as performance metrics in the processor pipeline design, leading to the following relationship:

$$IPC_N = \frac{IPS_N}{f_N}, \tag{3}$$

where IPC_N, f_N, and IPS_N are respectively the IPC, the frequency and the IPS of the N-core processor. Now, lets consider a hypothetical uni-core processor that performs the tasks in the same time as that of a N-core processor. The execution time of a uni-core processor is given by,

$$t_1 = \frac{IC}{IPC \times f}, \tag{4}$$

where IC is the number of instructions executed by the core. On the other hand, the executing time of a N-core processor is given by

$$\frac{IC_N}{IPC_N \times f_N} = t_N. \tag{5}$$

Since both execution times are equal and the workload tends to be equally divided among the cores in the multi-core processors, we get $IC_N = IC/N$ for the same program. If V_N denotes the operating voltage, the following relation [22] can be derived:

$$V_N = \frac{(V - 2 \times V_{th}) \times IPC}{N \times IPC_N} + 2 \times V_{th}. \tag{6}$$

The typical values of *IPC*, *V*, V_{th} are known, precisely these are: 1.0245, 1 Volt and 0.20 Volt respectively [23]. Substituting those values and using Eq. (3) yields

$$\frac{N \times IPS_N}{f_N} \leq 10.245. \tag{7}$$

If α is the unification degree under the variable stage pipelining, the following relationship can be obtained:

$$f_\alpha = \frac{f_{max}}{\alpha}. \tag{8}$$

Thus, from Eqs. (7) and (8), the following relationship between the power performance and the unification degree under the VSP approach can be obtained:

$$\frac{N \times \alpha \times IPS_N}{f_{max}} \leq 10.245. \tag{9}$$

Energy reduction with VSP

In this section, the energy reduction of the variable stage pipelining and its optimality are evaluated. According to the proposed VSP [3], the energy-delay product is used as the power-performance metric based on the VSP history information. Consider three metric α_1, α_2, α_3:

1. Higher workload Metric-α_1 and Corresponding unification degree-β_1
2. Medium workload Metric-α_2 and Corresponding unification degree-β_2
3. Light workload Metric-α_3 and Corresponding unification degree-β_3

In order to estimate the energy consumption using the VSP approach, we have investigated how the performance and power will change as the unification degree varies for the diverse processor cores. The relationship between various metrics and their power-performance implications have been derived as follows.

(a) *IPS* To find the optimal depth, Eq. (9) is considered as the basis for further analysis of the various power-performance metrics. For a *N*-core processor, the relationship of IPS with respect to the metric is obtained from Eq. (9) as:

$$\frac{\alpha \times N \times \beta_N}{f_{max}} \leq 10.245. \tag{10}$$

For $\beta_1 > \beta_2 > \beta_3$, and $\alpha_1 < \alpha_2 < \alpha_3$, assuming α and f_{max} are constant, the relationship between N and β_N is obtained as:

$$IPS = N \times \beta. \tag{11}$$

In Eq. (11), as the number of cores increases, the metric also increases. But still N cannot be increased beyond a certain value due to energy constraints. Thus, *IPS* is not as reliable as it should be since it does not take the power into account and it is considered as performance-only metric. This necessitates the need to analyze the Energy-Delay Product (EDP) to serve as the power-performance metric, which is suitable for most modern processor platforms such as laptops and mobile phones.

(b) *E-Metric* We assume the following for the Energy-Delay-Product (EDP) analysis: For a N-core processor, assuming that N, f_{max} and α are constants, the metric is proportional to the square of the unification degree. Therefore, for $\beta_1 > \beta_2 > \beta_3$, and $\alpha_1 > \alpha_2 > \alpha_3$. By considering Eq. (9) and by assuming that the voltage scales linearly with the frequency with respect to power, the following relationship is derived:

$$\frac{N \times \beta_N \times f_{max}^2}{\alpha^2} \leq 10.245. \tag{12}$$

In Eq. (12), we see that as the metric increases, the pipeline stages decrease. When N increases, IPS also increases since IPS/W is a power-performance metric; but on the other hand, the power increases too. Therefore, N cannot exceed a certain value. If that happen, it will result to a decrease in the metric. For higher metric, N should neither be high nor low.

(c) *EDP Metric* By considering Eq. (9) and assuming that voltage scales linearly with frequency with respect to $(Power \times IPS_N)$, the following relationship can be obtained:

$$\frac{N \times \beta_N \times f_{max}}{\alpha} \leq 10.245. \tag{13}$$

In Eq. (13), we see that as the metric increases, the pipeline stages decrease. When N increases, there is an increase in IPS due to IPS^2/W which in turn leads to increase in power too. Therefore, N cannot go beyond a certain value. If it exceeds the limit, results in decrease the metric. Hence, for the higher metric, N should be either high or low.

(d) *ED* ^{2} *P Metric* By considering Eq. (9) and by assuming that the voltage scales linearly with the frequency with respect to $(Power \times IPS_N{}^2)$, the following relationship can be obtained:

$$N \times \beta_N \times K \leq 10.245, \tag{14}$$

where K is a constant used in the experiment. According to $BIPS^3/W$, the CMP (chip multi-processor) configuration consists of large number of fairly narrow cores. Wider cores are considered to be too much power hungry to be competitive.

Proposed low power architectural technique

Our proposed approach VSP, saves the energy consumption in two ways: First, this method reduces the total load capacitance of the clock driver by stopping the clock signal to "by-passed" the pipeline registers. Second, it reduces the clock cycle count of the program execution by reducing the number of pipeline stages. In order to decrease the number of pipeline stages by by-passing a few stages, we enable the following functionalities for the clock/full-time clock, the part-time clock, and the unification signal:

- The full-time clock signal is always active regardless of the unification.
- The part-time clock signal is deactivated when the pipeline stages are unified. It is active when they are not unified.

- The unification signal indicates the pipeline stage unification. Since the pipeline register between two adjacent combinatorial logic circuits is inactive or by-passed, the two logic circuits operate together as a single stage.

In order to by-pass a pipeline register, we have used two methods. In the first one, the pipeline register logic is organized in such a way that a signal can pass through it regardless of the clock signal when the VSP is enabled. This solution can be implemented if the pipeline registers are made up of transparent latches. It is simple, but its drawback is the cost-effectiveness of using transparent latches in the pipeline. In the second method, the logic gates and multiplexors are involved. The multiplexors are meant to decide which pipeline registers are active and which ones are to be shutdown when the unification signal is applied. An example of this solution is shown in [24].

Pseudo-code of the variable stage pipelining algorithm

The implementation of the proposed VSP approach consists of four modules as shown in Algorithm 1. The execution of this algorithm is shown in Fig. 2 and its output is shown in Fig. 3.

Experimental modeling

In this section, the optimal unification degree and power-performance of the proposed VSP technique are quantitatively measured. To study the effect of different pipeline depths on various cores, we have varied the number of cores of a modern superscalar processor architecture, which has out-of-order execution simulator in the Simple-Scalar Toolset (so-called M-Sim [25]), Table 1 lists the processor configuration and clock

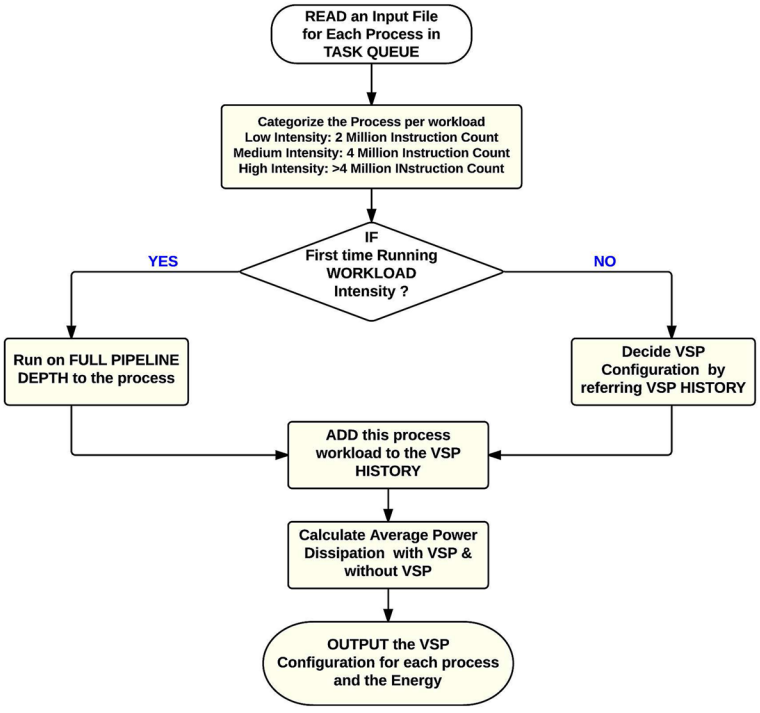

Fig. 2 Flow chart of proposed algorithm

Algorithm 1 Pseudo code of the proposed VSP implementation

```
/* Benchmark program parameters and VSP configuration */
Write_to_Mem()
{
instruction_count;
branch_frequency;
cache_access_frequency;
num_of_int;
float_point_op;
}
/* Applications run by the user */
while(execute_all_apps)
{
    if(first_exe)
    {
        decision_module()
        {
            parameter_extract();
            comparision();
            generate_VSPconfig();
            write_to_table() //present in VSP history
            {
                app_id;
                vsp_config;
            }
        }
    }
}
```

```
/* Decision Module */
decision_module()
{
comm_task_queue()
    {
fetch_next_app();
    }
}
/* VSP History Cache */
open_VSP_history_cache()
{
compare_app_id();
    if(cache_hit)
comm_VSP_control_module()
        {
send_VSP_config();
        }
else
move_app_id_to_cache();
}
write_to_VSP_history_cache()
    {
if(app_exe > threshold)
    {
write_app;
write_VSP_Config;
    }
        }
```

frequency assumptions. M-Sim is a multi-threaded micro-architectural simulation environment with a detailed cycle-accurate model for the key pipeline structures which is similar to the current processor architectures. For power estimations, M-Sim includes the Wattch framework as applied to Simple-Scalar. Various applications such as heavy, moderate and light are provided so that the processor can perform the power-performance analysis. The simulation results have been obtained for various cores such as 2, 4, 8, 16, and 32 cores. Based on these results [3], it has been concluded that beyond 32 cores, the number of non-architectural registers available for a core will be 0. Hence, the number of cores beyond 32 cannot be achieved in M-Sim. The pipeline depth is calculated for each scenario and the optimal pipeline depth-core combination is derived. Figure 3 shows the actual readings for power and energy consumption, as collected from the program execution of using VSP techniques.

Results of VSP and optimum analysis

We have studied the optimum number of cores among several pipeline stage unification degrees. In this subsection, we study the optimality by adopting the VSP technique. We assumed the total number of instructions as 10 KIPS, Issue-width = 8. In a few of the benchmarks (MiBench) the obtained simulation results for various cores such as 2, 4 and 8 are given below: Total simulation time = 10, 20 and 40 sec respectively; Total Power

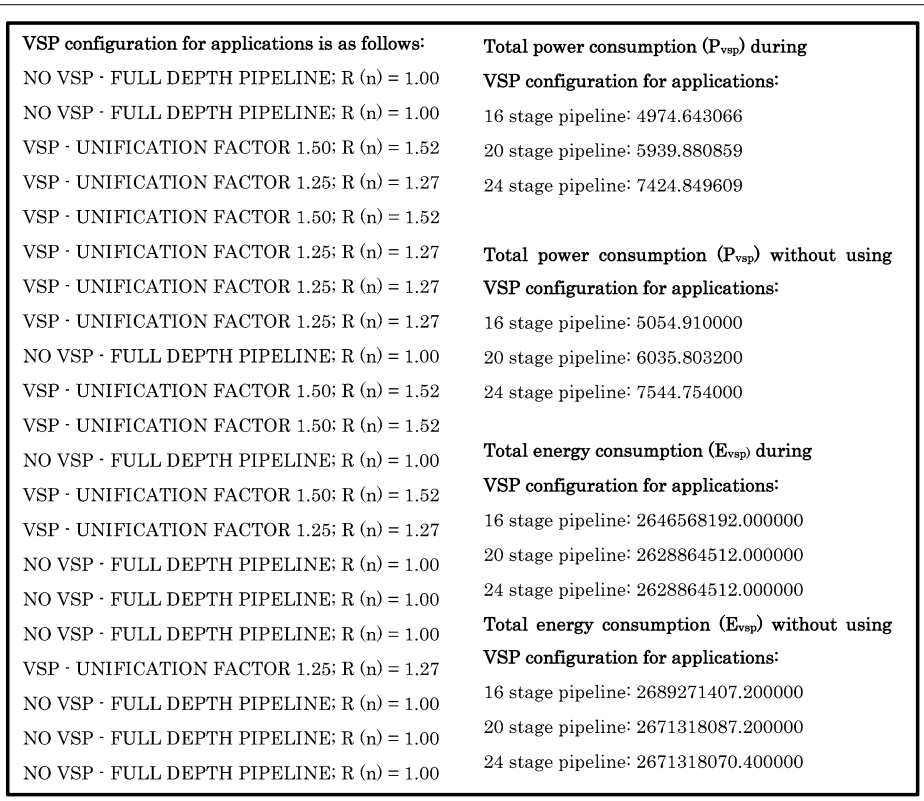

Fig. 3 Sample output of proposed VSP approach

Table 1 Processor configuration and Clock frequency assumptions

Parameters	Alpha 21264 processor
Fetch, issue, commit-width	4,4 (int), 2 (float), 11
Reorder buffer size	80
Issue window	20 (int), 15 (float)
Load/store queue	32 (load), 32 (store)
Register file	160
Floating-point ALU	1 adder, 1 multiplier
Integer ALU	4 adder, 4 multiplier
L1 Data, instruction-cache	512642
Dtlb, Itlb	164128, 132128 (fully associative)
Clock frequency rate $f(\beta_1 = 1, \beta_2 = 1.5, \beta_3 = 2)$	100 %, 66.7 %, 50 %

Dissipation = 718.5, 1465.1 and 3006.9; Throughput IPC = 1.69, 3.4 and 7.3 respectively. By using these, the performance and optimal pipeline depth for various metrics have been calculated as follows.

(a) *IPS Metric* Pipeline depths for various configurations such as 2, 4, 8 and their corresponding pipeline stage unification degrees are 2, 1.5, 1.25 respectively. For the obtained results of the *IPS* metrics and the f_{max} values of each configuration, we obtained the following relationship:

$$\frac{N \times IPS_N \times \alpha}{f_{max}} \leq 10.245. \tag{15}$$

(b) *E-Metric* The pipeline depths for various configurations such as 2, 4, 8, 16, 32 and their pipeline stage unification degrees are 1, 1, 1.25, 1.5, and 2 respectively. For the obtained results of the IPS/W metrics, and the f_{max} and power consumption (ρ) values of each configuration, we have obtained the following relationship:

$$\frac{N \times \beta_N \times k_1 \times f_{max}^2}{\alpha^2} \leq 10.245, \tag{16}$$

where $\beta = \frac{IPS_N}{\rho}$ and k_1 (where $k_1 = 2.8 \times 10^{-}3$) is the technology design parameter.

(c) *EDP Metric* The pipeline depths for various configurations such as 2, 4, 8, 16, 32 and their pipeline stage unification VSP degrees are 1, 1, 1.25, 2, and 3.5 respectively. For the obtained results of the IPS^2/W metrics, and the f_{max} and power consumption (ρ) values of each configuration, we have obtained the following relationship:

$$\frac{N \times \beta_N \times k_1 \times f_{max}}{\alpha} \leq 10.245, \tag{17}$$

where $\beta = \frac{IPS_N^2}{\rho}$ and k_1 is the technology design parameter.

(d) *ED^{2}P Metric* The pipeline depth for various configurations such as 2, 4, 8 and their pipeline stage unification degrees are 2, 1.5, and 1.25 respectively. For the obtained results of the IPS/W metrics, and the f_{max} and power consumption (ρ) values of each configuration, we have obtained the following relationship:

$$N \times \beta_N \times k_1 \leq 10.245, \tag{18}$$

where $\beta = \frac{IPS_N^3}{\rho}$ and k_1 is the technology design parameter.

Analysis and discussions

In this section, we present an analysis of the relationship between the number of cores and pipeline stage depths. The results have been recorded for the different metrics with their various VSP unification degrees as shown in Table 2.

In Fig. 4, it can be observed that there is an increase in the number (N) of cores, which in turn has led to an increase in the performance and unification degree as well. Thus, the optimal point shifts towards the deeper pipelines with higher number of cores. In Fig. 5, it can be observed that there is also an increase in the number (N) of processor core, which has led to an increase in the energy per delay (IPS/W) and a decrease in

Table 2 EDP results vs. VSP unification degree

IPS		E-Metric		EDP		ED^{2}P	
No. of cores	VSP degree	No. of cores	VSP degree	No. of cores	VSP degree	No. of cores	VSP degree
2	2	2	1	2	1	2	2
4	1.5	4	1	4	1	4	1.5
8	1.25	8	1.25	8	1.25	8	1.25
16	0	16	1.5	16	2	16	0
32	0	32	2	32	3.5	32	0

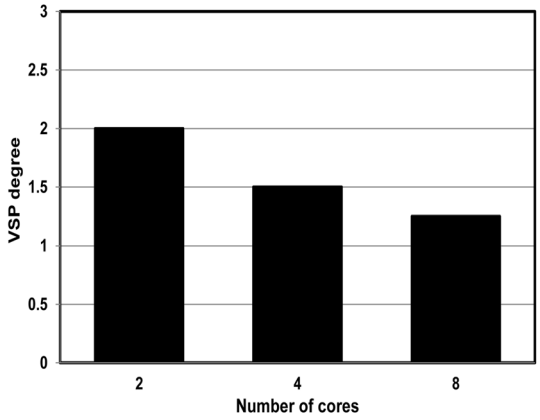

Fig. 4 IPS vs. VSP unification degree

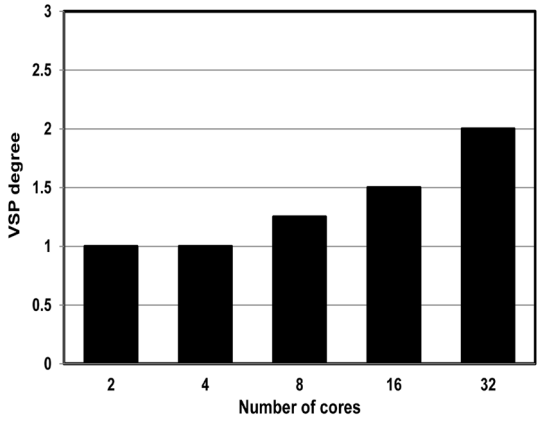

Fig. 5 E-Metric vs. VSP unification degree

the unification degree. Thus, the optimal point shifts towards the shallower pipelines, with a medium number of cores. Also, in Figs. 6 and 7, it can be observed that there is an increase in the number (N) of cores and a decrease in the unification degree, which have led to an increase in both the energy-delay product (IPS^2/W) and ED^2P (IPS^3/W). Thus, this metric's optimal point shifts towards the medium core with very low pipeline stages and a higher number of cores with shallower pipeline stages for the ED^2P. Overall, from the simulation results, it was observed that the optimal number of cores—pipeline stage depth combination is a 8 core processor with a VSP unification degree of 1.25 for all the metrics. Such a configuration will yield the maximum performance without compromising the power consumption. As observed from the results obtained for ED^2P, beyond 16 cores will give rise to inconsistent results due to the memory coherency.

From the analysis of the performance and the energy-delay product metrics, it can be argued that the optimal point varies for different metrics. By considering performance-only metrics such as IPS and ED^2P, the optimal number of cores lies between higher and lower pipeline stage unification degree. When considering power-performance metrics such as E and EDP, the optimal number of cores lies between low-medium number of cores with low-moderate number of pipeline stages. Thus, it can be concluded that when

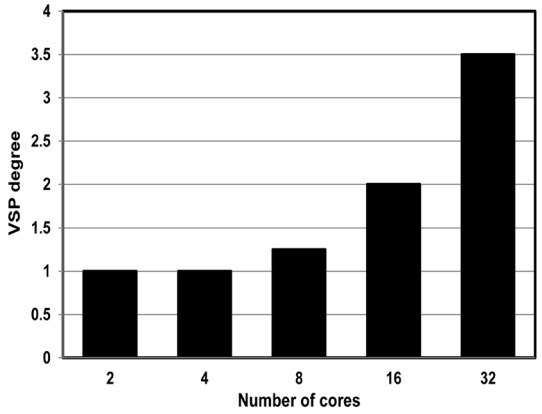

Fig. 6 EDP vs. VSP unification degree

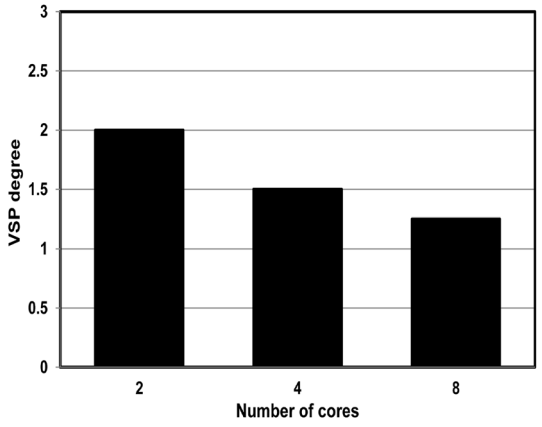

Fig. 7 ED^2P vs. VSP unification degree

both power and performance are taken into account, a medium number of cores with moderate number of pipeline stages will be the optimal configuration.

Limitations and future enhancement

The proposed VSP configuration requires the preferred workload characteristics. The successful detection of workload characteristics may help to optimize the power modeling. The current limitation of this proposed approach is to predict the workload characteristics by using the history information and the power reconfigurations are usually programmed well in advance of the actual program execution, similarly branch prediction also required in out-of-order processors. Hence, the future development of hardware prototype VSP power model is based on understanding the workload characteristics and optimized hardware scheduling algorithm in order to improve the power/performance efficiencies.

Conclusions

In this paper, a VSP-based micro-architectural power saving technique for balanced power-performance trade-offs is proposed and its efficiency in terms of several

performance metrics is demonstrated by experiments. An analytical model is also proposed to analyze the relationship between the number of cores and the pipeline stage depth. The proposed method can be applied to explore energy efficient design points in multi-threaded multi-core CPUs. The simulation findings have revealed that the optimal number of cores—pipeline stage depth combination is an 8 core processor with a VSP unification degree of 1.25 among all the studied metrics. Such a configuration will give a good trade-off between power and performance. In future, we intend to compare the proposed VSP-based scheme against some benchmark schemes, using the multi-threaded micro-architectural simulation environment, and performance metrics. Our estimates show that using VSP technique saves energy consumption approximately 2 % as shown in Fig. 3. Though the % improvement is moderate, VSP technique is quite useful as the technology progresses in future mobile, laptop and desktop processors.

Authors' contributions

VS investigated the state-of-the-art pipelining techniques in chip multi-processors (CMP) and proposed a way to utilize it. AA proposed the mathematical model to sustain the above-mentioned pipeline techniques. IW investigated the different simulation scenarios and helped in conducting the associated simulations. All authors read and approved the final manuscript.

Author details

[1] WINCORE Lab, ELCE Department, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada. [2] Computer Science Department, Ryerson University, Toronto, Canada.

Competing interests

The authors declare that they have no competing interests.

Received: 31 May 2015 Accepted: 23 August 2015
Published online: 21 September 2015

References

1. Chakraborty K, Roy S (2012) Architecturally homogeneous power-performance heterogeneous multicore processor. US Patent App. 13/495,961. (Online). https://www.google.com/patents/US20120324250
2. (2012) The ITRS Technology Working Groups, International Technology Roadmap for Semiconductors (ITRS). (Online) http://www.itrs.net/ITRS%201999-014%20Mtgs,%20Presentations%20&%20Links/2013ITRS/2013Chapters/2013ExecutiveSummary.pdf. Accessed 2013
3. Vijayalakshmi S, Anpalagan A, Woungang I, Kothari D (2013) Power management in multi-core processors using automatic dynamic pipeline stage unification. In: 2013 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), July 2013, pp 120–127
4. Kunkel SR, Smith JE (1986) Optimal pipelining in supercomputers. In: Proceedings of the 13th annual international symposium on Computer architecture, ser. ISCA '86. IEEE Computer Society Press, Los Alamitos, pp 404–411. (Online). http://dl.acm.org/citation.cfm?id=17407.17403
5. Hartstein A, Puzak TR (2002) Optimum Power/Performance Pipeline Depth. In: IEEE Computer Society, IBM-T. J. Watson Research Center
6. Hrishikesh MS, Farkas KI, Burgert D, Keckler SW, Shivakumar P (2002) The optimal logic depth per pipeline stage is 6 to 8 FO4 inverter delays. In: Proceedings of the 29th Annual International Symposium on Computer Architecture, pp 14–24
7. Srinivasan V, Brooks D, Gschwind M, Bose P, Zyuban V, Strenski PN, Emma PG (2002) Optimizing pipelines for power and performance. In: Proceedings of the 35th annual ACM/IEEE international symposium on Microarchitecture, ser. MICRO 35. IEEE Computer Society Press, Los Alamitos, pp 333–344. (Online). http://dl.acm.org/citation.cfm?id=774861.774897
8. Hartstein A, Puzak TR (2002) The optimum pipeline depth for a microprocessor. SIGARCH Comput Archit News 30(2):7–13. doi:10.1145/545214.545217
9. Borkar S (1999) Design challenges of technology scaling. IEEE Micro 19(4):23–29. doi:10.1109/40.782564
10. Srinivasan V, Brooks D, Gschwind M, Bose P, Zyuban V, Strenski PN, Emma PG (2002) Optimizing pipelines for power and performance. In: International Symposium on Microarchitecture (MICRO35), Nov. 2002. Selected as one of the four Best IBM Research Papers in Computer Science, Electrical Engineering and Math published in
11. Sprangle E, Carmean D (2002) Increasing processor performance by implementing deeper pipelines. In: Proceedings of the 29th Annual International Symposium on Computer Architecture, ser. ISCA,'02. IEEE Computer Society. Washington, DC, pp 25–34. (Online). http://dl.acm.org/citation.cfm?id=545215.545219
12. Ghasemazar M, Pakbaznia E, Pedram M (2010) Minimizing the power consumption of a chip multiprocessor under an average throughput constraint. In: IEEE ISQED, pp 362–371. (Online). http://dblp.uni-trier.de/db/conf/isqed/isqed2010.html#GhasemazarPP10

13. Hennessy JL, Patterson DA (2006) Computer architecture: a quantitative approach, fourth edn. Morgan Kaufmann Publishers Inc., San Francisco
14. Koppanalil J, Ramrakhyani P, Desai S, Vaidyanathan A, Rotenberg E (2002) A case for dynamic pipeline scaling. In: Proceedins of the 5th International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES'02), pp 1–8
15. Yao J, Miwa S, Shimada H, Tomita S (2007) Optimal pipeline depth with pipeline stage unification adoption. SIGARCH Comput Archit News 35(5):3–9. doi:10.1145/1360464.1360470
16. Vijayalakshmi S, Aniket S, Sudeep C (2014) Reducing power dissipation in multi-core processors using effective core switching. IJCIT 3(6):1435–1442
17. Wang A, Chandrakasan A, Kosonocky S (2002) Optimal supply and threshold scaling for subthreshold cmos circuits. In: Proceedings IEEE Computer Society Annual Symposium on VLSI, 2002, pp 5–9
18. Calhoun B, Wang A, Chandrakasan A (2005) Modeling and sizing for minimum energy operation in subthreshold circuits. IEEE J Solid State Circuits 40(9):1778–1786
19. Rabaey J (2009) Low power design essentials, 1st edn. Springer Publishing Company, Incorporated
20. Taur Y, Ning TH (2009) Fundamentals of modern VLSI devices, 2nd edn. Cambridge University Press, New York
21. Kim NS, Austin T, Blaauw D, Mudge T, Flautner K, Hu JS, Irwin MJ, Kandemir M, Narayanan V (2003) Leakage current: Moore's law meets static power. Computer 36(12):68–75. doi:10.1109/MC.2003.1250885
22. Herbert S, Marculescu D (2007) Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In: Proceedings of the 2007 International Symposium on Low Power Electronics and Design, ser. ISLPED '07. ACM New York, pp 38–43. (Online). doi:10.1145/1283780.1283790
23. Tsai Y, TPS (2005) University, Tools and Techniques for Leakage Power Analysis. Pennsylvania State University. (Online). https://books.google.com/books?id=W89BjkKz7hMC
24. Boucaron J, Coadou A (2009) Dynamic variable stage pipeline: an implementation of its control. INRIA, Rapport de recherche RR-6918, rR-6918. http://hal.inria.fr/inria-00381563/PDF/RR-6918.pdf
25. Joseph KG, Sharkey J, Ponomarev D (2005) Abstract M-SIM: a flexible, multithreaded architectural simulation environment. Tech. Rep