Energy per Instruction Trends in Intel® Microprocessors

Ed Grochowski, Murali Annavaram
Microarchitecture Research Lab, Intel Corporation
2200 Mission College Blvd, Santa Clara, CA 95054
edward.grochowski@intel.com, murali.m.annavaram@intel.com

Abstract

Energy per Instruction (EPI) is a measure of the amount of energy expended by a microprocessor for each instruction that the microprocessor executes. In this paper, we present an overview of EPI, explain the factors that affect a microprocessor's EPI, and derive a historical comparison of the trends in EPI over multiple generations of Intel microprocessors. We show that the recent Intel® Pentium® M and Intel® Core™ Duo microprocessors achieve significantly lower EPI than what would be expected from a continuation of historical trends.

1. Introduction

With the power consumption of recent desktop microprocessors having reached 130 watts, power has emerged at the forefront of challenges facing the microprocessor designer [1, 2]. The goal of modern microprocessors is to deliver as much performance as possible while keeping power consumption within reasonable limits.

Energy per instruction (EPI) is a measure of the power efficiency of a microprocessor. It records the average amount of energy expended per instruction processed by the microprocessor. EPI is measured in Joules/instruction. EPI is related to other commonly-used power-efficiency metrics *performance/watt* and *MIPS/watt*. Specifically, EPI is the reciprocal of IPS/watt. This relationship is shown in the following equation:

$$\frac{\text{Joules}}{\text{Instruction}} = \frac{\frac{\text{Joules}}{\text{Second}}}{\frac{\text{Instructions}}{\text{Second}}} = \frac{\text{Watt}}{\text{IPS}}$$

MIPS/watt is the ratio between two rates: the rate at which the CPU is processing instructions, and the rate at which energy is being expended. Because they are comparing two instantaneous rates (throughput performance and power), MIPS/watt and EPI are ideal metrics for assessing power-efficiency in environments where throughput performance is the primary objective. In order to deliver high throughput performance within a fixed power budget, a microprocessor must achieve low EPI.

It is important to note that MIPS/watt and EPI do not consider the amount of time (latency) needed to process an instruction from start to finish. Other metrics such as MIPS^2/watt (related to energy•delay) and MIPS^3/watt (related to energy•delay^2) assign increasing importance to the time required to process instructions, and are thus used in environments in which latency performance is the primary objective.

2. What Determines EPI?

Consider a capacitor that is charged and discharged by a CMOS inverter as shown in Figure 1.

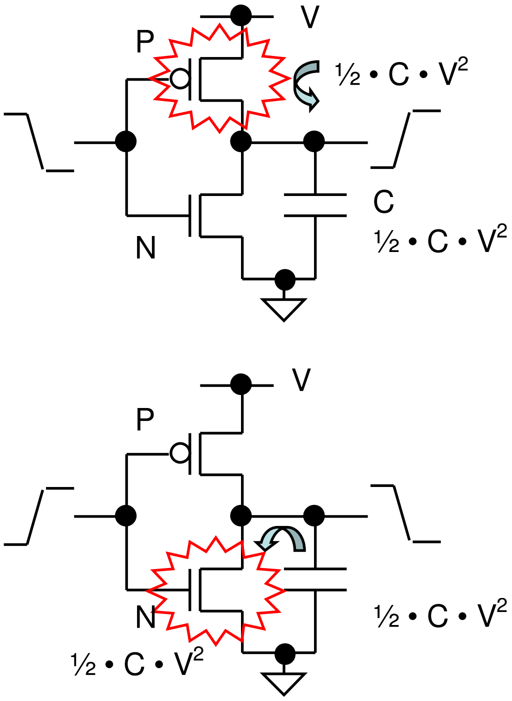

Figure 1: Energy Storage in a Capacitor

During the low-to-high transition of the output, the capacitor is charged with an amount of energy equal to

$E=\frac{1}{2}\bullet C\bullet V^2$. The P transistor dissipates as heat the same amount of energy. During the high-to-low transition of the output, the $E=\frac{1}{2}\bullet C\bullet V^2$ of energy stored in the capacitor is dissipated as heat in the N transistor.

A simple way to comprehend EPI is to consider the microprocessor to be a capacitor that is charged or discharged with every instruction processed. We can apply the capacitor equation:

$$E = \frac{1}{2} \bullet C \bullet V^2$$

to a microprocessor by defining the variables as follows:

E is the energy expended per instruction as the instruction is processed in the microprocessor pipeline from fetch, decode, schedule, execute, to retirement

C is the amount of capacitance toggled while processing the instruction (sometimes referred to as *switching capacitance*. This is equal to activity factor multiplied by total capacitance)

V is the supply voltage

With the capacitor model, energy per instruction depends on only two factors: the amount of capacitance toggled to process an instruction and supply voltage. Note that the capacitor model considers only switching (dynamic) power. In our analysis of Intel microprocessors, we consider both leakage power and dynamic power.

We now consider the EPI of a practical microprocessor. EPI is a function of several factors:

1) Design (microarchitecture, logic, circuits, and layout)
2) Process technology
3) Environment (supply voltage)

Design affects *C*; environment affects *V*; and process technology affects *C* and sets bounds on *V*. Each generation of process technology has reduced both *C* and *V* compared to the prior generation. On the other hand, each generation of design has tended to increase *C* due to the increasing complexity of the microarchitectures. To better understand the effects of design, we would like to separate the contributions due to design from the contributions due to process technology and environment. We do the separation as follows.

Consider two microprocessors, the Pentium® processor and Pentium Pro processor, both built on 0.6 um technology in 1995. The Pentium Pro processor running at 150 MHz delivered 6.08 SpecInt95 and consumed 29.2 watts at 3.1 volts. The Pentium processor running at 100 MHz delivered 3.33 Specint95 and consumed 10.1 watts at 3.3 volts. Note that the 1.5x increase in frequency was entirely a result of the deeper pipeline on the Pentium Pro processor since the process technology was the same for both. The Pentium Pro processor was 6.08/3.33 = 1.8 times faster than the Pentium processor on SpecInt95, and consumed $29.2/10.1\bullet(3.3/3.1)^2$ = 3.3 times the power. This comparison is applicable for *both processors being fabricated on the same process technology and operating at the same supply voltage*. Since energy per instruction is the ratio of power to performance, from these data values we conclude that the Pentium Pro processor consumed 3.3/1.8 = 1.8 times the EPI of the Pentium processor. A similar analysis can be performed for the microprocessors listed in Table 1.

Process	Product	Frequency	Performance	Power (watts)	Voltage (volts)
0.8 um	i486	66 MHz	39.6 SpecInt92	4.9	5
0.8 um	Pentium	66 MHz	77.9 SpecInt92	13	5
0.6 um	Pentium	100 MHz	3.33 SpecInt95	10.1	3.3
0.6 um	Pentium Pro	150 MHz	6.08 SpecInt95	29.2	3.1
180 nm	Pentium III	1.0 GHz	402 SpecInt2K	29.0	1.75
180 nm	Pentium 4 (Willamette)	2.0 GHz	681 Specint2K	75.3	1.75

Table 1: Performance and Power of Intel Microprocessors, 0.8 um to 180 nm

To compute the relative performance and power over more than two generations, we multiply together the ratio of the relevant parameters for the intervening generations. For example, we multiply together the ratio of the performance of the Pentium Pro processor to the Pentium processor by the ratio of the performance of the Pentium processor to the Intel486™ processor to compute an approximate performance ratio between the Pentium Pro and i486™ processors.

For the more recent microprocessors listed in Table 2, a different methodology is needed. The Pentium 4 microarchitecture changed significantly between the 130 nm and 90 nm process technologies, and the older microarchitecture was not fabricated on 90 nm to provide a baseline for comparison. However, since all microprocessors in Table 2 ran the same benchmark (SpecInt2K) and frequency improvements due to process technology were small, it is still possible to conduct a performance comparison between these microprocessors in a process-neutral manner. For a power comparison, we apply an ideal 0.7 scaling factor between the 130 nm and 90 nm technologies (the Pentium 4 microarchitecture remained largely the same on the 90 nm and 65 nm process technologies).

Intel®, i486™, Pentium®, Pentium® III, Pentium® 4, Pentium® M, and Core® Duo™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

To reflect the SpecInt2K benchmark being single-threaded, in the case of the Intel Core® Duo™ processor we consider the power used by only one of the CPU cores and the shared L2 cache. In the Core Duo processor, the power of a single CPU core plus the shared L2 cache is equal to 57% of the rated TDP.

Process	Product	Frequency	Performance	Power (watts)	Voltage (volts)
130 nm	Pentium 4 (Northwood)	3.4 GHz	1342 SpecInt2K	89.0	1.525
130 nm	Pentium M (Banias)	1.0 GHz	673 SpecInt2K	7.0	1.004 ULV
90 nm	Pentium 4 (Prescott)	3.6 GHz	1734 SpecInt2K	103	1.47
90 nm	Pentium M (Dothan)	2.0 GHz	1429 Specint2K	21	1.32
65 nm	Pentium 4 (Cedarmill)	3.6 GHz	1764 SpecInt2K	86	1.33
65 nm	Core Duo (Yonah)	2.167 GHz	1721 SpecInt2K	31	1.3

Table 2: Performance and Power of Intel Microprocessors, 130 nm to 65 nm

Assembling the performance and power data over several generations of Intel microprocessors produces the graph shown in Figure 2.

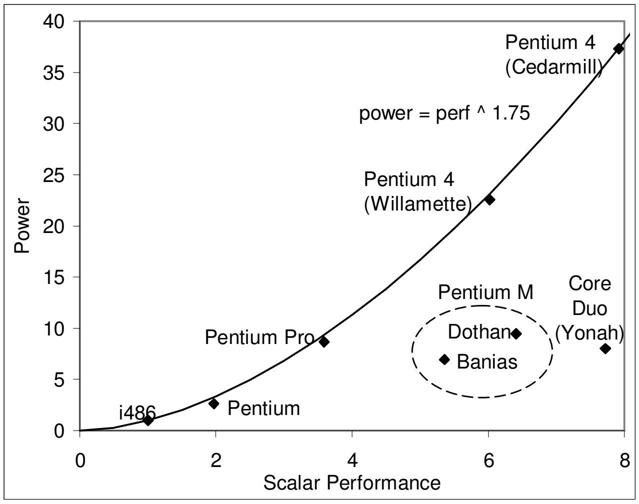

Figure 2: Normalized Power versus Normalized Scalar Performance for Multiple Generations of Intel Microprocessors

In Figure 2, both power and performance have been adjusted to factor out improvements due to process technology over time, and all data have been normalized to the i486™ processor. The results are as if all generations of microprocessors were built on the same process technology. To realize these performance deltas in practice, older microprocessors would need to be given appropriate high-speed memory systems in newer process technologies (i.e. an L2 cache would become necessary since a main memory latency of 5 clocks at 66 MHz would become 80 clocks at 1 GHz).

In Figure 2, the equation power = performance^1.75 has been fitted to the data points corresponding to several generations of desktop microprocessors. The data points corresponding to the Pentium M and Core Duo processors were not included in the fitted curve.

Relative energy per instruction may be computed by taking the ratio of power to performance. Absolute EPI may be computed by noting that the Pentium 4 processor (Cedarmill) achieves a retired IPC of approximately 0.5 on SpecInt2K [3]. Thus, the Pentium 4 processor's EPI is 86/(3.6e9•0.5) = 48 nJ. EPI for all other microprocessors may be computed based on their performance and power ratios relative to the Pentium 4 processor. This data is shown in Table 3.

Product	Normalized Performance	Normalized Power	EPI on 65 nm at 1.33 volts (nJ)
i486	1.0	1.0	10
Pentium	2.0	2.7	14
Pentium Pro	3.6	9	24
Pentium 4 (Willamette)	6.0	23	38
Pentium 4 (Cedarmill)	7.9	38	48
Pentium M (Dothan)	5.4	7	15
Core Duo (Yonah)	7.7	8	11

Table 3: EPI of Intel Microprocessors

3. Analysis

Figure 2 reveals that desktop microprocessors are expending large amounts of power for relatively small improvements in scalar performance. Relative to the i486 processor, the Pentium 4 processor (Cedarmill) delivers approximately 8 times the scalar performance (2.5x the IPC at 3x the frequency), but consumes 38 times more power. These numbers imply that the Pentium 4 processor is spending 5 times the EPI of the i486 processor if both were fabricated on the same process technology and operated at the same supply voltage. The steep relationship between power and scalar performance was first noted by Fred Pollack in his keynote at MICRO-32 in 1999. Fred made the now well-known observation that *we are on the wrong side of a square law.*

The reason for the dramatic increase in power is that the design techniques used in the desktop microprocessors tended to result in much more energy being expended per instruction due to the higher capacitance toggled to process each instruction. Deep pipelines [4], large out-of-order structures to process many in-flight instructions, and mis-speculation all contribute to high EPI.

In contrast to desktop microprocessors, the mobile Pentium M processor [5] and Core Duo processor exhibit much lower EPI. This is a result of more modest pipeline depths, moderately-sized out-of-order structures, aggressive clock gating, and micro-op fusion. The Pentium M and Core Duo processors deliver higher performance by performing more useful work in each clock cycle. This may be seen in the Pentium M processor doubling the SpecInt2K/GHz compared to the Pentium III processor (0.8 versus 0.4). The Pentium M processor employs micro-op fusion to improve IPC and also to amortize the energy costs of out-of-order allocation, dispatch, and retirement over two micro-operations.

As a result of micro-op fusion and other techniques, each core in the Core Duo processor delivers almost 8 times the scalar performance of the i486 processor while consuming only 8 times the power of the i486 processor. Thus, the Core Duo processor achieves roughly the same EPI as the i486 processor! Even though Core Duo is a much higher performance processor, the effective capacitance switched per instruction is roughly the same as the i486 processor. This is a remarkable achievement, one that reverses the trend towards ever-greater EPI as shown in Table 3. Although microarchitectural techniques for improving IPC and improving frequency may both be used to achieve higher performance, improving IPC has emerged as the more energy-efficient of the two techniques.

4. Acknowledgements

The authors would like to thank Jerry Bautista, Jeff McVeigh, Ronny Ronen, and John Shen for providing valuable feedback and suggestions.

5. References

[1] E. Grochowski, R. Ronen, J. Shen, H. Wang. Best of Both Latency and Throughput. In *Proceedings of the 22nd International Conference on Computer Design,* pages 236-243, October 2004.

[2] M. Annavaram, E. Grochowski, J. Shen. “Mitigating Amdahl's Law through EPI Throttling.” In Proceedings of the 32nd International Symposium on Computer Architecture, pages 298-309, June 2005.

[3] B. Davies, J.Y. Bouguet, M. Polito, M. Annavaram, iPART: An Automated Phase Detection and Recognition Tool, Intel Research Tech Report IR-TR-2004-1 (2004). ftp://download.intel.com/research/library/IR-TR-2004-1-iPART.pdf

[4] D. Boggs, et. al., "The Microarchitecture of the Intel® Pentium® 4 Processor on 90nm Technology" in Intel Technology Journal, Q1, 2004. ftp://download.intel.com/technology/itj/2004/volume08issue01/art01_microarchitecture/vol8iss1_art01.pdf

[5] S. Gochman, et. al., "The Intel® Pentium® M Processor: Microarchitecture and Performance" in Intel Technology Journal, Q2, 2003. ftp://download.intel.com/technology/itj/2003/volume07issue02/art03_pentiumm/vol7iss2_art03.pdf

6. Authors

Murali Annavaram is a Senior Researcher at the Microarchitecture Research Lab in Austin, Texas. His current research focuses on squeezing more performance from microprocessors without sacrificing power. In the past he has worked on server workload characterization, trace analysis and full-system simulation infrastructure for server workloads. Prior to joining Intel he spent the best part of his youth at the University of Michigan working with Prof. Ed Davidson on prefetching techniques for databases. He earned his Ph.D. in computer science from the University of Michigan. Murali has four patents pending in microarchitecture and variation-tolerant designs.

Ed Grochowski is a Senior Principal Engineer at the Microarchitecture Research Lab in Santa Clara, California. He joined Intel in 1986 and has had various technical and managerial responsibilities in the Intel® i486™, Pentium®, Pentium® II, and Itanium® microprocessor design teams. Grochowski currently works on the microarchitectural techniques needed for future energy-efficient chip-level multiprocessors. He received both his B.S.E.E. and his M.S.E.E. from the University of California, Berkeley. Grochowski holds over 30 patents in the areas of microarchitecture and logic design.