1.3: Microprocessors for the New Millennium: Challenges, Opportunities, and New Frontiers

Patrick P. Gelsinger

Intel Corporation, Hillsboro, OR

Overview

While Moore's Law appears to be alive and well for the next decade, fundamental power limitations, decreasing instruction-level parallelism, and overall design complexity limit the future of approaches based on its principles. The industry needs to consider new vectors to continue to deliver improvements that keep pace with market trends and provide value to end-users.

Evolution of Microarchitecture: Performance Unimagined

Gordon Moore observed that the total number of devices on a chip doubled every 12 months. He predicted the trend would continue in the 1970s but would slow in the 1980s, when the total number of devices would double every 24 months [1]. Known widely as "Moore's Law," these observations made the case for continued wafer and die size growth, defect density reduction, and increased transistor density as manufacturing matured and technology scaled. It also sparked a revolution in microprocessor architecture innovation and design techniques that would deliver enormous computing power to end-users.

Figure 1.3.1 plots the total number of transistors in Intel microprocessors, starting with the first microprocessor, 4004, to the recent Pentium® 4 microprocessor. The graph shows the total number of transistors on a microprocessor has doubled almost every 24 months, 1.96 years to be exact. Here we consider only "lead" microprocessors—not shrinks and compactions. A lead microprocessor employs a new microarchitecture and has substantially different features, like enhanced and deeper pipelines. The lead microprocessor is subsequently ported to the next-generation process technology by shrinking and compacting the physical layout. These shrinks and compactions take advantage of the increased transistor density; but do not substantially increase the transistors on the die.

To satisfy Moore's Law, the die size of lead microprocessors has been increasing 7% per year, or 14% every two years, or doubling every 10 years. Process technology has indeed scaled as Moore predicted, and defect density has decreased to make these larger dies manufacturable and economical. The operating frequency of theselead microprocessors increased rapidly at first. However, it flattened out in the mid-1980s until the nMOS-to-CMOS technology transition, in the 1990s, when frequency picked up, almost doubling every 24 months.

In 1989, using Moore's Law and extrapolating 20 years of trend data, Intel boldly predicted the microprocessor of the Year 2000, with a projected 50M transistors on a die measuring 1.0 inch square, would operate at over 250MHz and perform over 750M instructions per second (MIPS) [2]. While these projections seemed bold, if not aggressive, history has shown Intel clearly underestimated the potential for enormous gains in both frequency and performance that resulted from advances in process technology, microarchitecture, and design sciences. Microprocessor architects and designers have exploited increased transistor budgets by continually evolving microarchitecture over the last three decades and delivering higher performance. In the last decade alone, the technology scaled from 1.0μm to 0.18μm, and frequency increased by 50x, with about 13x of the frequency growth attributed to process technology and 4x to microarchitecture advances. Microprocessor performance, on the other hand, grew by 75x, as shown in Figure 1.3.2, out of which 6x growth is due to advances in microarchitecture exploiting higher transistor budgets.

The economic and physical challenges of maintaining Moore's Law continue to exist – modern fabs cost well over $2B and ever-finer geometries push processes to their limits. However, these challenges, as difficult as they appear, do not fundamentally deter Moore's Law for at least the next decade or before the 0.03μm generation of technology.

Challenges: No Longer "Business As Usual"

Microarchitectures are only becoming more complex and intricate with more breakthroughs in performance and scaling. If the trends of the last 30 years continue into the coming decade [3] – i.e., transistors on lead microprocessors double every two years, die size grows by 14% every two years, supply voltage scales meagerly, and frequency doubles every two years – then what does it mean for power consumption? Expected power consumption of such microprocessors through 2008 is illustrated in Figure 1.3.3. Notice these predicted powers are excessive and prohibitively large for any practical application, and it is clear total power consumption will become a limiting factor. Therefore, "business as usual" will not work in the future, and alternatives must be evaluated to continue to deliver computational performance.

Advances in microarchitecture, such as superscalar and superpipelined designs and out-of-order and speculative execution, delivered a 6X performance increase in the last decade alone, but at what cost? Pollack's Rule determines performance efficiency of microarchitectures, we apply Pollack's Rule [4]. Fred Pollack observed that in the same process technology, a lead microprocessor provides 1.4x performance but consumes 2x area over the previous generation microprocessor. Figure 1.3.4 plots growth in performance and growth in area of the lead microprocessor and a compaction of the previous generation microprocessor. For instance, in the 1.0μm generation, this is the ratio of the performance and area of the 80486 and 80386 compaction. Growth in performance in a new microarchitecture is declining in subsequent generations and is approximately the square root of the growth in area of the new microarchitecture in any given technology generation. Operation is on the wrong side of a square law!

At the platform level, external bus frequencies have not kept up with processor frequencies as shown in Figure 1.3.5. Microprocessor frequency has increased by 50x, but external bus frequency has increased less than 10x. This frequency gap would have caused an imbalance of computational rate and data bandwidth to the external memory, but bus width kept increasing the bus to improve total bandwidth. It also increases latency to the external memory, and at the platform level, main memory latency has remained almost constant even as internal processor frequencies have increased 50x and clock count per instruction has decreased. This results in the "opportunity cost" of a single main memory read growing substantially during this period – from a few instructions to hundreds of instructions.

It becomes extremely challenging to fill an execution pipeline this deep, and fewer and fewer applications have the necessary parallelism to keep the processor busy as it waits for main memory data. A growing disparity emerges between the potential maximum performance of the processor and the delivered performance when accessing main memory, which inevitably results in capping performance and tends toward a worsening of the performance vs. power metric. This problem is not for external memory alone; even internal caches exhibit similar behavior. Here we have looked at

 0-7803-6608-5/$10.00

the performance disparity of CPU execution vs. memory speed. Similar trends and increasingly large gaps exist for all forms of I/O, in particular disk and communications, for which caching is even less effective than for main memory.

Managing logic and circuit complexity is the biggest future design challenge. Logic complexity continues to increase as transistor integration capacity increases, and circuit design complexity will also increase due to the nature of deep sub-micron transistors. In the past, logic synthesis improved logic design productivity by elevating the level of abstraction. Similarly for circuit design, standard cells and cell-based designs raised the level of reuse. Simple extrapolations from Moore's law will lead to designing chips of greater than 1B transistors in the near future. Employing techniques like formal verification will become essential to improve coverage of logic validation, but this may not be enough. Designs will approach a level of system complexity requiring a new level of design abstraction and verification to design at least part of the chip, if not the full chip.

Opportunities: Architecture Innovation

There are several options to continue improving performance in the future. First, Pollack's Rule indicates that general-purpose logic performance is expensive in terms of power. Hence, it is more efficient to employ special-function logic blocks to deliver application-specific MIPS. We already have movement on these fronts. Several ISAs incorporate both integer and floating point SIMD instructions. For example, Pentium®, Pentium II, Pentium III and Pentium 4 processors employ MMX, ISSE, and ISSE II instructions via application-specific hardware. This hardware exploits inherent parallelism in the human interface and communication kernels. Looking forward, three classes of applications are becoming increasingly important – human interface, knowledge processing, and communications.

Fortunately, these three contain a high degree of inherent parallelism, that may be exploited further with dedicated instructions or increasingly special-purpose hardware engines working in parallel to deliver the performance for these intensive applications with lower power dissipation than general-purpose instruction processing. This is especially true for desktop and mobile computers, which are performance hungry, must be lower power, and need to fit into smaller form factors. With future silicon budgets approaching 1B transistors, integrating other platform components, such as memory and graphic controller, with the microprocessor needs to be considered.

Second, maximum power is consumed when the execution pipeline is full and all superscalar units are busy. Designers have to devise thermal and power delivery solutions for this maximum power. However, few applications spend any significant time operating at maximum power. The wider and deeper is the execution core, the greater is the inefficiency. Thus, with power constraints, focus needs to be on techniques that increase execution core efficiency and add only additional logic and power. One such solution is to employ multiple CPUs on a single die, as shown in Figure 1.3.6. Multiple CPUs provide near linear performance with die size on transaction workloads such as TPC-C, versus a single CPU with more advanced microarchitecture, as predicted by Pollack's Rule. Also, a large shared L2 cache is more performance efficient than two separate small L2 caches. It is extremely unlikely both CPUs will be at maximum power at the same time, and such a condition can be detected and throttled. It can simplify interconnect in a symmetric multiprocessing (SMP) system. This approach is more suitable for server class computers, which typically run thread or transaction workloads and other aspects, such as reliability, are of importance equal to that of single-execution peak performance.

Third, consider multi-threaded architecture, where a single CPU is augmented to look like two or more CPUs to software. It adds ~10% logic to the CPU design, increasing maximum power by < 10%, but can increase throughput by 30+%. Multi-threading also helps address increasing overhead of cache misses. When the CPU waits for a piece of data from external memory or a slower cache, the hardware is used by another thread.

Fourth, cache can be increased on the die. As noted previously, external cache memory misses are expensive in terms of instruction cost. By further increasing the size of the on-die cache further, one could improve performance. Also, memories consume an order of magnitude less power than logic, and thus help increase performance for considerably less power. Further, if DRAM technologies are used rather than SRAM, larger caches at even lower power are possible.

New Frontiers: Human- and Knowledge-Based Computing on the Internet

In the last few decades, the industry has pioneered significant advances in computer technology, architecture, design, manufacturing, and software. However, the fundamental computing model has not changed. It is still machine-based, and moving from machine-based to human-based computing models will be necessary in this decade to advance the use of computing power. For example, when the PC is used to do the mundane task of word processing, the software defines commands and directives, which humans have to give to the computer. In the case of human-based computing, the computer recognizes speech, pens, and gestures and is able to understand human commands, which are subjective at times, but nevertheless interpreted correctly by the computer in the given context.

This paradigm shift will also require "knowledge computing" rather than "data computing." Today's paradigm interprets data given to the application rather than the knowledge associated with the data. In knowledge-based computing, the data is incidental to support the knowledge. For example, XML raises the level of abstraction compared to HTML with tags, context, and relevance. Both human-based and knowledge-based computing will become pervasive in the coming decade as speech, natural-language recognition, and computer vision technologies become more mature and widespread. Undoubtedly, all these technologies will require tremendous computing power, which will be available with the advent of faster and function-rich microprocessors fabricated in the deep sub-micron technologies.

Connectivity will become standard on future computing devices. For wired connectivity, the Internet will continue its rapid increase in backbone bandwidth. More importantly, edge-of-the-network bandwidth will markedly increase as broadband technologies, such as cable, xDSL, satellite, and supporting networking hardware and software become standard. More important than wired connectivity will be rapid transitions in wireless communications. These will come as personal area networking such as Bluetooth, local area networks such as 802.11, and wide area networks such as GPRS and 3G are standardized and required features of all computing platforms. This communications renaissance is uniquely enabled by the silicon performance technologies described above.

Connectivity and mobility are key drivers of new computing models that will emerge in the years to come. In the future, peer-to-peer computing will move the Internet from a hierarchical client and server model to a dynamic, self-forming distributed computing network that will fully utilize the installed network of bandwidth and edge-computing resources. This will enable entirely new classes of distributed applications that were eco-

nomically impractical in the past. Further, many low-power, power-efficient "dumb devices" (e.g., a simple IP enabled cell phone) will be connected using wired or wireless networks accessing both peer-to-peer and client-server networks. These devices will be connected to appliance servers, not just in office and home, but also in cars and on the person. Knowledge-based smart computing should be available anytime and anywhere, with seamless connectivity to its wired and wireless peers and servers. The Internet is the backbone and data source for the knowledge base, which means knowledge-based computing requires much higher data rates than those available today.

Summary

Numerous challenges are faced in accelerating the general-purpose performance curve, with power management and design complexity being two key ones. Despite these challenges, breakthroughs in communications, human interface, and knowledge management technologies continue to have promising futures because they exploit inherent parallelism and could use special-purpose performance. There will always be a need for performance, and as was the case in 1989, it is boldly predicted that the microprocessor of the Year 2010 will have 1B transistors on a die, operate at 20 to 30GHz, and perform over 1T operations per second.

Acknowledgments:

The author thanks S. Borkar, B. Colwell, M. Hayward, D. Papworth, F. Pollack, J. Rattner, and I. Young for contributions to this paper. Intel and Pentium are registered trademarks of Intel Corp, and all other brands and names are the property of their respective owners.

References:

[1] Moore, G., "Progress in Digital Integrated Electronics," IEDM, 1975.

[2] Gelsinger, P., et. al., "Microprocessor circa 2000," IEEE Spectrum, Oct. 1989.

[3] Borkar, S., et. al., "Technology and Design Challenges for Low Power and High Performance," ISLPED 1999.

[4] Pollack, F., "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies," Micro32, 1999.

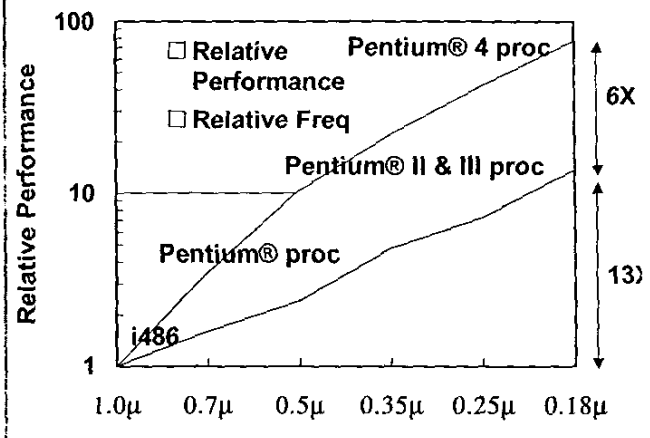

Figure 1.3.2: Growth in Microprocessor Performance.

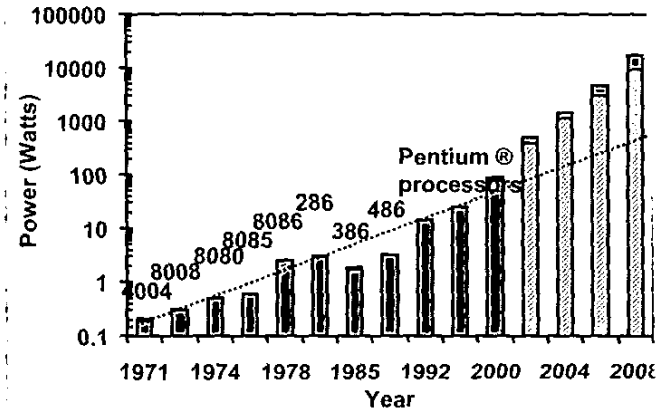

Figure 1.3.3: Lead microprocessor power increases dramatically beyond expected trend.

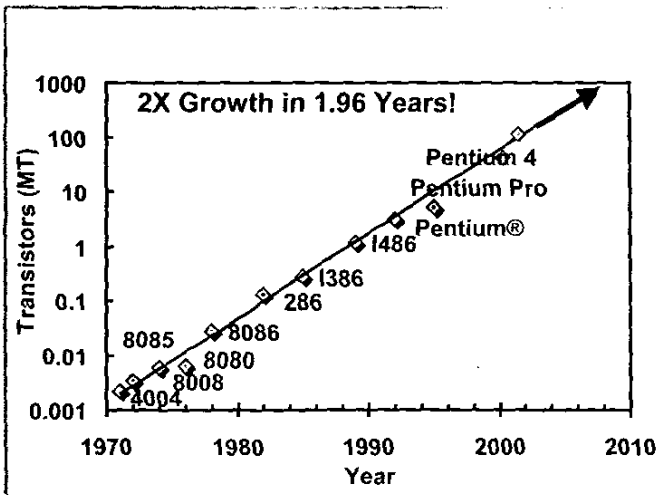

Figure 1.3.1: Moore's Law. Transistors on a chip double every two years.

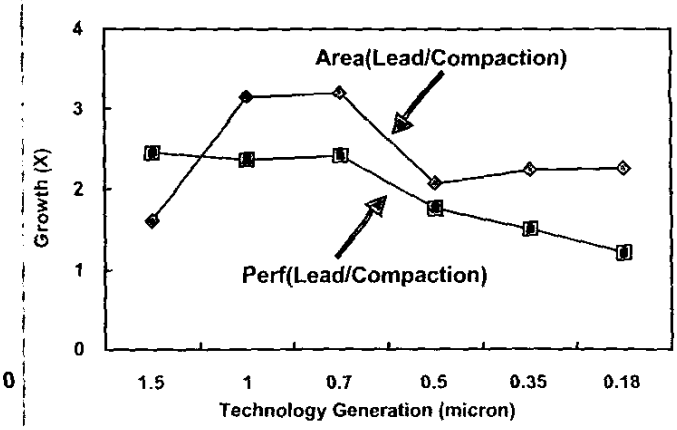

Figure 1.3.4: Pollack's rule.

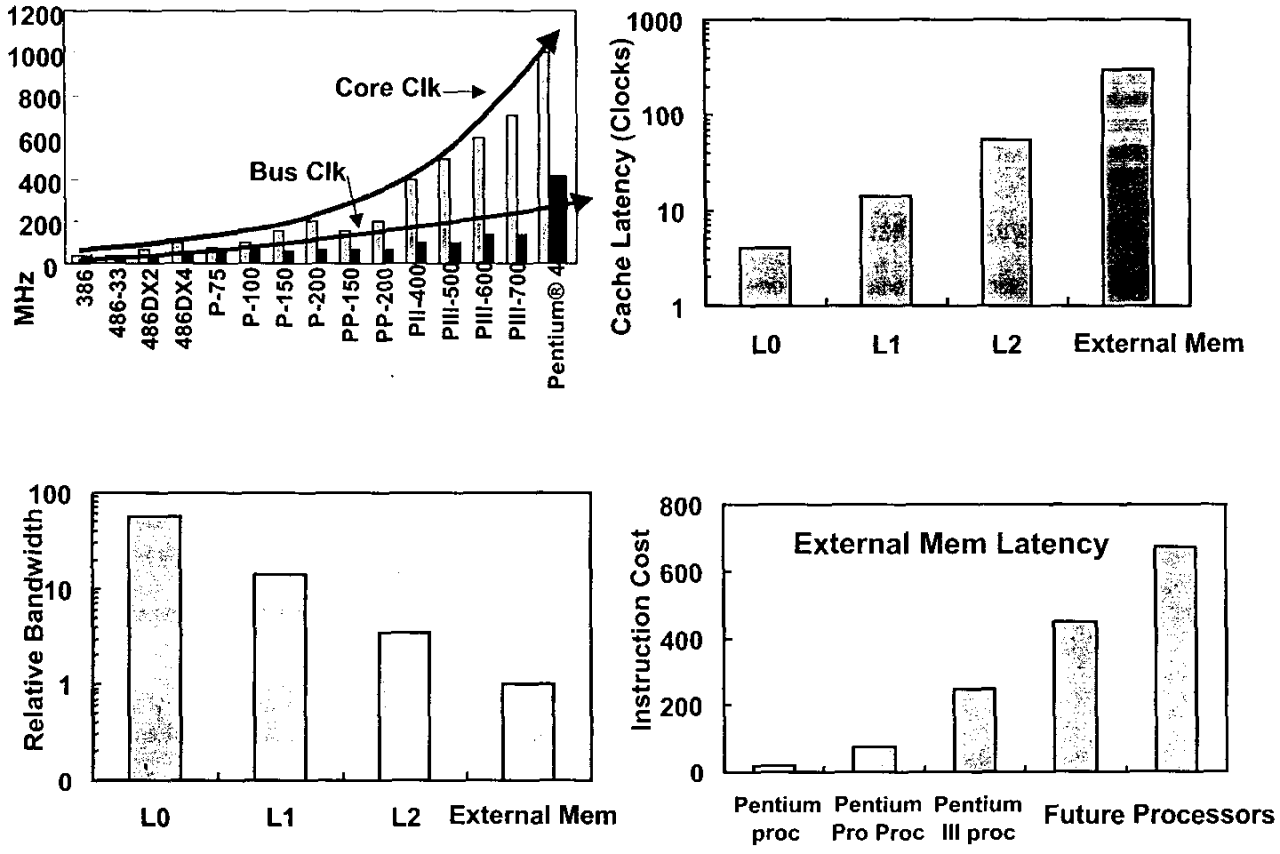

Figure 1.3.5: External bus bottleneck.

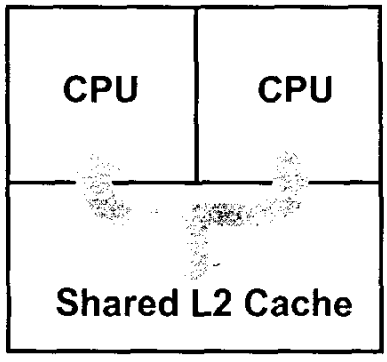

Figure 1.3.6: Multiple CPUs on a single die sharing a large cache.