Intel's Road to Multi-Core Chip Architecture

by Geoff Koch

What is the problem?

In 1965, Intel co-founder Gordon Moore observed an exponential growth in the number of transistors per integrated circuit and predicted that this trend would continue – a prediction today known as Moore's Law. Through Intel's ongoing research and development efforts, the doubling of transistors every couple of years has been maintained for more than 40 years.

For most of that time, Moore's Law made it relatively easy to improve performance. Most average users simply scaled up, relying on applications that took advantage of the increasing capabilities and speed of single processors. Users with the most extreme demands scaled out, turning to applications that divided processing tasks between two or more processors.

Lately, however, scaling up has become more difficult because of several less-friendly laws of physics.

First, memory speeds are not increasing as quickly as logic processing speeds. During the i486 CPU days in the late 1980s and early 1990s, the requirements were 6 to 8 clocks per cycle to access memory. Today's processors require more than 200 clocks per cycle. These wasted clock cycles can nullify the benefits of frequency increases in the processor.

Next, smaller, denser transistors on today's chips need to be threaded together with ever-increasing lengths of wire interconnects. As these interconnects stretch from hundreds to thousands of meters in length on a single processor, path delays crop up that can cancel the speed increases of the transistors.

Perhaps most important is an unsustainable increase in power density. The number of transistors per chip has ballooned in recent years, which presents an engineering challenge since each of these transistors is a working electrical device that consumes power and produces heat.

A 1993-era Intel® Pentium® processor had around 3 million transistors, while today's Intel® Itanium® 2 processor has nearly 1 billion transistors. If this rate continued, Intel processors would soon be producing more heat per square centimeter than the surface of the sun, which is why the problem of heat is already setting hard limits to frequency increases.

However, immutable laws of physics don't necessarily lead to hard limits for computer users. New chip architectures built for scaling out instead of scaling up offer enhanced performance, reduced power consumption, and more efficient simultaneous processing of multiple tasks.

The Intel multi-core chips available today embrace the scale-out approach to performance. This architecture in essence reflects a "divide and conquer" strategy. By splitting the computational work performed by the single execution core in traditional microprocessors among multiple cores, a multi-core processor can perform more work within a given clock cycle.

The result is a better overall user experience, and an extension of Moore's Law well into the future.

Why is it important?

Some people say that extending Moore's Law into the future isn't necessary, and that today's computer hardware and software is good enough. This is a dubious notion given the history of the information technology industry. All too often, statements about good-enough computing capabilities, or innovations that will never find a market in the first place, are proven wrong – no matter who's making the claim.

In the 1970s, for example, an Intel engineer came to Gordon Moore with an idea for a computer that would be used in the home. Moore asked what it would be good for, and all the engineer could think of was that such a device might be used by housewives to store recipes. Moore, who couldn't imagine his wife with her recipes on a computer in the kitchen, declared the idea not to have any practical application at all.

Today, there are more than half a billion PC users worldwide doing lots more than storing recipes.

Mobile users of notebook PCs and handheld devices want all this mobile gadgetry to work together better and be easier to use, especially when it comes to rich media applications involving music, photos and videos.

Business users of software for 3D modeling, scientific calculations or high-end digital content creation have their own growing list of performance-heavy requirements, as well. And in a world where cost counts and competition is global, the business crowd increasingly is seeking computing and communications infrastructure platforms for end-to-end solutions in businesses.

With PCs moving out of the den and into the living room, digital-home users are seeking computing and communications platforms that emphasize entertainment applications and easy interoperability with consumer electronics devices.

Finally, with health care costs rising in most industrialized countries and the post-war baby boomers aging into their retirement years, a huge digital-health market is emerging. Problems in health-care research, diagnostics and productivity, and personal health care are ripe for new solutions based on increasingly powerful computing platforms and applications.

For all these users, performance means more than wringing additional benefits from a single application.

First, there's the issue of power consumption and energy costs, dominant themes for policymakers, business IT managers and even families looking to trim their monthly expenses. In a *New York Times* article published Feb. 10, 2005, reporter Ian Austen wrote that "[w]ith their increased processing capabilities, computers today have become power gluttons." Austen went on to cite an estimate by the U.S. Environmental Protection Agency which states that operating a typical computer in power-saving mode would result in an annual savings of $100 per computer.

Then there's the issue of the evolving, interrupt-driven computer usage pattern. Users increasingly multitask, actively toggling between two or more applications, or working in an environment where many background processes compete with each other and with open applications for scarce processor resources.

In a 2002 Intel-sponsored Harris Interactive survey, 76 percent of computer users said they multitask at least occasionally or frequently on the PC, and nine out of 10 users reported that they've experienced problems while trying to perform multiple processor-intensive tasks on a PC. Problems include computer freezes, time lags, function shut-downs, choppy screens and audio distortion. Nearly 60 percent of survey respondents said they feel bored when a computer function makes them wait, so they do something else on the computer at the same time.

Aside from multitasking, users are faced with handling vastly growing quantities of application data. The amount of data in existence is doubling every 24 months – a sort of data equivalent to Moore's Law. Worldwide, data growth is increasing so quickly it's now measured by the exabyte. That's 10^{18} bytes, or a billion billion bytes. Other technologies associated with this explosion of data – storage, networking, input/output (I/O) and database – are advancing rapidly. But without another leap forward in processing capabilities, the opportunity to harvest and take advantage of this wealth of data will be missed.

No one expects a leap forward in processors' core execution engines, which are already at the edge of the manufacturing envelope. In a few years, these processor cores may have larger cache sizes and at least slightly faster clock speeds than those on today's chips. But because of physics – and engineering-related challenges, these advances will be hard won and won't keep pace with performance demands. Inevitably, meeting new demands will require a scale-out approach that divvies up computing tasks among multiple processing cores.

One way to do this divvying up is to build single computers that are powered by multiple processors. For years this has been the norm in high-performance computing and supercomputing, two industry segments in which single machines often contain hundreds or even thousands of processors.

But adding even one additional processor isn't always practical for mainstream PC users. Because of the additional socket and associated chips and circuitry, motherboards for dual-processor systems are more expensive than their single-processor counterparts. Also, since the processor is often the most expensive single component in a given computer system, trying to achieve scale-out performance by swapping a single-processor machine for a multiprocessor machine can be expensive.

What is the solution?

Another scale-out option, and the one Intel has embraced in its move to multi-core architectures, is to add two or more brains to each processor. Explained most simply, multi-core processor architecture entails silicon design engineers placing two or more execution cores – or computational engines – within a single processor. This multi-core processor plugs directly into a single processor socket, but the operating system perceives each of its execution cores as a discrete logical processor with all the associated execution resources.

Multi-core chips do more work per clock cycle, and thus can be designed to operate at lower frequencies than their single-core counterparts. Since power consumption goes up proportionally with frequency, multi-core architecture gives engineers the means to address the problem of runaway power and cooling requirements.

Multi-core processing may enhance user experience in many ways, such as improving the performance of compute- and bandwidth-intensive activities, boosting the number of PC tasks that can be performed simultaneously and increasing the number of users that can utilize the same PC or server. And the architecture's flexibility will scale out to meet new usage models that are bound to arise in the future as digitized data continues to proliferate.

Any application that will work with an Intel single-core processor will work with an Intel multi-core processor. However, to make the most of a multi-core processor today, the software running on the platform must be written such that it can spread its workload across multiple execution cores. This functionality is called thread-level parallelism, or "threading," and applications and operating systems that are written to support it (such as Microsoft Windows* XP, Windows* Server and various Linux* vendor offerings) are referred to as "threaded" or "multithreaded."

A multi-core processor can execute multiple and completely separate threads of code. This can mean that one thread runs from an application and a second thread runs from an operating system, or parallel threads run from within a single application. Multimedia applications are especially conducive to thread-level parallelism because many of their operations can run in parallel.

Despite the recent attention focused on multi-core architecture, Intel has been delivering threading-capable products for more than a decade. By 1994, the Intel Pentium processor already featured instruction-level parallelism, an architectural feature that extracted instructions in a single thread of code, executed the instructions in parallel, and then recombined them in the same order.

That year, Intel added “glue-less” dual-processing capability – two full processors that plugged into two board sockets – to provide a hardware-enhanced threaded environment for servers and workstations. The company expanded its efforts in 1995, providing glue-less multiprocessing capability with the introduction of the Intel® Pentium® Pro processor. The Intel Pentium Pro processor enabled the seamless connection of as many as four processors on a single board, providing servers and workstation-class products with the means to attain higher compute throughput in threaded software environments.

These efforts provided a springboard for delivering higher degrees of thread-level parallelism in a single processor on volume platforms. In the early part of the 2000s, Intel introduced Hyper-Threading Technology (HT Technology) into its Intel NetBurst® microarchitecture (for Intel® Pentium® 4 and Intel® Xeon® processors) as an innovative means to deliver higher thread-level parallelism on volume platforms. HT Technology enables processors to execute tasks in parallel by weaving together multiple threads in a single-core processor.

By fall 2004, HT Technology had shipped on well over 50 million Intel Pentium 4 products for desktops, servers and mobile PCs, offering new incentive for software developers to design applications capable of processing information in parallel for greater efficiency.

Intel has been offering multi-core processors since 2005. At the spring 2006 Intel Developer Forum event in San Francisco, the company disclosed details of the Intel® Core™ microarchitecture, the industry-leading foundation for Intel’s multi-core server, desktop and mobile processors. Intel Core microarchitecture products built with advanced 65 nanometer process technology deliver higher-performing, yet more energy-efficient processors that spur more stylish, quieter and smaller mobile and desktop computers and servers. Likewise, these new machines can reduce electricity- and real estate-associated costs, and provide critical capabilities such as enhanced security, virtualization and manageability for consumers and businesses.

With Intel Core microarchitecture, each core is equipped with a nearly complete set of hardware resources, including cache memory, floating point and integer units, etc. One programming thread can utilize all these resources while another thread can use all the hardware resources on another core. The same programming techniques that for years have been used to write threaded applications for multiprocessor systems, or more recently for platforms based on HT Technology, can be used to take advantage of multi-core processors.

Thanks to $28 billion in R&D spending through the tech downturn from 2001 to 2003, Intel has the manufacturing capacity in place to quickly increase its production of multi-core chips. This investment provides the foundation of Intel’s twin 2006 transitions – to a new 65-nanometer process technology in several of its fabrication facilities and to a powerful new microarchitecture. By the end of 2006, Intel Core microarchitecture will be at the heart of its PC and server platforms.

Further on the horizon, Intel plans to deliver additional processors with two or more cores for mobile, desktop and server platforms. At the spring 2005 Intel Developer Forum in San Francisco, Intel described goals to mass-produce chips with a hundred or more processing cores. Applications will have to be multithreaded to fully exploit these architectural innovations. The alternative – a single-threaded application using just $1/100^{th}$ of a processor's throughput – would likely be a tough sell, no matter what users are demanding in the future.

Intel's current multi-core products and platforms

Intel has more than 20 multi-core projects underway throughout the company's high-end server, volume server, workstation, desktop, mobile, and networking platform families. Six of these processors have already started shipping, while the balance is expected to start shipping over the course of 2006, 2007 and beyond. Among the recent highlights:

Desktop Multi-core – In the second quarter of 2005, Intel delivered the dual-core Intel® Pentium® processor Extreme Edition for PC enthusiasts and the dual-core Intel® Pentium® D processor for mainstream PC users.

Mobile Multi-core – Intel introduced its first mobile optimized dual-core processor, the Intel® Core™ Duo processor, in the first week of January 2006. The Intel Core Duo processor, codenamed Yonah, is an integral part of the Intel® Centrino® Duo mobile-technology based platform (codenamed Napa).

Server Multi-core – In the fourth quarter of 2005, Intel delivered both its first dual-core

Intel Xeon processor for dual processor servers, codenamed Paxville, as well as its first dual-core Intel Xeon processor for servers with four or more processors, previously codenamed "Paxville MP."

For more information, see "INTEL MULTI-CORE FACTS, FIGURES AND DECODER RING," a fact sheet prepared for the spring 2006 Intel Developer Forum.

Intel's multi-core evaluation program

Intel's evaluation program, which began Aug. 15, 2005, ultimately will deliver thousands of dual-core platforms based on Intel Pentium D processors, Intel Xeon processors, Intel Xeon processors MP, and Intel Itanium processors to early-adopter customers and software developers well into 2006.

IT evaluation cycles often take six to nine months, and evaluation systems are critical for IT managers to begin testing new technologies as early as possible. Having access to pre-production and production systems will allow IT managers to evaluate performance, test compatibility with in-house applications and determine future deployment plans.

To help software developers tune their applications to take full advantage of a multi-threaded processing environment, the evaluation systems also ship with a full suite of Intel® Software Development Products consisting of threading tools, compilers, performance analyzers and performance libraries.

Additional information is available at www.intel.com/software.

Multi-core platforms and content fuel digital entertainment

At the January 2006 Consumer Electronics Show in Las Vegas, Intel President and CEO Paul Otellini unveiled two platforms and several content alliances that provide the foundation for new experiences from digital entertainment and wireless laptops – and include the new high-performance, low-power Intel Core Duo processor.

Noting the transformation now underway in entertainment, Otellini introduced the company's new home platform – Intel® Viiv™ technology, some versions of which utilize multi-core processors – and several commitments from top U.S. and international entertainment companies including AOL, DIRECTV, NBC Universal, Turner Broadcasting's GameTap, ESPN, Televisa and Eros. ClickStar announced its first feature film, "10 Items or Less," with a plan for an Internet premiere within weeks of its theatrical release – an industry first. These and other developments will bring millions of songs, movies, programs and games to the PC in 2006.

Intel Viiv technology-based entertainment PCs will help make it easier for families to download, store, view, manage and share all kinds of digital entertainment and information on a choice of TV, PC, laptop and handheld viewing screens.

Intel Centrino Duo mobile technology improves performance and battery life for the fast-growing wireless laptop market segment. Otellini also introduced the groundbreaking Intel Core Duo processor – powerful dual core silicon supporting the Intel Centrino Duo and certain Intel Viiv technology models. The processor is well equipped to deliver performance-per-watt efficiency and sleek designs for entertainment PCs, notebooks and CE-like devices.

More information on Intel Viiv technology is available here; more on Intel Centrino Duo is available here.

Intel Senior Fellow on multi-core architecture

Richard Wirt is an Intel Senior Fellow, corporate vice president and general manager of the Software and Solutions Group. His job is making sure that the world's software takes full benefit Intel's newest platform features and technologies, including multi-core architecture. As an Intel Senior Fellow, he is also considered a top technical achiever within the company.

In 2005, Wirt spoke to a worldwide group of Goldman Sachs developers. The topics addressed included the transition to multi-core processors, the lineup of Intel's software development tools and the move to parallelism through threading and clustering. At the Intel Development Forum Spring 2006 event, Wirt revisited these topics in a software-focused lecture entitled "Harnessing the Capabilities of Intel Platforms."

Highlights of Wirt's Godman Sachs presentation, excerpted by the Intel® Software Network, are available here; a Webcast of Wirt's IDF lecture is available here.

Multi-core gaming

In November 2005, Intel helped inaugurate the era of multi-core PC gaming at the Intel-sponsored Cyberathlete Professional League (CPL) World Tour Grand Finals in New York City. The event brought together some of the best players, hardware and software.

Intel's multi-core processors, such as the Intel® Pentium® processor Extreme Edition, let software developers add features to their games that create more realistic experiences. The ability to run different threads – or components – of the game in parallel can not only make the game run faster, but can also make it more engaging.

Multi-core gaming titles are listed at www.intel.com/go/gaming; read how Intel® Threading Tools helped Toronto-based Alias – the company whose tools are behind blockbuster films such "The Lord of the Rings" trilogy and the "Star Wars" prequels – optimize its Maya* digital software in this case study.

Threading tools to help Apple developers

On Jan. 11, 2006, Intel announced new software development tools and resources, available through its Intel® Software Network, to help Apple developers take advantage of Intel Core Duo processor technologies and to maximize application performance on the new Intel-based Mac platforms.

The special beta versions of the Intel® Fortran Compiler, Intel® C++ Compiler, Intel® Math Kernel Library and Intel® Integrated Performance Primitives are available now. Intel also will provide other resources to assist with software optimization, dual-core threading and migration information.

More information on these programs is available at www.intel.com/software/apple.

About the Author

Geoff Koch is a science and technology journalist in Lansing, Mich. His articles on writing code for cellular or handheld devices include *Culture: The Next Big Thing in Code* and *One Plea for More Open Cell Phone Platforms*.

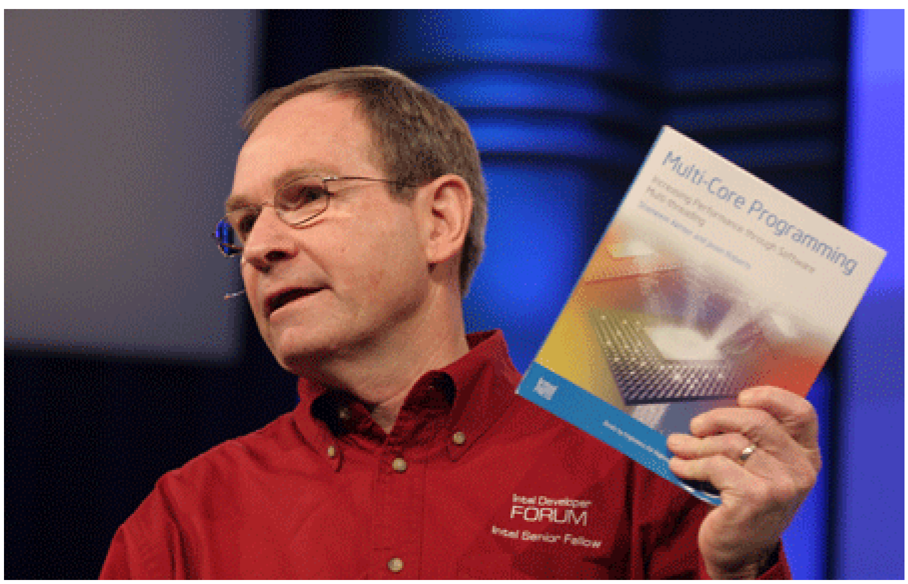
Additional Resources

The Intel® Software Network is a collection of software development products, tools, training and expert advice designed to help software developers bring more innovative products to market even faster on Intel platforms. Whether developing software for mobile, home, office or enterprise computing, the Intel Software Network is a portfolio of resources for the developer to gain access to information about Intel platforms and technologies. Visit

www.intel.com/software for more information.

Intel's Pat Gelsinger, senior vice president, general manager, of Intel's Digital Enterprise Group, holds 12 of Intel's new dual core chips during his keynote at the Intel Developer Forum Spring 2006.

Intel's Justin Rattner, chief technology officer, talks about Intel's new improved platform idle efficiency mother boards during the Intel Developer Forum Spring 2006.



Intel's Richard Wirt holds a copy of Intel's new Multi–Core Programming book as he talks to the large crowd attending the Technology Insights at the Intel Developer Forum Spring 2006.