Platform 2015: Intel® Processor and Platform Evolution for the Next Decade

Shekhar Borkar
Pradeep Dubey
Kevin Kahn
David Kuck
Hans Mulder
Steve Pawlowski
Justin Rattner

Table of Contents

(Click on page number to jump to sections)

Platform 2015: Intel® Processor and Platform Evolution for the Next Decade

By Shekhar Borkar, Pradeep Dubey, Kevin Kahn, David Kuck, Hans Mulder, Steve Pawlowski, Justin Rattner

Overview: What's Over the Digital Horizon?

Without a doubt, computing has made great strides in recent years. But as much as it has advanced in the last 10 years, in the coming decade, the emergence and migration of new workloads and usage models to mainstream computing will put enormous demands on future computing platforms: demands for much higher performance, much lower power density and greatly expanded functionality.

Given these seismic shifts in the uses of computing, how we define and architect future computing platforms will have to change dramatically, holistically comprehending and satisfying not only computation, but interface and infrastructure requirements as well. Intel's long-range vision for the evolution of these three fundamental platform elements, and the architectural innovation and core competencies driving that evolution, is what we call Platform 2015. This article focuses on the computational component of Platform 2015 and how we plan to get there over the coming decade. To read the more comprehensive overview of Intel's Platform 2015 vision, read the Platform 2015 white paper on the Intel Web site.

Microprocessor Architecture in 2015: The Intel Roadmap

Looking ahead, Intel processors and platforms will be distinguished not simply by their raw performance, but by rich and diverse computing and communications capabilities, power-management, advanced reliability, security and manageability and seamless interaction with every other element of the platform.[1] Intel's roadmaps encompass the following key characteristics of processor architecture circa 2015, or "Micro 2015" for short:

1. Chip-Level Multiprocessing (CMP)

Intel continues pioneering in one of the most important directions in microprocessor architecture—increasing parallelism for increased performance. As shown in **Figure 1**, we started with the superscalar architecture of the original Intel® Pentium® processor and multiprocessing, continued in the mid-90s by adding capabilities like "out of order execution," and most recently introduced Hyper-Threading Technology in the Pentium 4 processor. These paved the way for the next major step—the movement away from one, monolithic processing core to multiple cores on a single chip.

Intel is introducing multi-core processor-based platforms to the mainstream. These platforms will initially contain Intel processors with two cores, evolving to many more. We plan to deliver Intel processors over the next decade that will have dozens, and even hundreds of cores in some cases. We believe that Intel's chip-level multiprocessing (CMP) architectures represent the future of microprocessors because they deliver massive performance scaling while effectively managing power and heat.

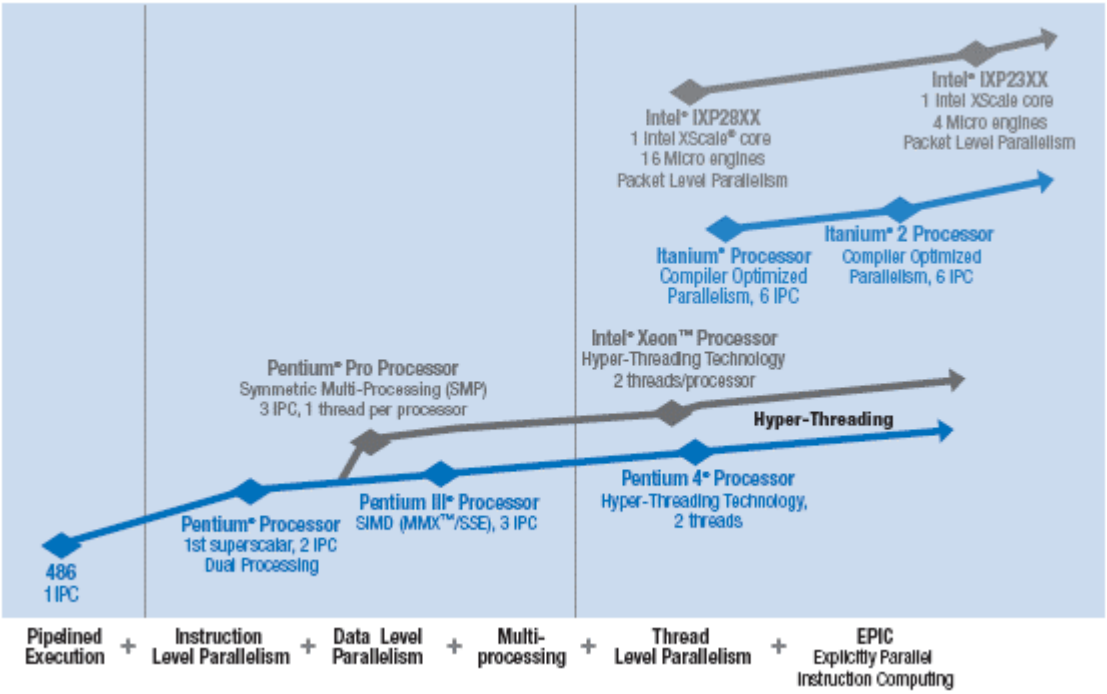

Figure 1. *Driving increasing degrees of parallelism on Intel® processor architectures.*

In the past, performance scaling in conventional single-core processors has been accomplished largely through increases in clock frequency (accounting for roughly 80 percent of the performance gains to date). But frequency scaling is running into some fundamental physical barriers. First of all, as chip geometries shrink and clock frequencies rise, the transistor leakage current increases, leading to excess power consumption and heat (more on power consumption below).

Secondly, the advantages of higher clock speeds are in part negated by memory latency, since memory access times have not been able to keep pace with increasing clock frequencies. Third, for certain applications, traditional serial architectures are becoming less efficient as processors get faster (due to the so-called Von Neumann bottleneck), further undercutting any gains that frequency increases might otherwise buy. In addition, resistance-capacitance (RC) delays in signal transmission are growing as feature sizes shrink, imposing an additional bottleneck that frequency increases don't address.

Therefore, performance will have to come by other means than boosting the clock speed of large monolithic cores. Instead, the solution is to divide and conquer, breaking up functions into many concurrent operations and distributing these across many small processing units. Rather than carrying out a few operations serially at an extremely high frequency, Intel's CMP processors will achieve extreme performance at more practical clock rates, by executing many operations in parallel[2]. Intel's CMP architectures will circumvent the problems posed by frequency scaling (increased leakage current, mismatches between core performance and memory speed and Von Neumann bottlenecks). Intel® architecture (IA) with many cores will also mitigate the impact of RC delays[3].

Intel's CMP architectures provide a way to not only dramatically scale performance, but also to do so *while* minimizing power consumption and heat dissipation. Rather than relying on one big, power-hungry, heat-producing core, Intel's CMP chips need activate only those cores needed for a given function, while idle cores are powered down. This fine-grained control over processing resources enables the chip to use only as much power as is needed at any time.

Intel's CMP architectures will also provide the essential special-purpose performance and adaptability that future platforms will require. In addition to general-purpose cores, Intel's chips will include specialized cores for various classes of computation, such as graphics, speech recognition algorithms and communication-protocol processing. Moreover, Intel will design processors that allow dynamic reconfiguration of the cores, interconnects and caches to meet diverse and changing requirements.

Such reconfiguration might be performed by the chip manufacturer, to repurpose the same silicon for different markets; by the OEM, to tailor the processor to different kinds of systems; or in the field at runtime, to support changing workload requirements on the fly. Intel® IXP processors today provide such capability for special purpose network processing. As shown in **Figure 2**, the Intel IXP 2800 has 16 independent micro engines operating at 1.4 GHz along with an Intel XScale® core. Another related Intel research area is focused on development of a reconfigurable radio architecture, enabling processors to dynamically adapt to different wireless networking environments (such as 802.11b, 802.11a, and W-CDMA).

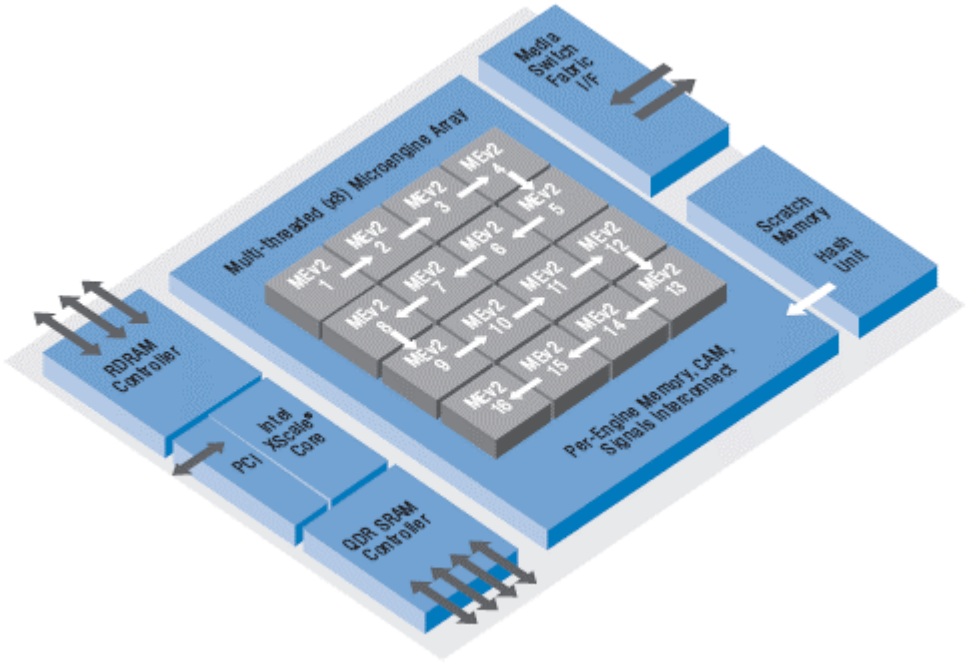

Figure 2. *Block-level diagram of Intel® IXP 2800 with 16 independent micro engines and one Intel XScale® core.*

2. Special Purpose Hardware

Over time, important functions once relegated to software and specialized chips are typically absorbed into the microprocessor itself. Intel has been at the forefront of this effort, which has been the driving force behind our business model for over 35 years. By moving functions on chip, such capabilities benefit from more-efficient execution and superior economies of scale and reduce the power consumption drastically. Low latency communication between special purpose hardware and general purpose cores will be especially critical to meet future processor architecture performance and functionality expectations.

Special-purpose hardware is an important ingredient of Intel's future processor and platform architectures. Past examples include floating point math, graphics processing and network packet processing. Over the next several years, Intel processors will incorporate dedicated hardware for a wide variety of tasks. Possible candidates include: critical function blocks of radios for wireless networking; 3D graphics rendering; digital signal processing; advanced image processing; speech and handwriting recognition; advanced security, reliability and management; XML and other Internet protocol processing; data mining; and natural language processing.

3. Large Memory Subsystems

As processors themselves move up the performance curve, memory access can become a main bottleneck. In order to keep many high-performing cores fed with large amounts of data, it is important to have a large quantity of memory on-chip and close to the cores. As we evolve our processors and platforms toward 2015, some Intel microprocessors will include on-chip memory subsystems. These may be in the gigabyte size range, replacing main memory in many types of computing devices. Caches will be reconfigurable, allowing portions of the caches to be dynamically allocated to various cores. Some caches may be dedicated to specific cores, shared by groups of cores, or shared globally by all cores, depending on the application needs. This flexible reconfigurability is needed to prevent the caches themselves from becoming a performance bottleneck, as multiple cores contend for cache access.

4. Microkernel

Microprocessors will need a sizable dose of integrated intelligence to coordinate all this complexity: assigning tasks to cores, powering up and powering down cores as needed, reconfiguring cores to suit changing workloads, and so on. In Intel's CMP architectures, with their high capacity for parallel execution, the processor itself will have to perform some amount of invisible user-level threading, breaking applications into multiple threads that can be processed simultaneously. One way to efficiently handle all this is through a built-in microkernel, relieving higher-level software of these complicated hardware management tasks.

5. Virtualization

Future microprocessors will need several levels of virtualization. For example, as shown in Figure 3, virtualization is needed to hide the complexity of the hardware from the overlying software. The OS, kernel and the software should not have to deal with the intricacies of many cores, specialized execution hardware, multiple caches, reconfiguration and so on. Rather, it should see the processor as one or more unified virtual machines through its global interfaces—virtualization provides the necessary abstraction. Virtualization will also be used to ensure manageability, reliability and security. For example, the processor can be partitioned into multiple virtual processors, some dedicated specifically to management and security tasks and some to manage applications. Some of these features are already in Intel's roadmap as disclosed recently[4].

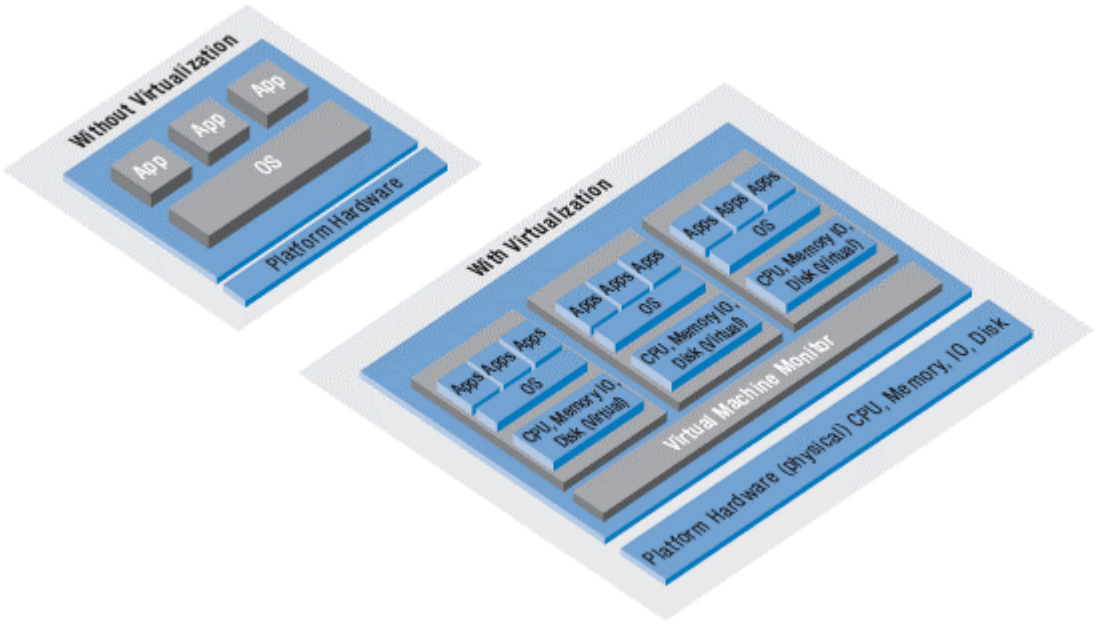

Figure 3. *Intel virtualization scheme providing an abstraction layer to hide hardware complexity from software.*

6. Silicon and Process Technology

Silicon CMOS process scaling is expected to continue at its current pace at least through 2015, and probably beyond. The on-going trend of introducing new materials and new structures will continue. Examples under development are high-k/metal gate [PDF] dielectrics and tri-gate transistors. Farther out, III-V transistors, carbon nanotubes and silicon nanowires are being investigated. All of these have a goal of continuing to improve device speed, maintaining or reducing power consumption, and further dimensional scaling. In addition, the integration between chip architecture and process technology will become ever tighter as we pack billions of transistors onto a single die. This is very critical—the architects and the process technologists will need to work even more closely to develop the future microprocessor and platform architectures.

7. Compatibility and Ecosystem Enabling
Intel's commitment to maintaining compatibility with existing and older IA software even as we advance the performance and capabilities of our processor and platforms is well proven. Intel will deliver future processors and platforms with teraflops of supercomputer-like performance for new applications and workloads without sacrificing compatibility with the installed base of software and the vast developer ecosystem around it.

Meeting the Challenges

What will it take to realize this processor vision for 2015? Some major challenges loom, in both software and hardware. It should be mentioned that raw transistor count is not, in itself, likely to be a major hurdle for Intel. No other company has proven better at delivering to the goals of Moore's Law, and we can confidently predict that Intel processors will pack tens of billions of transistors on a 1-inch square die: enough to support the many cores, large caches and other previously described hardware over the next 10 years. There are, however, other challenges to be met.

1) Power and Thermal Management
Currently, every one percent improvement in processor performance brings a three percent increase in power consumption. This is because as transistors shrink and more are packed into smaller spaces, and as clock frequencies increase, the leakage current likewise increases, driving up heat and power inefficiency. If transistor density continues to increase at present rates without improvements in power management, by 2015 microprocessors will consume tens of thousands of watts per square centimeter.

To meet future requirements, we must cut the power density ratio dramatically. A number of techniques hold promise. As explained earlier, Intel CMP designs with tens or even hundreds of small, low-power cores, coupled with power-management intelligence, will be able to significantly reduce wasted power by allowing the processor to use only those resources that are needed at any time.

In addition, Intel CMP designs will support ultra-high performance without ultra-high clock speeds, thus mitigating some of the current leakage problems that increase with frequency. Further, Intel CMP designs will also exploit variability in transistor speed made possible by future high-density process technologies. Slow and fast transistors can utilize different supply voltages, with time-critical tasks assigned to fast, high-power cores and other tasks to slower, low-power cores. The ultimate goal is to build fully power-aware architectures that can automatically reconfigure to meet power and workload requirements.

Additionally, a variety of techniques at the circuit level—including body bias, stack effect and sleep transistors—can be used to control current leakage. Increasing on-die memory (through large caches) not only increases performance but, by reducing off-chip memory accesses, also slashes power consumption and power density. Specialized hardware such as TCP/IP processing engines can also reduce power consumption because they can execute these functions more efficiently (with less circuitry and fewer clock cycles) than general-purpose hardware.

2) Parallelism
Taking advantage of Intel's future CMP architectures requires that tasks be highly parallelized—for example, decomposed into subtasks that can be processed concurrently on multiple cores. Today's single- and multi-core processors are able to handle at most a few simultaneous threads. Future Intel CMP processors will be able to handle many threads—hundreds or even thousands in some cases. Some workloads can be parallelized to this degree fairly easily by developers, with some help from compilers, such that the processor and microkernel can support the necessary threading.

In image processing, for instance, the image can be subdivided into many separate areas, which can be manipulated independently and concurrently. Around 10 to 20 percent of prospective workloads fit into this category. A second group of workloads—around 60 percent—can be parallelized with some effort. These include some database applications, data mining, synthesizing, text and voice processing. A third group consists of workloads that are very difficult to parallelize: linear tasks in which each stage depends on the previous stage.

Category 1 represents the early workloads on the parallelism roadmap. The real challenge lies in category 2, which represent the bulk of applications. Parallelizing these tasks requires human intervention and a number of software technologies, currently the focus of multiple industry efforts. These technologies include:

- Arrays of adaptive software libraries and algorithms to handle parallelism seamlessly, including sophisticated tools to generate these libraries.
- Extensions to existing programming languages, and in some cases a new class of programming languages, to help enable parallelism.
- A sophisticated threading model that is compatible with compilers, the operating system and user-level applications, and exploits multiple cores.
- A set of program expression, performance, and verification tools that advance the parallel engineering process.
- Innovative new data representations that support massive multithreading.

Intel and our co-travelers are applying an enormous amount of research in this area, and we are confident in our ability to develop the software technologies that will take advantage of our evolved architectures[5].

3) Complexity Management

Future Intel CMP architectures represent a big leap in operational complexity. Consider what's needed to coordinate the large number of concurrent processing activities, assign threads to cores, manage power, reconfigure the hardware as necessary on the fly and more—all dynamically at high speed. Efficient operation requires that the processor hardware take on more of these management tasks, rather than leaving them to the operating system and managed runtime software (such as the Java* virtual machine)[4]. In other words, the processor itself must have the built-in intelligence to manage and make the best use of the underlying hardware.

Such capabilities will be achieved through an intelligent microkernel. The microkernel will work with the OS to schedule threads and assign them to cores, turn off cores when idle to conserve power, monitor operations and shift jobs among cores if any are in danger of failing, manage special-purpose hardware, and reconfigure cores, caches and interconnects as needed. The microkernel, in turn, will present a unified interface to OS and application software. Operating systems and applications will be able to use the functionality of the microarchitecture, specifying parameters such as power-usage or performance requirements, and let the system decide how best to reach those goals using the processing capabilities at hand. Another benefit of this abstraction is that it will allow applications to be ported easily from one class of computing devices to another.

4) Security and Manageability

Intel's microprocessor evolution over the next 10 years will address the need for new, more-robust mechanisms to ensure security and manageability. While such functions are handled today by software and human beings, future systems will need more advanced security and manageability built in. New firmware and hardware mechanisms are being developed for Intel platforms. For example, virtualization can be used to create special partitions for security and management functions, and to implement layers of protection that prevent tampering, intrusions and virus attacks[4]. Hardware reconfigurability will allow such virtualization to be implemented by the chip manufacturer, the device manufacturer and/or during runtime.

5) Variability and Reliable Computing

This is likely to be a major focus of research as we approach the end of the decade. As future transistor sizes drop to 20nm and below, we are likely to see increasing variability in the behavior of these transistors. Intel is exploring several new mechanisms needed to compensate for this underlying variability in transistor behavior:

- Hardware-based self-monitoring and self-management (which can, for example, detect when a core or a circuit is likely to fail, and preemptively shift the work to another core or circuit). Over the coming years, at least 5 to 10 percent of a processor's 10 billion-plus transistors will be used for circuits and logic, dedicated to ensuring reliability.
- Firmware-based fault prediction based on error history.
- Statistical computing techniques, which use probabilistic models to derive reliable results from unreliable components.

6) High Speed Interconnects

Intel's CMP architectures will circumvent the bottlenecks and inefficiencies common in other architectures, but may encounter new performance challenges. Chief among these is the communication overhead of shuttling data among the numerous cores, their caches and other processing elements. High-speed interconnects are therefore needed to move data rapidly and keep the processing from bogging down. Intel's approaches include improved copper interconnects, and ultimately optical interconnects (which move data at light speed).

The challenge lies not only in the interconnect material, but also the interconnect architecture. Ring-type architectures are being applied successfully in designs of up to 8 to 16 cores. Beyond that, new interconnect architectures capable of supporting hundreds of cores will be needed. Such mechanisms will have to be reconfigurable to handle a variety of changing processing requirements and core configurations. Interconnect architecture is an area of active and extensive research at Intel, at universities and elsewhere in the technology industry.

Summary

This article and the accompanying paper has outlined an ambitious vision which, in some respects, is a wide departure from present-day processors and platforms. But in reality, this vision is based on a continued evolution of Intel's drive for increased parallelism, and our proven investment, research, development, manufacturing and unparalleled ecosystem enabling capability that, when taken together, will continue to lead us into an era of more powerful, versatile and efficient processing engines and platforms containing those engines. Ultimately, this evolution is driven by usage—what people want from technology and what they do with it. And though no one can precisely predict the future course of technology, the developments now underway point to some likely outcomes.

Based on current requirements and trends, we at Intel believe that processor and platform architecture needs to move toward a virtualized, reconfigurable CMP architecture with a large number of cores, a rich set of built-in processing capabilities, large on-chip memory subsystem and sophisticated microkernel. This architectural evolution, delivered with volume computing economics and an adherence to maintaining compatibility with thousands of already existing applications, will ensure that Intel processors and platforms will continue to power a breathtaking array of sophisticated new applications over the coming years, transforming business and daily life in ways we can only begin to imagine.

More Info

Read the more comprehensive Platform 2015 white paper, which can be found on the Intel Web site.

You can also learn more by visiting the following areas of the Intel Web site:

- Platform 2015
- Architectural Innovation
- Technology and Research at Intel

References

1. Ramanathan, R M, *Intel's T's Deliver New Platform Enhancements Beyond Gigahertz,* Intel Corporation, 2004.

2. Annavaram, Murali, Grochowski, Ed, Shen, John, Mitigating Amdahl's Law Through EPI Throttling, Intel Corporation, 2005.

3. Architecting the Era of Tera, Intel Corporation, 2004.

4. Ramanathan, R M, Bruening, Francis, Virtualization - Bringing Flexibility and New Capabilities to Computing Platforms, Intel Corporation, 2004.

5. Kuck, David, Platform 2015 Software: Enabling Innovation in Parallelism for the Next Decade, Intel Corporation, 2005.

Author Bios

A number of Intel's top technologists contributed to this article. Their names and titles are listed below. You can read their full biographies on the Intel Web site:

Shekhar Borkar, Intel Fellow, Corporate Technology Group; director, Circuit Research
Pradeep Dubey, senior principal engineer and research manager, Platform Architecture Lab
Kevin C. Kahn, Intel Senior Fellow, Corporate Technology Group; director, Communications Technology Lab
David Kuck, Intel Fellow, Software and Solutions Group; director, Parallel and Distributed Solutions Division
Hans Mulder, associate director, Intel Research; director, Intel Research network of university labs
Steve Pawlowski, Intel Senior Fellow, Digital Enterprise Group; chief technology officer and director, Platform Planning, Architecture and Technology
Justin Rattner, Intel Senior Fellow, Corporate Technology Group; manager, Corporate Technology Group

—End of Technology@Intel Magazine Article—