Recognition, Mining and Synthesis Moves Computers to the Era of Tera

Pradeep Dubey
Senior Principal Engineer, Manager of Innovative Platform Architecture
Microprocessor Technology Lab
Intel Corporation

Table of Contents

(Click on page number to jump to sections)

Recognition, Mining and Synthesis Moves Computers to the Era of Tera

Pradeep Dubey
Senior Principal Engineer, Manager of Innovative Platform Architecture
Microprocessor Technology Lab
Intel Corporation

"The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don't have a clue as to what any of that data actually means."
– Stephen Cass, "A Fountain of Knowledge," *IEEE Spectrum*, January 2004

Overview: New Ways to Sort Data

World data is doubling every three years and is now measured in exabytes—a billion billion bytes. This much data will require computing platforms that can deal with terabyte-level (1,024 gigabytes) workloads. Intel is investigating three fundamental classes of processing capabilities in order to make meaningful use of this enormous sea of information: Recognition, Mining and Synthesis (RMS).

These three classes provide a framework for investigating and developing systems and software that can address tomorrow's computing workloads. This article explains RMS classifications, gives many examples of how they could be applied, and discusses the implications of RMS for the platform architectures of the future.

The Coming Era of Tera

Today, the world's data outstrips our ability to comprehend it, much less take maximum advantage of it. According to the How Much Information project at the University of California at Berkeley, print, film, magnetic and optical storage media produced about 5 exabytes of new information in 2002. That's the equivalent of 37,000 new libraries with book collections the size of the Library of Congress (17 million books). Approximately 92 percent of the new information was on magnetic media, mostly on hard drives.

The information explosion is not slowing down either. It's speeding up. According to the Berkeley project, new information stored on paper, film, magnetic and optical media doubled just between 1999 and 2002. As more of the world's population comes onto the Internet and uses technologies such as instant messaging, digital photography, digital video and weblogs, the rate of data increase will continue to accelerate.

Looking ahead, Intel predicts the Era of Tera is coming quickly. This will be an age when people need teraflops (a trillion floating point operations per second) of computing power, terabits (a trillion bits) per second of communications bandwidth, and terabytes (1,024 gigabytes) of data storage to handle the information all around them. You can already buy terabyte hard drives.

To handle all this information, people will need systems that can help them understand and interpret data. Search engines will not be enough. Tapping the Web alone, today's search engines often turn up thousands of documents in a single search, but many with minimal relevance. What's more, 50 million new or changed Web pages are added every day. And it's not just text; it's videos, photos, and various other kinds of media. Digitalization has given us infinite ways to create, store and display information. Now we need computers to be able to "see" data the way we do, to look beyond the 0's and 1's and identify what is useful to us and assemble it for our review.

Obviously, another leap forward in computing capabilities is needed to harvest and take advantage of the digital wealth being created. The ability to have computers intelligently understand this data and help us use it in business, medicine, sociology, science, hobbies, and virtually every other realm of human endeavor could have enormous benefits and initiate countless revolutionary advances in human knowledge. It could tremendously increase our ability to understand and make better use of the world around us.

To spur this next leap forward in computing capabilities, Intel is investigating classes of processing capabilities necessary to deal with tera-level workloads. Intel classifies these processing capabilities into three fundamental types: Recognition, Mining and Synthesis—or simply RMS. Intel believes most processing capabilities for today's and tomorrow's computing workloads can be classified under these three categories. Through a thorough understanding of RMS, Intel expects to gain a head start in designing platform architectures for the future.

The First Step: Computers Need To Work with Models

To advance what we do with computers, computers need to move beyond what they do for us today. Currently they are superb at storing data, processing it, and moving it around. What they don't do well is help us understand what the data is or what it means. We know a JPEG is a picture, but we don't know exactly what it's a picture of unless we open it. We can scan a picture of a person into a computer, but the computer can't search to find similar pictures of that person based on that scan.

The problem is ordinary computers don't model things. Aside from supercomputers, today's computers aren't capable of developing mathematical models of complex objects, systems or processes. Nor are they powerful or fast enough to perform such tasks at speeds people demand. We can't plug in a statistical model for a rare malignant tumor or the behavioral pattern of a shoplifting employee and search for similar instances of the model in a data set. To benefit from the wealth of data building up in the world, we need to be able to communicate with computers in more abstract terms (high-level concepts or semantics). We need to speak in terms of models.

To communicate using models, computers must have the performance capabilities and software to be able to construct, manipulate, and evaluate mathematical models. We need computers that can model events, objects and concepts based on what we show the computers and the data accessible to them.

Working with Models: Recognition, Mining and Synthesis

To work with mathematical models, computers must be able to understand and recognize models. This is what Intel calls ***Recognition***.

Recognition is a type of machine learning. In this case, we mean computers must be able to examine data and images, and construct mathematical models based on what they "see." Depending on the data provided, that model could be of a valuable vase, a terrorist's behavior pattern, the right time to sell a particular type of stock, or the qualities needed by an actor to successfully play the part of Othello. Recognition is the "what is." It's identifying that a set of data constitutes a model and then constructing that model.

Recognition must be a continuous process because data is always being created, always coming in. Through constant model building, computers will get better at enriching a model with further data and eliminating the data that's not necessary.

Once a computer has recognized the "what is" and turned that data into a model, the computer must be able to search for instances of the model. This is what Intel means by ***Mining***. Mining refers to searching a data set, such as the Web, and asking "is it?" to find instances (such as good stock trading opportunities or the best actors to play Othello) of a model. The better computers are able to build models (Recognition), the better computers should be at finding instances that fit these models and our needs (Mining). For example, a robust model for say, a car, will be able to locate its instances even when the car images are nonlinearly transformed or hidden in part.

Synthesis is discovering "what if" cases of a model. If an instance of the model doesn't exist, a computer should be able to create a potential instance of that model in an imaginary world. In other words, Synthesis is the ability to create an instance of a model where one doesn't exist (see **Figure 1**). If two treatment options are being considered for a particular medical patient, Synthesis could project for a doctor the probable success of each option based on the patient's medical history and current health. If an advertising agency is considering switching celebrities in an advertising campaign, Synthesis will show how that new celebrity would appear in the existing campaign—and possibly predict the success of making the switch.

Figure 1. *This diagram shows how RMS can be used to create a model, find instances of that model, and predict what a model instance might be like where there isn't one.*

Our description of RMS suggests a linear relationship between the three components, but other forms of synergistic interactions also hold great promise. For example, Mining can be used to refine the Recognition model. Consider a system for classifying news articles. This system might use semantic clusters to classify new articles as they appear based on how former articles were classified. The more articles classified, the better the system becomes at matching new articles to the established classifications. Yet another example is spam filters. These filters improve their Recognition abilities through "observing" human actions on the email that comes in. Every time we "tell" the filter something is spam, the filter's ability to identify similar spam improves.

Like Mining, Synthesis can also be used to improve a Recognition model. Synthesis can be used to create additional models from a single model. These additional models can then be used to improve Mining results. An example would be face recognition. By applying two portrait photos of a subject to a physical model of a face, forensics science could be used to render additional views of the subject and generate hundreds of facial expressions based on the original photos. The resulting models could be used to find all instances of a particular person in a photo database.

The Power of RMS

"Our society is creating massive amounts of complex data as the world continues to go digital, but it doesn't have the capability to enjoy the full potential of this rich resource. There is a critical need for scalable, adaptable and programmable computing architectures that have the capability to recognize, mine and synthesize all of this digital data." **–Intel Senior Vice President Pat Gelsinger**

Enabling mass-market computers to perform RMS at speeds beyond those of today's high-performance or super-performance computers will dramatically impact our lives, society, and the world at large. It will elevate the level and language of human/computer interaction, allowing us to extract maximum value from the data around us. The scope of this article is only to highlight a few examples. Like many new technologies, RMS will spawn many new uses and spread through numerous fields once it becomes available.

Medicine

Figure 2 shows how, through RMS, a tumor could be:

1. Recognized as a model
2. Identified through mining patient data as the type of tumor in a particular patient
3. Synthesized in a way that would predict the effects of the tumor's progression for a particular patient and whether treatment is advisable

Images courtesy of Surgical Planning Laboratory, Department of Radiology, Brigham and Women's Hospital

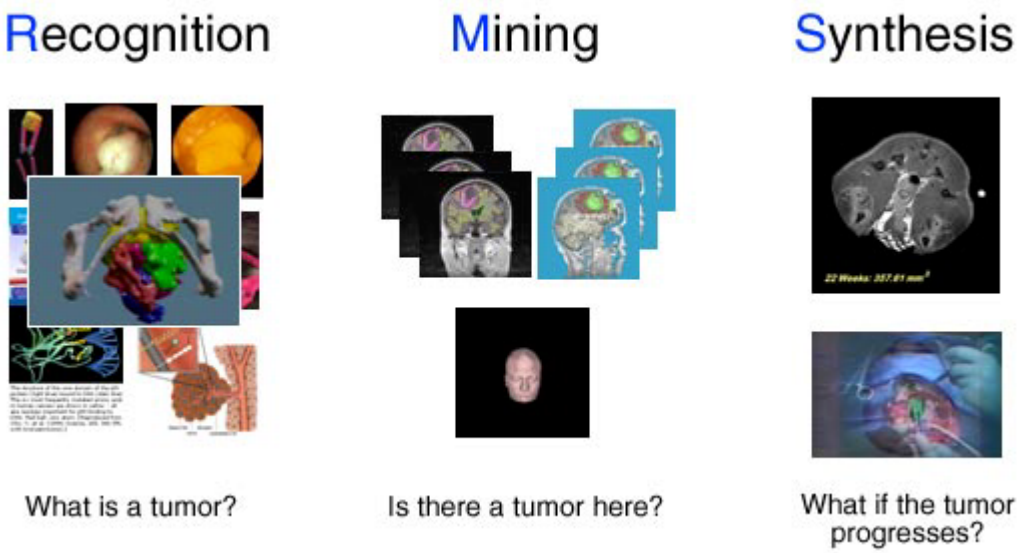

Figure 2. *An example of how RMS might be used in medicine.*

Synthesis could also be used as a means to determine the efficacy of various treatment options. One could take what is known about each treatment model and perform Synthesis with the patient's medical history and condition. For instance, on a cancer tumor, Synthesis could be used to simulate the effects and prognosis of treatments by chemotherapy, radiation, surgery, combined approaches, or no intervention at all. By examining the outcomes of each treatment, both the doctor and the patient could decide the best course of action.

In the near future, individual genetic profiles will enable "personalized medicine." Using RMS, a patient's genetic profile could be modeled and kept on file. When a patient comes in with a specific condition, such as high blood pressure, a doctor could mine drug databases using the patient's profile for the best drug options. The doctor could then test through Synthesis for possible reactions by that particular patient to the various drugs suggested. With technology like this, doctors could more safely prescribe medicines.

Investment

Selecting smart equity investments involves more than analysis of a company's financials. An investor should consider industry trends and a wide spectrum of potential factors that could include currency rates, trends in buying habits, oil prices, recent research on Asian attitudes toward American foods, and much more. There is so much possible data to consider that in the end people often just rely on standard indicators, such as a stock's price-to-earnings ratio, recommendations by friends or a broker, and their gut feeling.

RMS could radically change that. Using a model of a successful investment, people could mine a data set for other potential successful investments. Synthesis could then enable a person to examine "what if's" having to do with various investment periods and the potential effects of various events, such as a rise in interest rates or one company acquiring another. While RMS wouldn't take all the uncertainty out of an investment, it could allow investors to make use of a much larger data set in determining which equities to buy. RMS could create new expectations for what better portfolio management software should do.

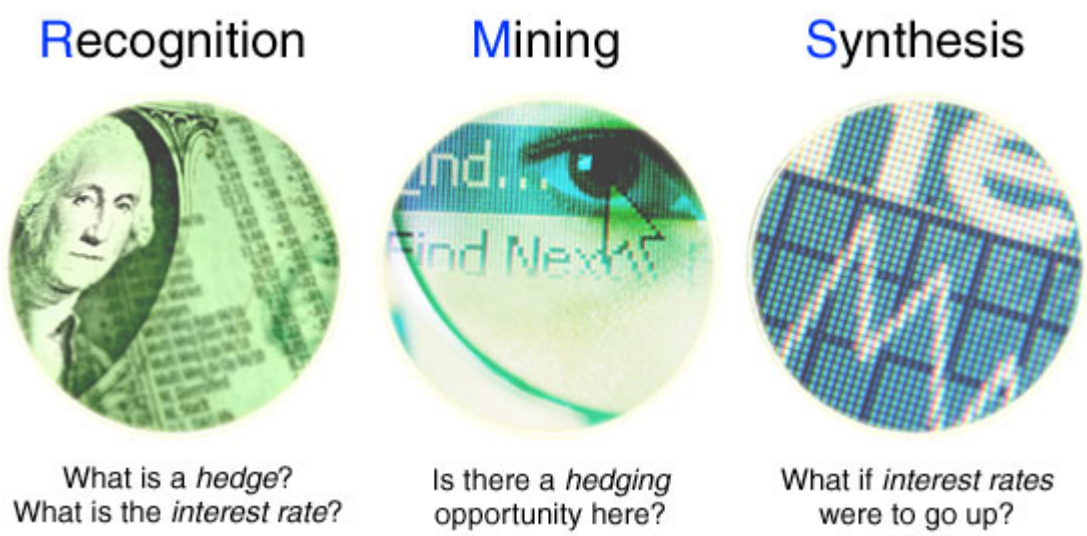

Figure 3. *RMS could potentially be used to "mine" for instances where, according to a particular hedge fund's strategy, it would make sense to invest. Synthesis could be used to predict what might happen to a particular hedge fund investment if certain conditions, such as prevailing interest rates, changed.*

Business

Successful businesses with large cash reserves often begin looking for acquisitions to broaden their product lines and their markets. Some work out; others don't. Some take years to pay off; others end up in divestiture. Even though financial analysts and other specialists pore over the prospective company's books for months before a transaction is completed, the result can still go sour over "intangibles" and "unexpected turns of events." Through RMS, many of these intangibles could become tangible and unexpected turns of events could be anticipated. Using an acquisition model and mining the vast amount of data on the two companies and past acquisitions, outcomes could be more accurately predicted using Synthesis. Other acquisition targets could also be more accurately "tried" before announcing any interest or a potential purchase.

Small businesses could profit from RMS as well. Independent retailers, such as a local auto parts store, have limited shelf space and need to optimize their inventory for both better profitability and customer service. Using a model of what a good inventory item is (good margin, fast mover), a local parts store could mine national car service databases by manufacturer to determine what parts to carry.

Surveillance could be another key business use of RMS. Today's surveillance cameras simply record what they see. Imagine if their images were run through a computer that could recognize and provide the right kind of alert for each instance of trouble. For example, a bank's surveillance system, trained through models to recognize a gun, could immediately alert police when someone has entered the bank with one. On a larger scale, RMS could help monitor freight. Today, only four percent of the container traffic coming into the U.S. on ships is inspected. Imagine if through RFID tags, surveillance cameras, biological sensors and other monitoring systems, computers could monitor imports using a database of models to identify which might require an alert and further inspection.

Another application for business could be hiring. Through Recognition, a model of a successful employee for a particular job position could be created. Using Mining, all particular potential candidates who fit this model could then be identified. Finally, through Synthesis, an idea of how each candidate might perform in a number of job-specific situations could be examined. Hours of sourcing and interviewing could be reduced to interviewing only a few candidates who, according to the Synthesis results, "performed" best in the simulated job situations.

Gaming

Computer games are constantly seeking to add more realism in their imagery to draw the player deeper into the conceit of the game. One application of Synthesis is having the game recognize at any particular moment where a player is, where the ambient light sources are, and simulating in real-time all the points or rays of light—including reflections—that are hitting the surfaces of objects in the scene.

This is called global illumination and currently is done only on expensive Hollywood films, such as the *Shrek** movies. Global illumination currently takes hours of rendering for each frame of a scene. Through platform innovations, the gaming computers of tomorrow may be able to render photorealistic scenes with the perfection of global illumination, providing the ultimate player experience.

Another exciting RMS application for gaming could be putting the actual player in the game. In this case, a model of the player could be created. The Synthesis or "what if" is the placing of this model as a character in the imaginary world of the game. Imagine a future generation of a game like *The Sims** in which models of a player and models of their friends actually inhabit a simulated world.

Speaking of multiparty gaming, tomorrow's computers will be powerful enough to run multiparty gaming and collaboration on their own. This should enable greater proliferation of gaming, plus inspire new forms of games and collaboration. A model of one game you like might be used to find similar instances.

Home

As the popularity of digital audio, photography and video grows, so does the number of MP3 files, photos and clips on people's hard drives. Having busy lives, most people do a poor job of naming and categorizing these files. The result is a great deal of time and frustration searching folders for particular audio files, photos or video clips.

In the future, it's only going to get worse, especially as hard drives continue to grow in capacity and people continue to save digital media onto them in a haphazard manner. Imagine if, using RMS, a person could easily assemble for a 25th wedding anniversary a collection of photos and video clips of just the married couple. Through Recognition and data Mining, these files could be easily and quickly found amongst hundreds of thousands of files. Using Synthesis, someone could even surprise the couple by showing what they might look like on their 50th anniversary or how they'd look now in their original wedding photos.

Consider, too, how people might use RMS as they gain access through the Web to entire libraries of music, film and television shows. It could unleash new channels of creativity and investigation as people use Synthesis to see what one film star might have been like in another film star's part, or how using a different pitcher might have changed the outcome of a baseball game. Or imagine shopping on the Web and being able to use Synthesis to interactively combine models of you and models of various articles of clothing. You'll be able to very quickly "try on" lots of outfits and colors before you ever click "buy"—and accurately determine how they fit.

Another excellent use of RMS would be information monitoring and sorting to help people stay current with a particular subject or interest. For example, a person's computer could constantly model specific information to look for based on what a person views and uses on their computer. This would be somewhat similar to how Amazon and other vendors on the Web today make product recommendations based on buying habits, but RMS would enable it to be applied to a much bigger universe of information.

This universe could even include the multitudes upon multitudes of blogs on the Internet today. Imagine having your computer monitor and collect on your hard drive all the items (text, video, photographs) in this information universe that should be of interest to you. Using models, a computer could mine the entire Web continuously for information of value to a person's business, educational pursuits and hobbies.

RMS Design Implications for Future Platforms

Processor platforms today are not designed or optimized for RMS applications. Mainstream desktop and server platforms today are optimized for traditional office applications, audio/video encode/decode, or database transactions. From a platform perspective, RMS workloads are very different from such workloads. For example, compared to these traditional workloads, RMS applications require much higher compute density, higher external bandwidth, offer significantly more coarse-grain (thread-level) parallelism; inputs are more often streaming in nature, and datasets are more often unstructured. However, various RMS applications also differ from each other. Compute-to-bandwidth ratio, for example, can differ significantly across recognition, mining and synthesis phases.

For RMS to work on a single platform then will require designing platform capabilities that efficiently handle system requirements of this class of applications. This also means developing a platform flexible enough that it can be used for statistical computing, collaborative filtering, physical simulation, behavioral modeling, Internet searching, real-time rendering of photorealistic images, and much more.

A New Architectural Paradigm for the Tera Era

There is a common core of computing across the RMS applications: *multimodal recognition and synthesis over large and complex datasets*. Building an RMS-optimized platform is about exploiting the inherent platform convergence implied by this workload convergence. With tera-levels of performance, it should be possible to bring these workloads together on a single architectural platform optimized for computing kernels at the core of these applications. This would be a major accomplishment, eliminating the need to optimize architectures for one workload type at the expense of the other. This requires a fundamental rethinking of how to deliver new levels of performance.

The superscalar platform architectures of today's desktop computers are not designed to provide this level of performance. Monolithic architectures running single threads of execution will not satisfy future workload demands when the amount of work being done per clock cycle will need to increase 10 times, or even 100 times. Instead, we must rely on multiple levels of concurrency and execution units to provide this performance.

Instead of a chip with a single execution unit, Intel is developing a "many-core" chip architecture (with potentially hundreds of compute threads). Imagine instead of one processor, there are 16, 32, 64, and so on, up to perhaps 256 processor cores on a single die. The advantage of multiple processors over one big processor is that more data can be put through them by simultaneously issuing instructions to a host of functional units. A single big processor could also be built with the same number of functional units, but it would be much more expensive (it might be much larger than the small processors put together), consume much more power, and could not be clocked nearly as fast. Numerous small cores are more efficient than a single big core.

In addition to having multiple cores with multiple threads, Intel's many-core architecture will have three key attributes:

- ***Scalability.*** This is the ability of the platform to exploit multiple levels of concurrency, based on the resources available, and to scale performance of the platform to meet the increasing demands of the RMS workloads. Historically the industry has scaled up by increasing the capabilities and speed of single processing cores. For RMS, scalability would also mean adding multiple cores and threads of execution, and scaling correspondingly the rest of the platform to deliver increased application-level performance.
- ***Adaptability.*** An adaptable platform adjusts to the intrinsic workload type and execution phases. Future platforms must be designed to adapt to any type of RMS workload. For example, they must be able to run a statistical algorithm and then quickly adapt to perform real-time rendering of a photorealistic image. Thus, they must be able to dynamically configure and reconfigure for the workload at hand.
- ***Programmability.*** The challenge in bringing high-performance computing to the desktop has been in defining parallelizable applications and the need for the software environment to understand the underlying architecture. The solution is a programmable system that has coarse-grain workload characteristics (like concurrency, data structures, synchronization, and communications requirements) communicated to the hardware. Simultaneously, architectural characteristics—like allocation of the cores and threads according to the resource requirements of the workloads—will need to be communicated back up to the applications.

Some of the Challenges Ahead

For RMS, the future holds key challenges in the areas of the applications and algorithms themselves, the platform architecture, and the programming model.

For RMS applications and algorithms, three big challenges are:

1. How to create a Recognition model.
2. How to use a Recognition model to Mine a data set.
3. How to apply a model to perform Synthesis.

Obviously, there won't be one solution for each of these three challenges. Complex model building, including simulations, will help Recognition. Mining will benefit immensely from creative solutions to high-dimensional indexing problems. Synthesis will require its own solutions, such as ray tracing and physics-based animation. While a lot of

impressive work has been done in these areas, researchers are still coming up with better algorithms and mathematical models everyday.

Other RMS challenges exist on the platform front. A many-core architectural platform robust enough to run RMS applications must surmount challenges in scalability, adaptability and programmability as described earlier. Progress is being made. At the Fall 2004 Intel Developer Forum, Intel demonstrated working silicon for the next generation Intel® Itanium® processor for supercomputing that featured a multicore design. In 2005, Intel plans to roll out its first dual-core processors for servers and mobile computers. Looking toward the future, Intel should be able to double the number of cores on a chip every 18 to 24 months.

Besides developing a robust-enough, many-core architectural platform, another major hurdle in RMS is writing the necessary software. Right now, writing code for highly parallel architectures is a specialty, requiring skills held by only a small subset of software developers. Developers must configure their algorithms to solve problems in parallel, and software tools must be developed to enable programmers to extract the inherent parallelism out of their programs. Intel is actively working with leading academic institutions, as well as aggressively funding efforts within the company's research and product groups, to address the challenges of parallel, many-core architectures.

Summary

RMS promises to have far-reaching affects on our ability to capitalize on the wealth of digital information that continues to grow all around us at a furious pace. It could spur profound improvements in business, medicine, science, entertainment, and many other fields. A key prerequisite for this revolution in Recognition, Mining and Synthesis is tera-level computing.

To achieve this level of computing, Intel is working on the development of many-core processors that could conceivably one day run hundreds of threads through hundreds of processors within a single die. The company is also helping the industry develop the modeling algorithms necessary for RMS, as well as the programming techniques for enabling a wide range of applications to tap the power of parallel processing. RMS represents a huge challenge for Intel and the industry. But as we enter the Tera Era, this new paradigm for using computers to enrich our understanding and use of data promises rich rewards.

More Info

To learn more, visit the following areas of the Intel Web site:

- Architecting the Era of Tera: Digital Transformation Fuels New Opportunities
- Technology and Research at Intel

References

Cass, Stephen (2004) "A Fountain of Knowledge," in IEEE Spectrum, January, pp. 60-67.

Author Bio

Pradeep Dubey is a senior principal engineer and manager of Innovative Platform Architecture (IPA) in the Microprocessor Technology Lab, part of the Corporate Technology Group. His research focus is computer architectures to efficiently handle new application paradigms for the future computing environment. Dubey previously worked at IBM's T.J. Watson Research Center, and Broadcom Corporation. He was one of the principal architects of the AltiVec* multimedia extension to Power PC* architecture. He also worked on the design, architecture, and performance issues of various microprocessors, including Intel® i386™, i486™, and Pentium® processors. He has published extensively, and served on various conference committees in the areas of computer architecture, multithreading, and multimedia processing. He holds 24 patents. Dubey received a B.S. in electronics and communication engineering from Birla Institute of Technology, India, an M.S.E.E. from the University of Massachusetts at Amherst, and a Ph.D. in electrical engineering from Purdue University. He is a Fellow of IEEE.

—End of Technology@Intel Magazine Article—