IBM

Global Technology Outlook

A comprehensive analysis of long-term trends in business technology by IBM Research.

2008

The View From Here

Few organizations are afforded the expansive view of the ever-shifting technology landscape as IBM Research. Our eight global labs employ 3,000 of the world's top scientists. Our own technology often drives the changes that relentlessly remake this industry year after year. And through our relationships with clients and partners in more than 170 countries around the world, we are privy to exciting new uses of technology.

Our unique global position spanning the worlds of technology, business and society allows us to develop and share what is perhaps the world's most comprehensive analysis of ongoing, long-term technology trends. We call it the Global Technology Outlook.

Our capabilities as a globally-integrated research organization allow us to develop a breadth and depth of insight that is unmatched. From the mature economies of the world to the emerging regions such as India, China and Eastern Europe, the Global Technology Outlook weaves together a truly global perspective on business and technology.

In producing this year's report, IBM Research collaborated more closely than ever with our colleagues in the IBM Academy of Technology, our community of Distinguished Engineers, our clients and partners, and our services and industry teams. The report highlights major shifts in technology architectures, starting with the physical limitations we are now seeing at the processor level, to the devices people use to access technology and the new kinds of datacenters required to deliver those applications and technologies. There is not a more complete view of how technology can affect business than the Global Technology Outlook.

Innovation that matters—for our company and for the world—is one of IBM's core values. The Global Technology Outlook, and the insights you will read about on the following pages, reflect that value. This report is designed for your organization to benefit from them as much as we have here at IBM.

About the Global Technology Outlook

NO ORDINARY EXERCISE IN TECHNOLOGY PREDICTION

IBM is, of course, in the technology business. We invent it, sell it, integrate it and maintain it. And we've been doing this for nearly 100 years. So it's not surprising that we care deeply about how technology is changing and where future opportunities lie.

As such, it would be easy to assume that the Global Technology Outlook (GTO) is nothing more than an elaborate tool that IBM uses to inform its corporate and product strategies. But the GTO goes far beyond the typical product development exercise.

The GTO takes an unflinching look at trends that are well outside of IBM's own offerings and expertise, some of which may even threaten entire IBM product lines. It uses history as a guide and takes a long-term view, looking out five or ten years further than most industry experts. It endeavors to understand the cultural and business contexts in which new technology will be used. The GTO solicits ample outside counsel from around the world when making its predictions. And unlike any other corporate strategy exercise, the GTO shares the results with clients, academics and even competitors.

Indeed, the GTO is unique among the many sources of technology prediction today. So for the purposes of this report, rather than list the many things the GTO is, it is perhaps more instructive to discuss what the GTO is not:

The GTO is not new. Though it has gone by many different names over the years, the Global Technology Outlook has been a regular part of IBM's research since 1982. Each year the process has been refined and adjusted to meet the changing nature of technology and the marketplace. Today, we believe it is the most comprehensive and effective technology prediction tool anywhere in the world.

The GTO is not perfect. Predicting the future of technology is not easy. It's so hard, in fact, that not even IBM can always get it right. We sometimes predict trends that are too nascent and not yet fully understood. As Nobel Prize-winning physicist Neils Bohr once said, "Prediction is very difficult, especially if it's about the future."

The GTO is not speculative. The trends that the GTO identifies are being driven by business needs. They are not IBM's vision of how the world should be. These trends are big, with major global business implications. And though we may not always know what to call them or how they will play out, we do know they are real.

The GTO is not ignored. Technology predictions come and go, and many IT leaders take note of them and then quickly return to their daily routines. But the GTO commands the attention of a broad range of IT influencers, both inside and outside of IBM itself. In fact, IBM placed a series of so-called Big Bets—$100 million investments—based on the findings of this year's GTO.

In the end, the GTO is a thorough, well-sourced, and useful guide to where technology is trending. It is both reflective of existing trends, and predictive of future trends. And while it informs IBM's own research direction, it also provides valuable insight to clients, academic institutions, business partners, and other research organizations.

Inside the Global Technology Outlook

THE NEVER-ENDING PROCESS THAT IS THE GTO

The business of predicting the future of technology is like trying to hit a moving target. At times it seems the entire industry can change direction overnight as disruptive technologies appear suddenly and alter the entire landscape with shocking speed. That is why the process of defining and refining the insights that make up the Global Technology Outlook has no beginning and no end. It is an endless, iterative cycle that accommodates change without resistance.

By the time the insights from one year are being finalized early in the calendar year, fresh ideas are already being solicited from IBM's 3,000 top researchers. Shortly thereafter, the ideas come pouring in from all corners of the globe: Switzerland, Israel, India, China, Japan and the United States.

The ideas come in all shapes and sizes. Some are broad, big-picture trends. Others are much more specific, built around a particular technology or process.

"We literally give thousands of people a chance to raise their hands and say, 'Hey, I've got an idea.' It requires a tremendous effort to filter and consolidate all the ideas that come in, but at least we can say we didn't leave anything unsaid."

Nick Bowen, Vice President of Technology, IBM

From there the ideas are vetted by six GTO strategists who bring broad industry perspective to the table. They have the ability to dismiss, embrace and connect the dots between seemingly disparate ideas. Only about 200 ideas make it through the first cut, at which time they are judged on 20 different criteria, such as the impact they will have on customers, whether they will spawn new businesses, and how long they will take to develop.

At this point in the process, input is solicited from nearly every conceivable perspective within the company. All product divisions, sales and distribution heads, and corporate strategists are asked to weigh in. IBM Fellows, Distinguished Engineers, even retirees are consulted. And though history in no way guides the process, prior GTO results are given consideration.

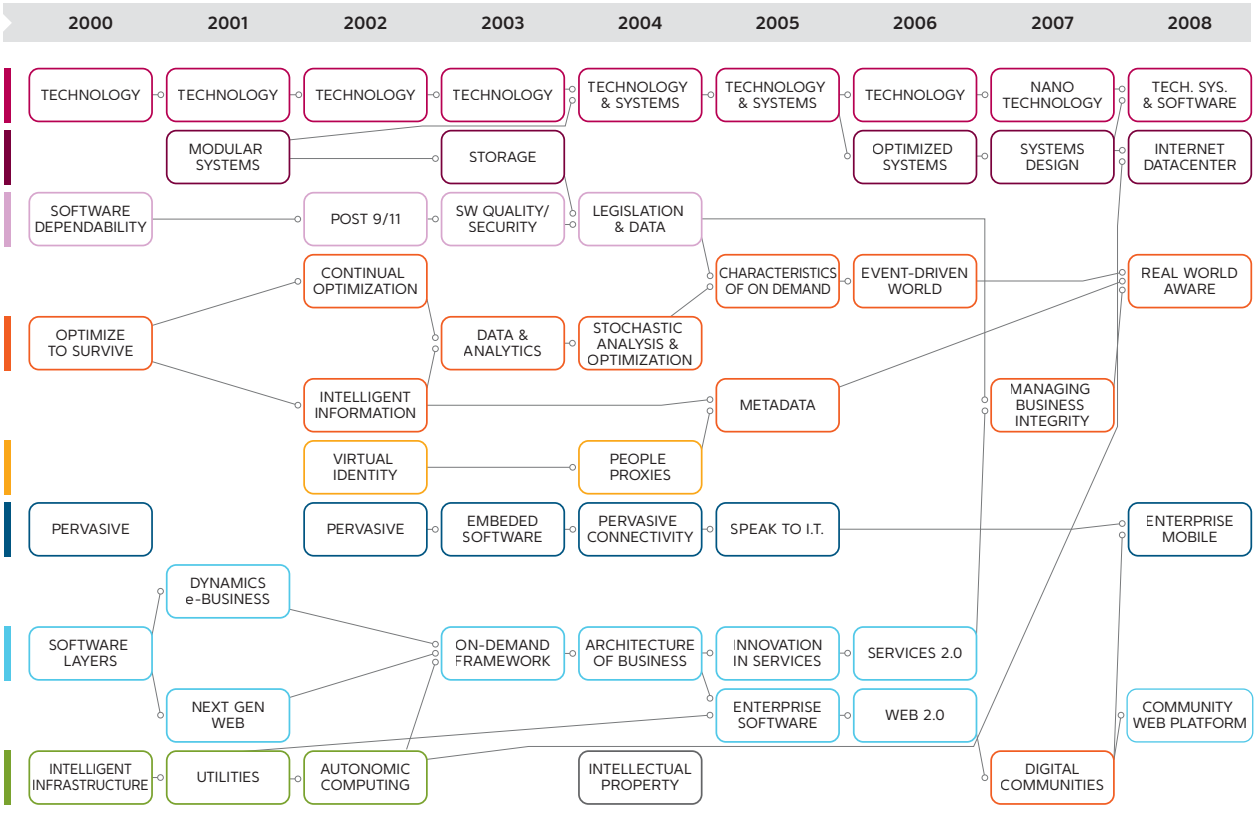
Historically, IBM Research was the source of this massive collective input. But when it comes to predicting trends as broad and far-reaching as those in the GTO, there is no such thing as too much perspective. So last year, for the first time, the GTO process was opened up to IBM's broader ecosystem of clients, business partners and fellow researchers from academic institutions around the world.

"These are some of the most difficult trends to predict, and we have the smartest clients in the world, so it just makes sense to get their views on where technology is going," says Bowen. "These changes don't happen in a technological vacuum. They happen because there is a business need for new technology. And that demand is coming from our clients and business partners."

From there, the GTO team spends months considering the many perspectives, vetting the ideas again, assigning probability metrics to each, consolidating like ideas, and refining the final insights. By the end of the year, usually less than half a dozen predictions ultimately make it through the process. And finally, the insights are presented to IBM's senior executives, including Chairman Samuel Palmisano, in a day-long meeting in December that amounts to a thesis defense.

The predictions are tweaked right up until the last minute, and that process of constant refinement is the defining characteristic of the GTO. Technically, the GTO process takes one year. Realistically, it never ends.

Figure 1 //
History of the GTO
Global Technology Outlook topics over the past eight years.



Real-World Aware

MAKING SENSE OF A COMPLEX WORLD ... IN REAL TIME

Since the beginning of time, smart people have understood the value of timely information. The ancient Greeks employed their fastest runners to deliver messages between city-states. History's greatest armies invested heavily in strategic communications systems. And over the years, technologists have worked tirelessly to improve those delivery systems, dramatically reducing the time it takes to get information into the hands of people who can act on it.

But in the Internet Age, we have taken this need for speed to a whole new level—instant messaging, overnight delivery, speed dating. Call it Generation Now: the society that wants all of its questions answered yesterday. If the communications revolution of the last 20 years has taught us anything, it's that there is no such thing as "fast enough."

That sentiment is driving the move toward IT systems that we are calling "Real-World Aware." In today's globally competitive business environment, information needs to be gathered, analyzed and acted upon in real time. And it needs to be collected from myriad sources, some traditional, some nontraditional; some structured, some not. Currently, this kind of data collection and analysis is done in batches, in daily, weekly, sometimes monthly time increments. And it quite often requires an ever so time-consuming human review before it's put to good use. But with real-world aware systems, the time frame is severely compressed—and the human element may be eliminated entirely.

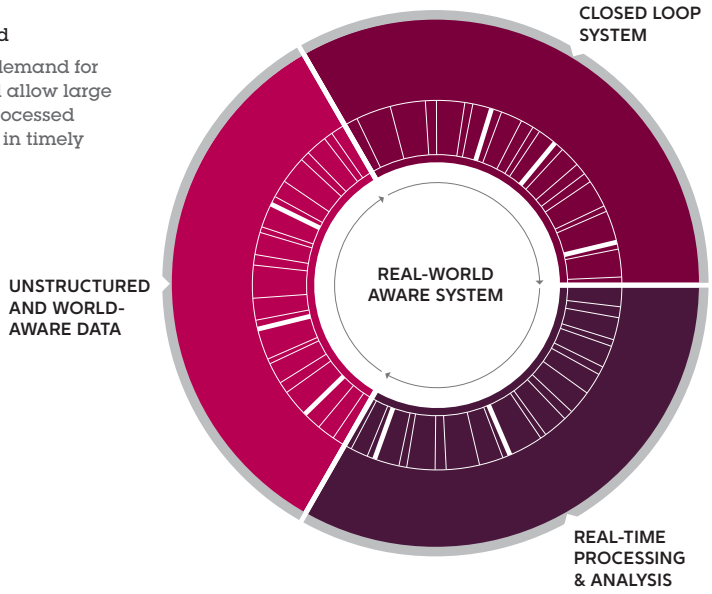
When new data is generated by an event, that data is captured and analyzed, compared with other data (both current and historical), and acted upon, all within a few milliseconds. *(Figure 2)*

The sources of the data being fed into these systems are limitless. Dozens of new data points are being created every day, through the increasing use of sensors (which can record data as varied as temperature, moisture, movement, even heart rates), audio and video surveillance, mobile and location-based devices, and more. In fact, we are generating more data than we can possibly store, analyze or even make sense of.

To turn this data deluge into something more useful, several evolving technologies will be needed. One of the critical systems technologies that will fuel this trend is called "stream processing," or the ability to process information as it is being generated, or with very low latency. Not surprisingly, this is not easy to do. For one thing, it takes a tremendous amount of bandwidth and

Figure 2 //
Analytics and Control Closer to the Real World

Real-world and IT business events will drive demand for real-world aware systems. These systems will allow large volumes of current unstructured data to be processed and analyzed with almost no delay, resulting in timely reactions in the closed-loop operation.



- REAL-TIME RESPONSE TO REAL-TIME ANALYTICS: HAVE MARKET PERMISSION HEALTH CARE, UTILITY
- NEW TECHNOLOGIES: FINANCIAL, CALL CENTER
- NEW OPPORTUNITIES: GOVERNMENT, RETAIL

computing power to sift through vast amounts of data in real time. Another problem is that many of the new sources of that data are so-called "unstructured" data. In other words, it's the kind of intuitive information that makes sense to humans, but not necessarily to machines—things like news reports, video and business trends.

That's why real-world aware IT systems require a rethinking of the traditional computer architecture, a change that could be as fundamental as the transition to the Web browser. Because real-world aware systems will gather so much data from so many sources, there needs to be more intelligence built in at the edges of the network. Handling this amount of processing in a central location would require computers the size of a small country.

These systems need to be architected hierarchically, so that the data that needs to be acted upon most urgently is analyzed and processed at the edges of the network, and the data that requires more human analysis is stored and mined centrally by very powerful supercomputers like IBM's Blue Gene.

"The combination of the Stream Computing architecture and the Blue Gene supercomputer allows enhancements to our real-time messaging and analytical capabilities while simplifying the underlying infrastructure. This taps into IT innovation that can benefit our business."

Rizwan Khalfan, CIO of TD Wholesale Banking

Part of the intelligence that must be built into the edges of the network is the ability to allow only the most relevant bits of data to pass through to the next level of the network. In other words, not all data is created equal, and IT systems that can make quick decisions on which data is worthless and can be discarded, and which data has value, will ultimately achieve greater efficiency. These systems also need to be able to recognize patterns in unstructured data. For example, if a video surveillance program is trained

to recognize license plate numbers, it needs to be able to quickly determine if the same license plate number has run three consecutive red lights in a row. This kind of "metadata" then becomes easier to work with and extremely valuable.

The final step of the real-world aware loop is tying back into the legacy systems on the network. All IT systems are a combination of old and new technology, so marrying real-time data with existing systems is critical for the effectiveness of any real-world aware system. For example, having a camera in a retail environment that tracks the store shelf for product placement information is nice. But allowing that system to tap into the SAP inventory application and automatically place orders for hot new products would complete the automated cycle.

Financial institutions are already experimenting with real-world aware technologies, but there is also early potential for this technology in health care, law enforcement, retail and risk analysis environments. As the systems that facilitate this real-world awareness are improved, and the cost is reduced, there is no end to the potential applications. And that's good news for Generation Now. ■

Enterprise Mobile

THE MOVE TO THE MOBILE WEB WILL CREATE NEW BUSINESS MODELS AND REDEFINE ENTIRE INDUSTRIES

To the average person, it may appear as though the wireless revolution has already taken place. After all, mobile phones are nearly ubiquitous, with more than 3.3 billion subscribers worldwide. They outnumber fixed-line phones. And there are three times as many mobile phones in use than PCs.

You might think the wireless revolution is complete. But you'd be wrong. Because this is just the beginning of what could possibly be a more disruptive technological change than that of the PC (or perhaps the Internet itself). Our current wireless environment is analogous to the dial-up era of the Internet: wildly exciting but relatively unsophisticated and unbearably slow. We are, quite simply, wireless infants. *(Figure 3)*

In fact, perhaps the most important step in the evolution of mobile society has yet to even occur: the convergence of wireless and the Web. And this convergence will radically alter the way mobile devices are used by consumers and businesses alike.

"We are at an epic point in telecommunications history, when the mobile platforms and the Internet platforms that have enabled such spectacular growth and innovation are poised, if we manage this well, to merge."

Paul Bloom, Telecommunication Research Executive, IBM

Today, accessing the Web on a mobile device is nothing less than excruciating. For those of us accustomed to the high-speed Internet connections afforded by hard wire, or even WiFi, the sclerotic wireless transmission rates and baffling user interfaces of current mobile phones are simply unacceptable. But there are two critical trends that are conspiring to change all that. First, true wireless broadband is becoming a reality. And second, the power and capabilities of mobile devices are taking a quantum leap forward while prices decline.

Many people think of the shift to wireless broadband as an incremental improvement. They think we'll be able to do the same things we can do today, only faster. But that's a misconception. The move to wireless broadband is not incremental, it's fundamental. Today we access the Web over cellular networks at data rates that are measured in the single-digit megabits-per-second range. But through true IP-based wireless technologies like WiMAX or Long Term Evolution, wireless bandwidth is expected to jump to 100 megabits per second in just a few years, and one gigabit per second in less than a decade. At speeds like this, it's difficult to even imagine the myriad ways mobile devices will be used.

"In the early days of the Internet, people could not even conceive of the kinds of applications we now run over the network. The mobile Web will be just like that. Voice, text and e-mail will seem laughable compared to what will be possible. The way that enterprises communicate with their customers, employees and partners will radically change, creating fundamental changes to their business processes."

Paul Bloom, Telecommunication Research Executive, IBM

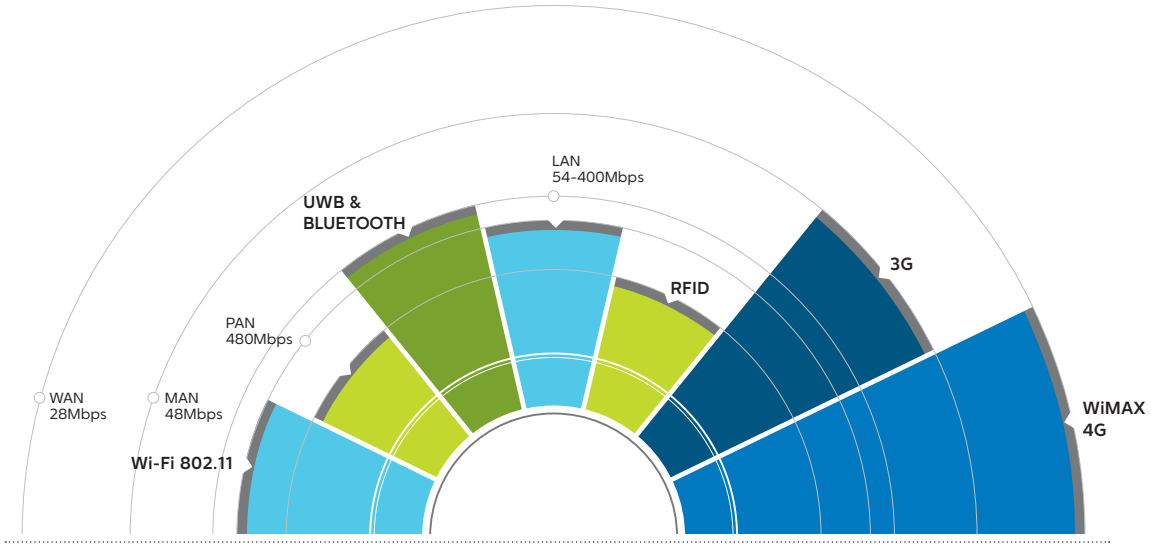
Complementing the wireless broadband evolution, mobile phones and smart phones are becoming legitimate alternatives to the PC, especially in the developing world. These devices are increasingly capable, and in the near future could realistically be thought of as mobile business terminals. And they are being designed to take full advantage of the Web convergence being made possible by more bandwidth.

So what does all this new technology mean? It means the telecommunications landscape will never look the same again. Just as the Web forced major upheaval in the fixed-line telecommunications industry, the mobile Web will force wrenching change upon the current wireless establishment. Today, your wireless carrier (in conjunction with your mobile phone maker) decides which applications your phone can access and how much it costs. They are the de facto wireless gatekeepers.

But with the advent of the mobile Web, the many Web applications you currently access over your PC (Yahoo! Mail, Google, etc.) will also be available on your mobile device, regardless of which provider or device you use. And that will drive a new era of openness in mobile communications, enabled by standards, open applications (software and services that can be used on any device) and open devices (that can be used on any wireless network). In other words, the mobile Web will begin to look a lot more like the regular Web. And that will result in massive investment in mobile Web infrastructure and an explosion of ideas for new applications.

Figure 3 //
Spectrum of Wireless Technology

Unlike the seamless system we will have in the future, today's wireless landscape features multiple wireless technologies, each of which is optimized for a different use.



Already we've seen some ingenious new applications for mobile devices. Mobile banking, location-based marketing, social networking and telemedicine are just a few. The time is coming for enterprise application providers to deliver products and services that are optimized for the mobile device interface. Speculating about potential mobile applications has been a parlor game among the digerati for years, but suffice it to say that wireless applications are slated to grow 17 percent a year between now and 2011. That's compared to wired applications, which will grow a mere 1.5 percent during that same time.

58% of the world's mobile phone subscribers live in developing nations

The effect the mobile Web will have on developing countries will be more profound. In many emerging markets, where fixed-line telecommunications infrastructure is poor or nonexistent, the growth of wireless subscriptions has been exponential. Increasingly it seems that mobile phones, and not cheap laptops, will be the device that brings the rural people of Africa, India, China and elsewhere more fully into the Internet age. In many of these areas, voice is still the killer application, as most people lack the literacy skills needed to operate PCs. But many regions are developing highly sophisticated mobile applications out of need, like the M-PESA mobile money transfer system operated by Kenyan wireless provider Safaricom. Through this mobile revolution, existing businesses will be redefined, and countless new businesses will be created.

Of course, all of this technological change will require some major advancement in the IT systems that support these mobile services. New mobile application programming models will emerge, as will the underlying IT architectures. Mobile security (both at the device level and the network level) and device management will play increasingly important roles, particularly in the enterprise. These are important issues that can, and will, be addressed so as to allow the rampant innovation the mobile Web promises. ■

Community- and Information-Centric Web Platforms

A LITTLE OPENNESS GOES A LONG WAY

American poet and essayist Ralph Waldo Emerson put it best when he said "our best thoughts come from others." Emerson was referring to the basic human need to collaborate and share ideas among a community with common interests. That this sentiment came from a man who spent most of his time extolling the virtues of individualism and self-reliance makes it all the more powerful.

Fortunately, the Internet Age brings with it tools that make forming communities and sharing information easier than ever. The wild success of social networking sites like Facebook and MySpace attest to the draw of online communities. But it's not just consumers that can take advantage of these mechanisms. These "community- and information-centric Web platforms," have tremendous potential for enterprises of all stripes as well. And we've only just begun to understand the potential.

The evolution toward community- and information-centric Web platforms has taken place over several decades. Traditional computer systems and the applications they ran were bought, customized and maintained all within the four walls of the company that used them. With the advent of the Internet, companies begin to outsource their applications to companies that "hosted" them, accessing them only when needed through a dedicated or virtual network. This reduced the cost, but limited the control and customization to which enterprises had become accustomed. This was known as "Software as a Service."

Today, a new trend is emerging. Over the last few years ecosystems have begun to develop around certain popular applications. For example, Salesforce.com offers a Web-based customer relationship management (CRM) application, which is available on a subscription basis, making it far less expensive than traditional CRM applications. But to get around the lack of flexibility this model offers the customer, Salesforce.com has opened up some of its code to the outside software developer community. That community has responded by developing hundreds of new applications that complement and enhance their core service.

In so doing, Salesforce.com has created a Web platform, a vibrant community that includes the developers and users of their service. The result is exponential growth of their service offerings at a relatively low cost to their customers.

The concept is similar to the "user-generated content" trends of recent years, just with an enterprise spin. By opening up selected parts of their valuable data, enterprises can create virtuous cycles of content and

value creation with their customers. Google has done this by fostering a community around its Google Maps applications and the so-called "mashups" it facilitates. Amazon built an ecosystem around its vast treasure trove of merchandising data. This trend is completely redefining the way that software and applications are created, and it will only continue to grow and expand in the coming years.

The success of these communities depends on a certain level of openness. Like the trend of open source software itself, these community-based Web platforms require that companies share at least some of their most valuable assets. In some cases that may mean granting access to the tools and code that underpin a particular application or service.

Here's an example of how it works today in the banking industry. The Operational Riskdata eXchange Association, also known as ORX, is a consortium of more than 45 banks that share some of their most valuable data. Because they are heavily regulated, banks are required to do extensive risk analysis, which in turn requires a vast amount of loss data. However, the loss data coming from one bank may not be sufficient to yield a statistically sound risk analysis. So ORX acts as a broker between the banks, allowing them to anonymously and securely share loss data, yielding a more accurate risk analysis. *(Figure 4)*

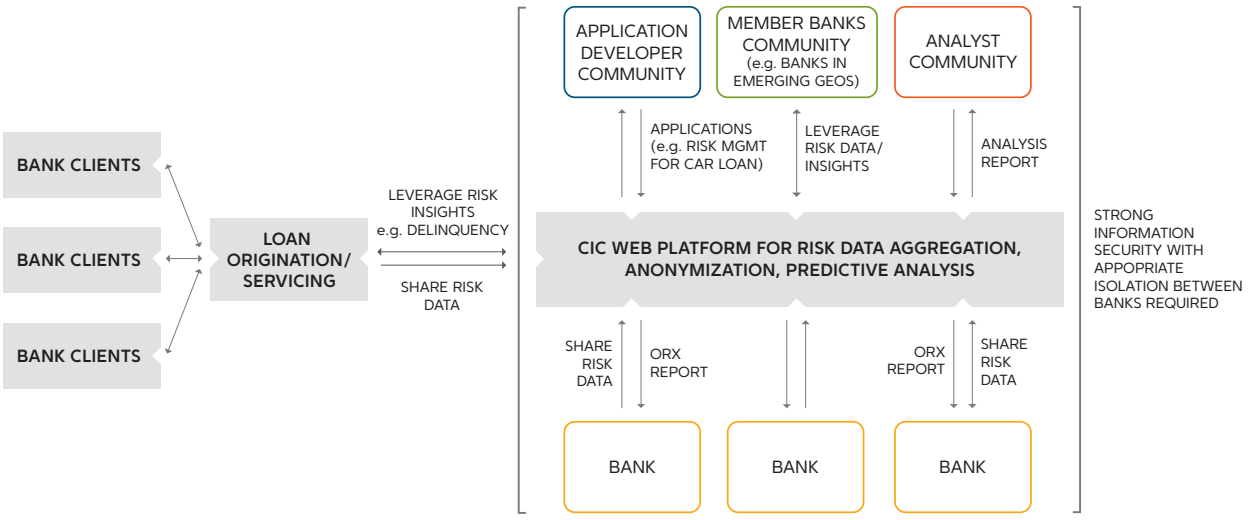
This superset of data has an added bonus as well: it allows the developer community within the banks to thrive, creating new models and tools to analyze risk.

"There are fundamental issues in operational risk measurement that we have begun to tackle together, that banks had not been able to tackle on their own. What you do by default is create a community, one that learns together, and in some ways is as valuable as the data itself."

Simon Wills, Executive Director,
Operational Riskdata eXchange Association

Figure 4 //
Sharing Operational Risk Data within the Banking Industry Community

Membership banks pooling risk data and models through Operational Risk eXchange (ORX).



It's easy to see how this kind of community-based enterprise data could emerge in other industries as well. Projects are already underway in fields as diverse as health care and petroleum exploration. But building these communities is not technically insignificant. It requires a seamless integration of data from both within and outside the enterprise. These systems will also need to protect the data that enterprises do not wish to share.

Many of these issues can be addressed through the development of service-oriented architectures that can limit access or visibility into certain data types. But there are a few other technologies and techniques that will be required. For example, businesses will need to perform extensive value/risk analysis to determine which services and data would be beneficial were it made open. Once a community is formed, collaborative intelligence technology would analyze the trends and contributions of that community, managing the collective knowledge base and helping to improve the service. Then there is the rise of "in vivo" development, which refers to the emerging process of creating software through an iterative process which has no end and no beginning, just continual improvements to constantly evolving applications. And finally there is the rather complex technology that can support the "massive multitenancy" that these Web-based communities require. Aside from raw horsepower, massive multitenancy's unique ability to deliver a single instance of software to multiple clients requires an end-to-end rethinking of system architecture.

These community- and information-centric Web platforms are going to have a profound effect on the way enterprises view their data, their industry, and the way they go about acquiring and using applications. There is an unlimited amount of potential in these solutions for enterprises with the foresight to exploit them. ■

Internet-Scale Datacenter

BUILDING SYSTEMS WITH PURPOSE AND INTENT

A computer, like anything else, works best when it is built and used for a specific purpose. Though far more complex than a hammer or saw, a computer is a tool just the same. And all tools must be designed to a task.

These days it is easy to forget this simple fact. As computer systems grow increasingly complex, particularly in the datacenter, we find that many are being used in unintended ways—for example, running software they were never meant to run, or being housed in buildings that were designed for other purposes. As hardware and software additions to the datacenter require more and more connections between new and old technology, eventually the industry will hit a complexity wall in which datacenters become unmanageable. The result will be widespread power and utilization inefficiencies at a time when energy and efficiency are at a premium.

"At AT&T, we're extremely sensitive to power and cooling issues in our datacenters. It's not just a financial issue, but an environmental issue as well. That's why we're in the process of rethinking datacenter design from the ground up and working to achieve maximum efficiency throughout our centers."

Saeid Shariati, Vice President of Global Internet Datacenters, AT&T

That is why we believe that the datacenters of the future will be vastly different creatures than the power-guzzling, administrative headaches of today. In coming years, the hardware systems that occupy these massive compute farms will be designed in concert with the software they are intended to run. Indeed, even the buildings that house the datacenter will be custom-built for the type of work load and processes the systems will handle. These new datacenters will use less power, produce better results and require less administration.

When it comes to datacenters, complexity is the enemy of efficiency. And at the moment, thousands of enterprise datacenters find themselves in the midst of mind-boggling complexity. One technology that holds the promise of greatly reducing that complexity is virtualization, the process that pools disparate computing resources—processors, memory, storage—to appear as one. But so far virtualization has been mostly about consolidating servers. This is helpful, and it improves utilization, but it does not cut down on software complexity or significantly reduce administrative costs.

There are three technologies that are emerging that will radically alter the virtualization landscape. One is the concept of a Virtual Machine Image, or VM Image.

A VM Image is the bundling of the operating system, middleware and application into a self-contained, fully operational package. These images have instructions attached to them (metadata) that enable them to simply drop into a datacenter environment, find the necessary resources and execute.

The second technology is VM Scheduling. This is akin to system provisioning, in which an administrator can decide when and where to run a particular VM Image. It allows for rapid scheduling and prioritization of shared resources among other VM Images on the same system, for more dynamic and efficient environments. The third technology is VM Mobility, which is the ability to move virtual images around the datacenter while they are actually running, without skipping a beat. Though there is still work to be done on the standards and licensing fronts, these technologies have the potential to greatly improve the dynamism and efficiency of the datacenter. *(Figure 5)*

This new world of virtualization will require some significant changes to the datacenter architecture itself. The major new concept that will emerge is something we are calling an "ensemble." These ensembles are essentially collections of homogeneous hardware, or clusters, that have systems management capability built in; everything from workload optimization to restart and recovery. The key to these ensembles is their autonomic abilities. In other words, they will monitor their own utilization, heat production and power consumption, dynamically allocating resources as needed. By using the principles of autonomic computing—Monitor, Analyze, Plan, Execute (MAPE)—these ensembles require very little by way of administration.

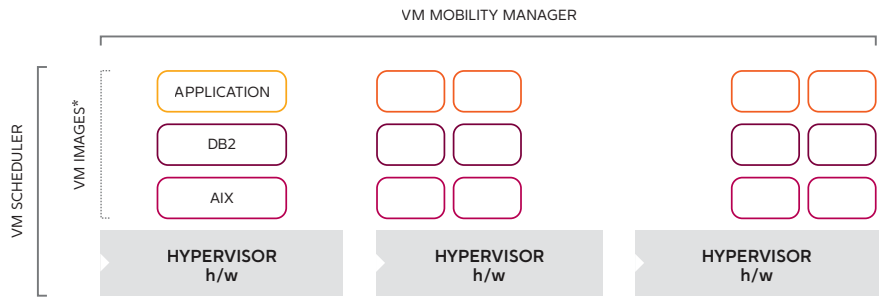
The goal of this re-architecting is to simplify the datacenter. Though the dynamic scheduling of workloads is actually a fantastically complex process, the interface that is exposed to the administrator is quite simple. By using the ensemble structure, managed by a service-oriented virtual machine interface, the datacenter becomes a system of self-contained components that interact with each other on an as-needed basis. *(Figure 6)*

There is one more element that needs to be rethought before datacenters can reach their full potential. Though it may seem sometimes that datacenters exist only in the world of ones and zeros, they are actual physical structures that require tremendous amounts of power and cooling in order to operate. In this way, datacenters are not unlike factories. And the lease, maintenance and power consumption are all factored into the cost of finished goods.

Like a factory, there is an optimal efficiency that can be reached in a datacenter by matching the machines to the building (or vice versa). By applying some of the same economic principles that measure the efficiency of factories to the economics of datacenters, we have arrived at some surprising recommendations for

Figure 5 //
Emerging Virtualization

Virtualization technology allows disparate computing resources to appear as one.



VM IMAGES HAVE POTENTIAL TO GREATLY REDUCE LABOR COSTS

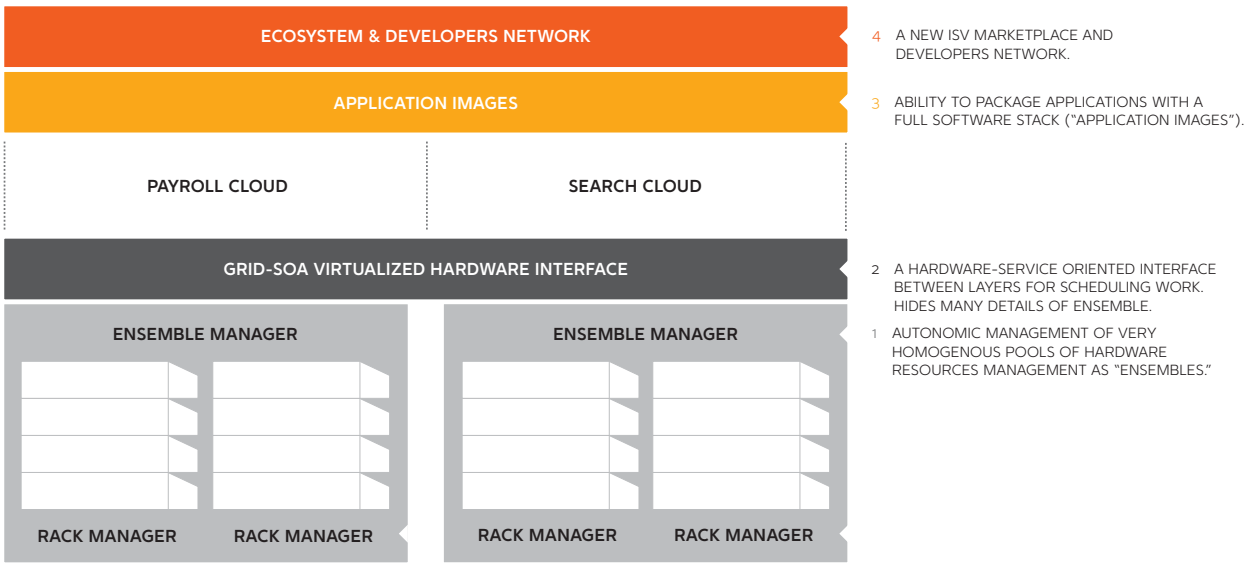
VM SCHEDULING ALLOWS BINDING OF THE VM IMAGE TO MACHINE AT RUN TIME

VM MOBILITY ALLOWS THE CREATION OF FREE SPACE WHEN NEEDED

** Purchased applications awaiting to run*

Figure 6 //
A New Datacenter Architecture Is Emerging

The datacenter architecture of the future will be simpler, faster and far more efficient, from the hardware to the software.



optimizing datacenters. For example, bigger is not always better. The cost benefits of scaling datacenters begin to diminish if it is too large and requires too much electricity. Maximum efficiency points will develop based on the workload of the datacenter and surrounding environment.

By optimizing all of the components that reside in this building block, and monitoring power and heat with sensors that feed back into the systems management capabilities discussed earlier, the datacenter can attain maximum efficiency. At a time when energy use carries heavy costs, both financial and environmental, every ounce of efficiency is highly valuable.

In short, the datacenter of the future will be a much more integrated, purpose-built machine. The type of workload will dictate the design of everything from the software to the building itself. And there may be a variety of different types of datacenters, based on their respective purpose. Not as simple as a hammer, but just as efficient and effective. ■

5 Insight

Technology, Systems and Software

LOOKING BEYOND THE MICROPROCESSOR FOR SPEED AND FUNCTION

Over the last 37 years, since the introduction of the first commercial microprocessor, society has come to expect a certain pace of technological change. Every year we expect computers to get smaller and use less power. We expect them to be faster, do more and cost less.

For decades we have taken these rapid evolutionary steps for granted. We have held fast to the belief that Moore's Law, which states that the number of transistors on a chip of a given size will double about every two years, will never be broken. And through the years, as transistor counts have increased, so too has the performance of microprocessors and the applications that run on them.

"It's a historical fact that whenever compute power is increased, and the costs of it are decreased, the world generates entirely new ideas about how to use it. This business is about smaller, faster and cheaper."

Thomas Theis, Director of Physical Sciences, IBM

But the science of computer design has always been a race to exploit the laws of physics. While it is still technically possible to double the number of transistors on a chip every two years, it is no longer possible to do so with proportionate increases in performance. The power consumed and the heat produced by those transistors are beginning to take a measurable toll.

The consequence is an urgent need to rethink how increases in performance can be derived not only from the core processors, but from the systems and software built upon them.

This isn't the first time that the accepted technology for fabricating microprocessors has run its course. The Bipolar Era, characterized by semiconductors with bipolar junction transistors (BJT), delivered improved system performance throughout the 1970s and 1980s, before it finally gave way to transistors based upon complimentary metal oxide semiconductors (CMOS). During the 1990s transition from Bipolar to CMOS, the performance of single processors degraded for a short while, but CMOS allowed for more transistors, less power and less heat than its predecessor. When combined with both hardware and software innovations, systems performance continued to increase throughout the CMOS Era.

Today, a new semiconductor technology is not ready and waiting. That is not to say that there are no promising technologies to boost system performance for decades to come. Carbon nanotubes, molecular chip design and quantum computing all could revolutionize the chip design industry. However, none of these technologies will be available for practical application

for at least ten years. This leaves us in a "Transition Era" between CMOS and the next major semiconductor technology, which necessitates some urgent innovation in the way chips are packaged, the architecture of the systems built with them and the software that runs on them. *(Figure 7)*

We have already seen the early stages of this Transition Era with the introduction of microprocessors with multiple cores, which achieve higher aggregate performance than single-core processors while decreasing the power and heat requirements. Another important technology that can work in conjunction with multicore systems is 3-DI, or 3-Dimensional Integration. Traditional chips are two dimensional. But by stacking processors, cache, accelerators, and other functional components atop one another within a single package, and by connecting them vertically, designers will be able to integrate performance in line with the historic trends. *(Figure 7)*

Another advantage of the multicore processors described above is the ability to tailor each core to specific functions. These "heterogeneous nodes" would each be optimized for designated tasks—for example, decryption, XML parsing or pattern matching. *(Figure 8)* Faster, heterogeneous nodes add complexity to the software development, because applications must be written specifically to operate in this environment.

The speed with which these chips are fed data also needs to improve to alleviate one of the most vexing bottlenecks in computer design. This involves the I/O systems (input/output), memory and caching. One way to boost these speeds is to shift from wires to fiber optics as the means of shuttling that data across the chip, processors, memory and accelerators. Traditionally used on the edges of networks, there will be a progression of optics getting ever closer to the processor itself, ultimately resulting in a chip with integrated photonics, or optical fiber communications.

In addition, new memory technologies will emerge. A fast, dense, low-power, nonvolatile technology known as phase change memory (PCM) shows particular promise. The nonvolatile nature of PCM also leads to new opportunities for optimizing the structure of the systems.

Figure 7 //
CMOS Era Simple Scaling Benefits

Throughout time, changes in technology properties have defined distinct eras. We are currently transitioning from the CMOS Era into a future Nanotechnology Era.

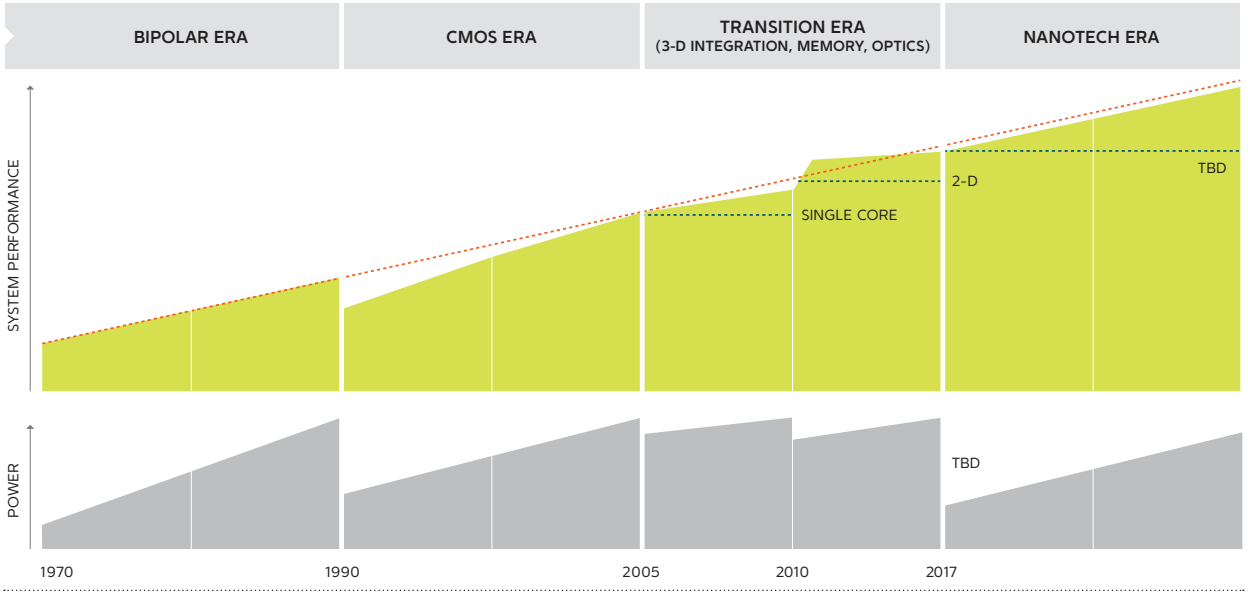


Figure 8 //
Transition Era in Comparison to CMOS and Nanotech Era

New technologies have historically driven up performance and power consumption, relying on transitions to keep power usage down. The Transition Era is challenged to preserve this trend until the arrival of the Nanotechnology Era.

	CMOS ERA	TRANSITION ERA	NANOTECH / QUANTUM ERA
LITHOGRAPHY	STANDARD LITHOGRAPHY	COMPUTATION LITHOGRAPHY	IMPRINT / SCANNING PROBE EUV LITHOGRAPHY, SELF ASSEMBLY
DEVICES, MATERIALS	SOI, ETSOI, FINFET ... HIGH-K, AIR GAP	NANOWIRES, COMPOSITES	NANOPHOTONICS, CARBON NANOTUBES, MOLECULAR COMPUTING, SPRINTONICS
I/O, PACKAGING	SINGLE & MULTICHIP MODE	3-D INTEGRATION, OPTICS	TBD
SUBSYSTEMS	SINGLE CORE, MULTICORE	MEMORY, HETEROGENEITY, MULTICORE	TBD
SYSTEMS	LOW-, MID- & HIGH-END	HIGH-END, INTERNET DATACENTER	TBD
SYSTEMS SOFTWARE	MONOLITHIC, HYPERVISOR, VIRTUALIZATION	DISAGGREGATED, FLEXIBLE, HYPERVISOR, VIRTUALIZATION	TBD
PROGRAMING MODEL	HOMOGENOUS, SEQUENTIAL, IMPERATIVE, GENERIC	HETEROGENEOUS, CONCURRENT, DELARATIVE, DOMAIN SPECIFIC	TBD
APPLICATIONS	PORTABLE, SCALE-OUT	VERTICALLY INTEGRATED, HYBRID SCALING, HETEROGENEITY	TBD

Operating systems and software applications built around these multicore, heterogeneous systems will need to be rethought to take full advantage. Independent applications will need to be consolidated. The principles of parallel programming (in which multiple tasks are carried out simultaneously) will need to be applied to both new and existing applications. Architectures will also need to be designed to support real-time analysis in data-intensive environments.

All of this means that the fastest and most effective systems will be those that are designed with processor, subsystems and software all designed in concert. There will likely be three different market segments that emerge over the coming years:

Cost-optimized systems These systems will be optimized for power and likely based on mid-size symmetric multiprocessors designed for virtualization. System software will evolve to take advantage of this platform.

High-end systems These systems will continue emphasizing single-thread performance at higher cost and power, but will require integrated hardware and software design to achieve system performance growth through increased parallelism.

Specialized domain systems These systems will have specific needs and require new thinking about the hardware and software they employ.

Ultimately, by optimizing not just the semiconductor technologies themselves, but also the architectures, subsystems, and software that surround them, performance at the system level will continue to increase at the historic rate during this transitional phase. "Computation will continue to get a lot faster and cheaper for a long time to come," says Tilak Agerwala, Vice President of Systems at IBM Research. "And with real-world aware systems, stream computing, and the mobile Web all imminent, we're going to need every last bit of it." ■

INTERNATIONAL BUSINESS MACHINES CORPORATION
NEW ORCHARD ROAD, ARMONK, NY 10504