



Intel[®] Technology Journal

Compute-Intensive, Highly Parallel Applications and Uses

Understanding the Platform Requirements of Emerging Enterprise Solutions

Understanding the Platform Requirements of Emerging Enterprise Solutions

Krishnamurthy Srinivasan, Digital Enterprise Group, Intel Corporation
Raj Ramanujan, Digital Enterprise Group, Intel Corporation
Michael Amirfathi, Information Services and Technology Group, Intel Corporation
Enrique Castro-Leon, Software and Solutions Group, Intel Corporation

Index words: emerging solutions, platform pathfinding, platform architecture, usage models, deployment models, performance characteristics

ABSTRACT

Given the long lead time to put an entire platform solution together, enterprise platform architects must be able to predict the intersection of evolving usage models, deployment models, and platform technology trends in order to meet platform requirements several years into the future. Platform architects and technologists are generally very familiar and comfortable with predicting platform technology trends but the same is not true for deployment and usage models. Hence, we find that most platform architecture development is primarily incremental and evolutionary until one comes up against a wall of some sort along one or more of the vectors. Being able to articulate the intersection upfront has the advantage of influencing all three vectors, thus resulting in an optimum solution. Due to this potential for inter-dependency, the process for developing the intersection is inevitably iterative and complex. In this paper we concentrate primarily on the two vectors that are least understood by platform architects: usage and deployment models. We present a list of key solutions being adopted in different vertical industries based on an extensive interaction with industry leaders. We discuss the business usage model trends and the technology deployment model trends across the industries. We describe how the emerging models are different in their characteristics from those prevalent today, and using several real-world examples, explain the platform implications. Two key trends in the data centers of large enterprises are “scale-out” and grid computing. Scale-out allows application solutions to be deployed over a multiple independent set of resources that are networked together, while grid computing allows flexible and dynamic provisioning of these resources to scaled-out applications. Both these trends are driven by usage and deployment vectors focusing on lowering initial costs as well as improving utilization, scalability, and availability

of data center resources. As enterprise compute and communication needs become increasingly complex, platform solutions from Intel have a crucial role to play in determining the optimum solution for these emerging models.

INTRODUCTION

As a general rule, it takes about four years to design, produce, validate, and take to market computer platforms with the microprocessor development taking the longest lead time. Hence, it is critical for platform architects to gain an understanding of the requirements of the solutions that need to be deployed that far into the future. In this paper we focus on the emerging enterprise IT solutions that are expected to have significant market adoption in 2009-2011 and discuss their expected performance characteristics.

There are two forces driving the rapid evolution of IT in the enterprise. The first is the challenge of continuous and rapid introduction of new or improved business processes to gain and retain an advantage in an extremely competitive market. This in turn translates into a need for accelerated deployment of new information systems capabilities (e.g., real-time decision support). The second driver, motivated by the prevailing trend of flat or dropping IT budgets, is to either cut or at least contain IT costs. Enterprise IT departments, faced with the challenge to provide new business capabilities at lower costs, are adopting various software and hardware technologies to develop, deploy, and maintain a larger portfolio of solutions at a reduced cost. Among these technologies are eXtensible Markup Language (XML), automated data center management, and server virtualization that uses an abstraction layer that decouples a consistent logical view of the server to the application from the actual physical resources that are utilized.

In this paper, we present the key new business solutions emerging in different vertical industries such as manufacturing and retail (e.g., intelligent inventory management in retail and collaborative product development in manufacturing) based on an extensive survey by Intel of key players in those industries. We identify key *usage model* categories that are common to these solutions (e.g., real-time supply chain management, collaboration, and image processing). We also briefly discuss some of the emerging *deployment models* such as XML and grids in the context of the usage models for which they are relevant. We discuss the significant new characteristics of these usage model categories. For example, real-time supply chain management often requires running of both transactional and decision support operations in the same environment, and synchronizing the databases underlying disparate business solutions more frequently. The key difference from the prevalent solutions of today is that all these operations will be running concurrently in order to enable a business to react very quickly to a rapidly changing environment.

These new characteristics of the emerging usage models are changing how enterprise IT departments do capacity planning. We present several specific examples based on real-world solutions adopted by leading enterprises in different industries.

KEY EMERGING ENTERPRISE SOLUTIONS

Physicist Niles Bohr said, “Prediction is very difficult, especially about the future!” This is particularly true about IT solutions where high expectations of new technology are often replaced by disillusionment and practical compromise. Significant time and costs are involved in optimizing large solutions and obtaining repeatable performance characteristics. The penalties of optimizing a general-purpose platform for the wrong requirements are even larger. Hence, we need to balance the prediction horizon with the accuracy to ensure that we look far enough ahead to accommodate the platform design lead times while being accurate enough with our predictions. To achieve such a balance, we focus on the solutions that have made it successfully past the phase where only the “innovators” with an extraordinary tolerance for costs and failure are interested in them; and are being adopted by the mainstream early adopters who pragmatically balance the technical risks with business benefits (Figure 1) [1]. For these solutions, mainstream adoption is expected in 2009-2011.

Intel’s Solutions Marketing Group works very closely with the leaders in various vertical industries such as manufacturing and retail to understand the key business capabilities they are trying to enable. These leaders are the

early adopters in their industries (Figure 1) whose lead would be followed by others in the next five to ten years. Intel plays the role of a “trusted advisor” in helping these businesses develop an IT roadmap to deploy these capabilities, and it also acts as a catalyst to drive the IT vendor ecosystem to remove any obstacles in the roadmap.

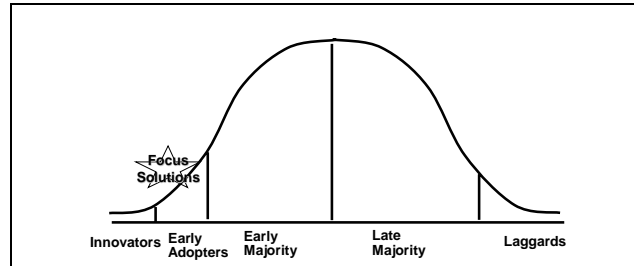


Figure 1: Phases of solution market adoption

Figure 2 shows the list of key solutions emerging in different industries Intel has identified. The list is based on the value of the functionality provided by these solutions to the businesses and their customers in the industry and the IT investment expected in deploying these solutions. While most solutions are industry specific, the requirements for mobility solutions are shared across industries.

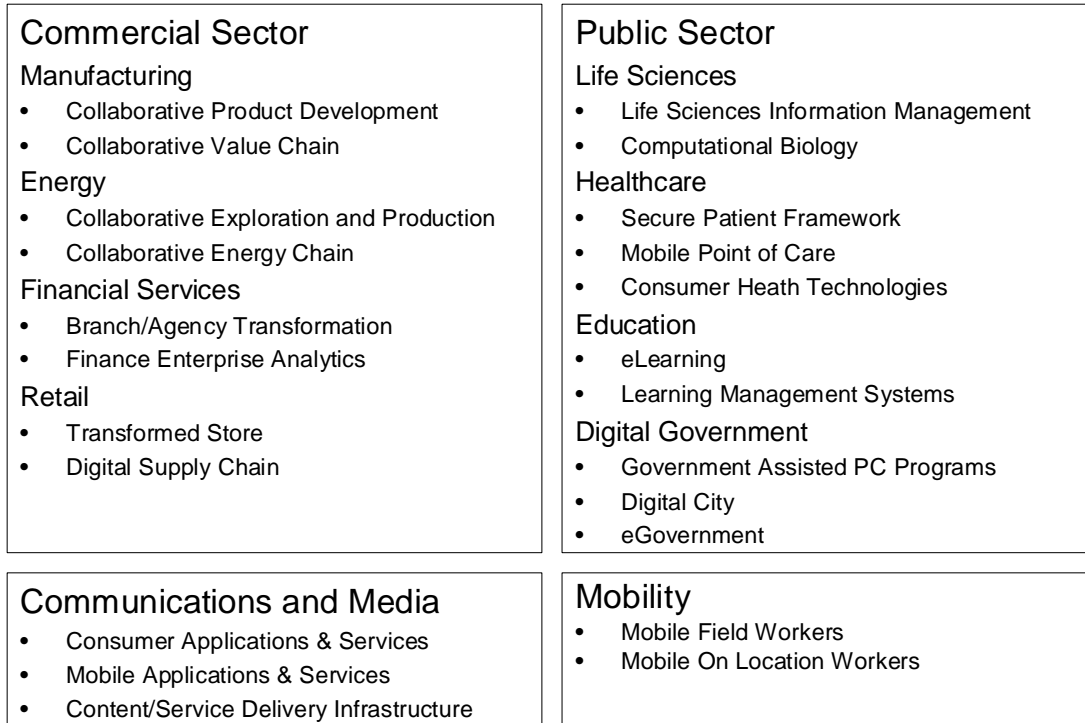


Figure 2: Key industry-specific and cross-industry solutions

USAGE MODEL CATEGORIES AND CHARACTERISTICS

Each solution shown in Figure 2 comprises two or more usage models. For example, consider the Digital Supply Chain solution in the retail industry (Figure 3). It consists of two usage models: the operational side of real-time

inventory management and supply chain integration; and the analysis and decision-support side of making pricing and inventory decisions, and detecting trends and anomalies. Figure 3 shows two other solutions from the financial industry, their constituent usage models, and their categories.

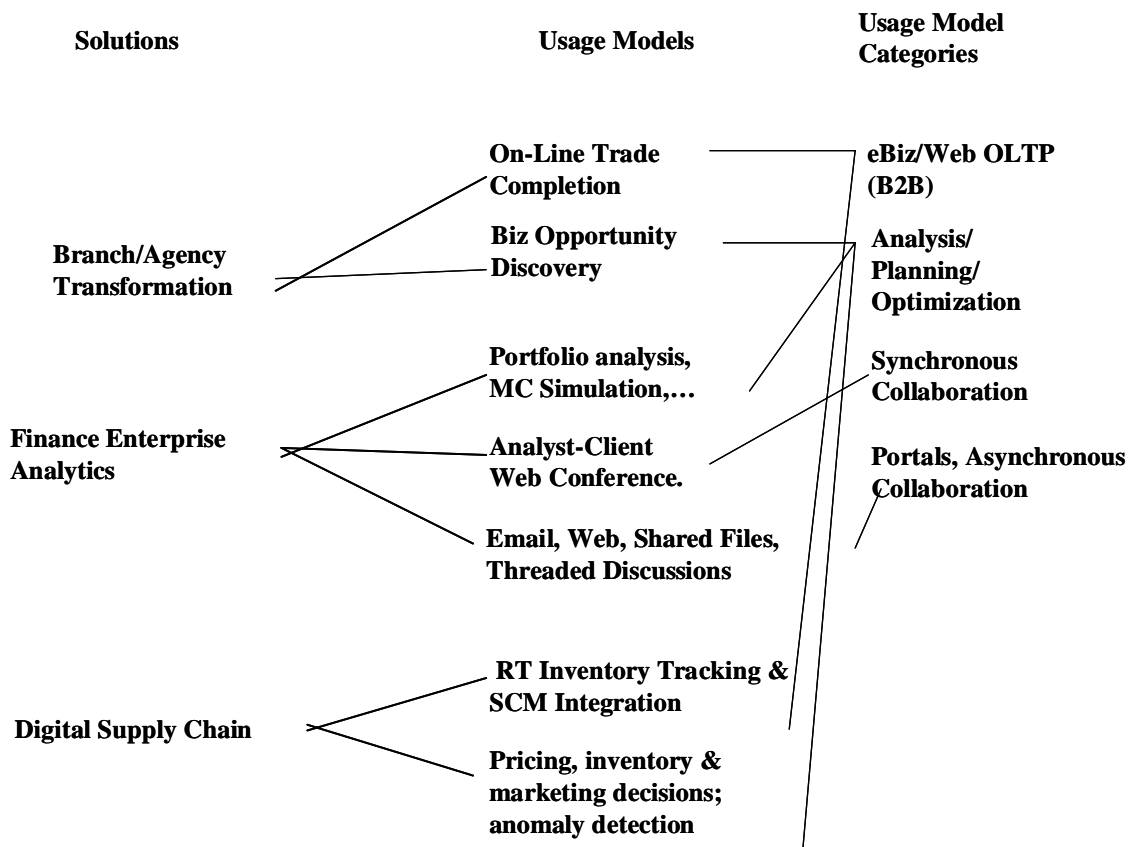


Figure 3: Identifying constituent usage models of solutions and categorizing them

Rationale for Usage Model Categorization

The usage models primarily help in identifying commonality across the solutions used in different industries. The solutions share a significant part of the software stack used to build similar usage models. For example, the eBiz/Web On-Line Transaction Processing (OLTP) usage model category invariably consists of Web servers, XML-based messaging, and links to enterprise applications and databases. Industry-specific applications (e.g., for on-line stock trading, student registration, or trading in electricity) might access these stacks in different patterns. These variations need to be comprehended in characterizing the usage model categories.

The second reason is the ability to map to industry-standard benchmarks from groups such as Transaction Processing Council (TPC) and Standard Performance Evaluation Corporation (SPEC). These benchmarks are widely used by IT hardware and software vendors today to evaluate the performance of the current enterprise platforms. Hence, it would be beneficial to use these benchmarks as the baseline and determine if and how the usage models differ in their characteristics.

Common Usage Model Categories

The most common usage model categories (based on how many solutions shown in Figure 2 they were part of) are shown in Figure 4. For each model category, we have also listed the forward-looking features that distinguish them from today's solutions. We have also selected specific solutions as representative of the usage model categories based on two criteria:

1. How aggressive a given industry is in pursuing the solution (e.g., the retail industry is a leader in adopting real-time inventory management while the financial services industry has been pushing the envelope in real-time analytics).
2. The IT market impact of the solution based on the IT spending expected for the solution.

The above information is based on an extensive survey of the industry leaders, conducted jointly by Intel and a research firm.

Usage Model Category	Vertical Industry	Solution	New Characteristics
eBiz/Web OLTP (B2B)	Retail	Digital Supply Chain	Real-time (Increased data volume, processing)
	Manufacturing	Collaborative Value Chain	Frequent DB syncs concurrent w/ OLTP Complex queries; app-level security, routing
Analysis	Financial Services	Channel Experience Consistency	Real-time DSS (Interactive, frequent/continuous ETL)
		Wealth Management & Compliance	Query across DB and XML Data mining
	Government	Homeland Security	Active data warehouse
	Manufacturing	Collaborative Value Chain	
Collaboration/ Portals	Government	e-Government Services, Digital City	Dynamic data gathering from DBs, Web services
	Financial Services	Wealth Management	Rendering large content (>10s of MB) using XSLT
	Manufacturing	Collaborative Value Chain	Security to address privacy and trust concerns Voice interactions Federated ID to allow single sign-on
Digital Content Processing & Delivery	Government	Homeland Security	Real-time/interactive image analysis
	Health Care	Secure Patient Framework	Secure image storage & processing
	Comm/Media	Content Delivery Infrastructure	
Technical Computing	Energy	Collaborative Exploration and Production	Real-time analysis Grids/Clusters
	Manufacturing	Collaborative Product Development	Increased data volume/XML
	Life Sciences	Computational Biology	

Figure 4: Common usage model categories

REAL-WORLD EXAMPLES

In this section, we describe real-world examples of solutions that embody some of the new characteristics. We also discuss their infrastructure needs and the software and hardware capabilities that would help in meeting the needs.

Real-Time Inventory Management in a Retail Chain

SAP described the increase in communication and computing needs of a retail enterprise with 1000 shops in migrating from a batch-oriented inventory management model using SAP's proprietary application messaging protocols to real-time inventory management using open, XML-based protocols [2].

Assuming a total of 33 million sales with three items per sale, the mySAP application communication from the 1000 retail shops to the central data center amounts to 2 GB per day, when the sales data are aggregated and communicated, once every 24 hours, using SAP's proprietary interface protocol.

Still aggregating and sending the data once every 24 hours, if the retail chain moves to an open XML-based interface, the data that need to be communicated and processed increases to approximately 20 GB.

Let us say the retail chain wants to react to sales trends faster. Aggregation delays the response. If information on every sale is sent to the data center to enable immediate response, about 200 GB of XML data has to be communicated and processed every day.

Here are some of the potential hardware and software capabilities apart from the increases in processing speed, cache, memory capacity and speed, and networking bandwidth that could help the retail chain implement the real-time inventory management capability:

- XML acceleration appliances, separate from the servers, may not help in all cases, particularly if large XML documents need to be moved. On-chip acceleration using special instructions or dedicated cores in a multi-core chip avoids communication latency overheads [3].
- Technologies such as InfiniBand Architecture (IBA) and Intel I/O Acceleration Technology could reduce the overhead in moving large amounts of data within the data center [4].
- Software changes to minimize or accelerate data type conversions or better leverage the hardware-specific features for handling different data types; and increased concurrency in XML processing could accelerate the solution.

Near-Real-Time Planning in Logistics

A logistics software vendor recently assessed the performance implications of supporting their customer's need to plan for filling orders on almost a continuous basis while simultaneously providing sub-second responses to the interactive RF terminals. This scenario exemplifies a couple of the new characteristics of the emerging eBiz/Web OLTP usage model:

1. Databases running more frequent synchronizations concurrently with the operational queries to ensure currency of data.
2. The same database also supporting complex, report-generation queries to support decision-making based on real-time data.

The minimum success criteria for this logistics software were 2.5 records/sec. for data download, 2.0 lines/sec. for order planning, and sub-second response time to the interactive users. Data download consisted of deleting a large number of rows in a database and inserting new ones. It was observed that a significant amount of time was being spent in deleting the rows. Using multiple threads for data deletion provided a significant reduction in time, particularly on Intel® Itanium®-based servers with EPIC architecture.

Currently, platforms are optimized for either throughput in transaction processing environments or for single job completion in analysis and planning environments. The industry benchmarks also emphasize one or the other. With the increased need to support analysis and planning in real- or near-real-time, such stove piping will not provide the best results.

Interactive Wealth Management Services

Traditional investment brokerage firms are faced with severe competition from Internet discount brokers who can charge lower transaction fees. The traditional firms are trying to leverage their strength in providing valuable financial advice to gain an advantage.

Charles Schwab, a pioneer in this field, wanted to provide a larger number of clients with objective financial advice at a fair price over multiple channels. They were faced with two requirements:

1. The wealth management solutions need to perform at interactive speeds for the financial advisors to work online with their clients.
2. The solutions need to be deployed on platforms based on standard, high-volume building blocks to keep the costs low.

A cluster of multiple IBM eServer xSeries 330* grid-enabled servers (using Intel® Xeon™ processors) using the Globus Toolkit for Linux*, reduced the processing time on the application of eight to ten minutes (and sometimes hours) to just 15 seconds.

® Itanium is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

* Other brands and names are the property of their respective owners.

® Intel and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

The use of clusters of servers to solve complex problems, where processing capacity can grow incrementally with demand, is finding increasing application in various areas including CAD/CAE and chemical analysis and search. Such a scale-out model lends itself more easily to solving some problems more than others (refer to section entitled “Scale-Out Virtualization: HPC for Enterprises,” below for details).

Secure Document Sharing for Global Manufacturing

A large manufacturer with factories and customers across the globe wanted to improve how it shared its product and process-related documents both internally and with its customers. It was faced with three challenges:

1. Delivering tailored documents to users instead of drowning them in documentation.
2. Minimizing the costs in generating and continually updating the documents with the rapid changes in products and processes.
3. Increasing concern over intellectual property and globalization demand a high level of security. However, poor usability and performance of the security solutions cause poor compliance by employees.

All these challenges were exacerbated by globalization. For example, this manufacturer dealt with a relatively small number of customers in North America and Europe. In Asia, they had to deal with a much larger number of smaller customers (more than 1000X).

To meet the first two requirements, the manufacturer moved towards a dynamic document-generation model. The data were retrieved from databases and Web services in XML format and rendered in HTML, PDF, or other formats for users to view.

Disk access was the first bottleneck in such a dynamic document delivery system. This was avoided by caching the XML content in memory. Rendering XML to HTML, etc., became the new bottleneck. These XML documents were typically in the 1 MB to 5 MB range. The eXtensible Style sheet Language Transformation (XSLT) scripts used to convert XML into HTML, etc., were typically in the 50 KB to 100 KB range.

The overheads in using specialized XSLT acceleration appliances prevented any improvements to them. However, scaling the solution out by adding Web servers to run the XSLT scripts was effective, as shown in Figure 5.

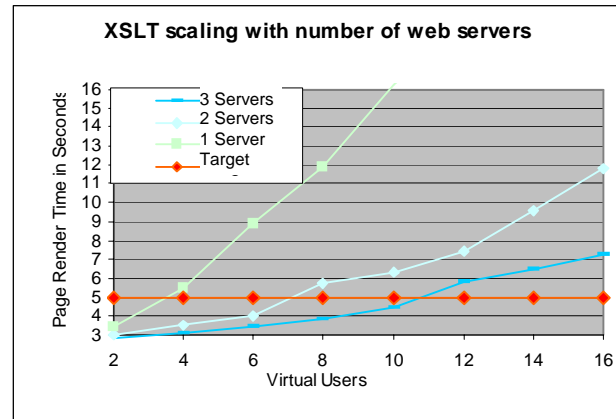


Figure 5: Response time improvement with the number of Web servers

To meet the needs of all its product groups and rapidly growing numbers of users across the world at a reasonable cost, significant speed-up in XSLT processing is needed. XSLT compilers and hardware support (without the overheads in moving the data around) have the potential to help.

The performance needs for security are equally significant. Currently, about 3000 documents are in a secure repository used by about 10,000 users. It is expected to grow to 100,000 documents used by 80,000 users in the near future. The personnel responsible for the solution think most of the users would use the secure repository only if the document access times are not excessive compared to non-secure access. Hence, they believe crypto acceleration is key to the success.

Other key requirements relate to ensuring the trustworthiness of the platform. They include securing private keys on both the servers and clients and ensuring that the client software stack is not compromised. The Trusted Platform Module [5] and La Grande technology [6] are expected to meet these requirements.

Scale-Out Virtualization: HPC for Enterprises

Grid computing technology has undergone a significant evolution over the past three or four years: the grid has been gradually moving from its High-Performance Computing (HPC) roots in university and government labs to more “mainstream” enterprise applications such as financial models and graphical rendering for motion pictures. We call grids applied in this domain “enterprise grids.” Most of the enterprise grids are run on server clusters within the data center.

The concept of enterprise grids is literally server virtualization turned inside out, and it represents the next

step in workload disintermediation (decoupling applications from the physical platforms that run them): server virtualization allows multiple logical servers to run in one physical server. Each logical server runs one application. Conversely, in an enterprise grid environment, it is possible to apply more than one server, a *node* in grid parlance, to an application. We call this “Scale-Out Virtualization.”

Enterprise grids of various sizes are getting deployed in different areas. Grids with 8-64 nodes are common for Computer Aided Design (CAD) and Electronic Design Automation (EDA) applications. Larger grids with up to 256 nodes are common in financial services, oil exploration, and pharmaceuticals.

Some of these problems are characterized as “embarrassingly parallel.” It is possible to partition these problems so that computation and the associated data sets for parts of the problem could be isolated to individual nodes, and hence there is very little communication between the nodes. Monte Carlo simulation for investment portfolio analysis is an example of such a problem. Today’s commercial servers and networks (100 MB or Gigabit Ethernet) could be used to solve these problems using large grids.

Other problems may not be so easily partitioned due to the need to move data between nodes or from memory to the CPU on a single node. EDA applications are examples of such problems. Messages could be “bundled” to minimize the penalty due to high network latency. This would require rewriting some of the applications. Alternatively, expensive interconnect technologies may be required.

Based on the data from several enterprise problems, we have derived a heuristic we call the Rule of 10: *the degradation in latency and bandwidth between two consecutive layers in the hierarchy should be no worse than a factor of 10 for all layers in a grid.* The on-CPU cache, main memory, disks, and network are the layers. For the “embarrassingly parallel” applications, this rule may not apply. For applications with significant data movement, the factor may have to be as low as 6. However, the Rule of 10 seems applicable to many cases.

Let us consider a cluster of servers connected by a Gigabit Ethernet as an example. The actual bandwidth for Gigabit Ethernet is about 100 MB/s. At the next lower level in the hierarchy, memory bandwidths of 3.2 to 6.4 GB/s are typical today in commodity servers. Hence, the bandwidth is degraded by a factor of 32 to 64. The degradation is even higher for latency: with memory latencies in the order of 100 to 150 nanoseconds, and the Ethernet network latencies around microseconds, the degradation factor is 700-1000. Hence, solutions with significant network communication need to be rewritten to bundle the

messages, or the Ethernet has to be replaced by networking technologies with lower latency, such as IBA.

There are ways to work around the Rule of 10. Intel future platforms, for example, are expected to offer on-CPU caches comparable to today’s memory in size, thus reducing the need for memory access and hence the importance of reducing the memory access latency for many solutions [3].

CONCLUSION

Wayne Gretzky, the legendary hockey player, once said that great hockey players go where the puck is going to be, not where it is. Well, the same applies to computing platforms. Great platforms meet what the solution requirements are going to be. Intel Corporation has identified the important solutions in vertical industries through extensive interactions with the industry leaders. These solutions were identified based on the significant value they offer to both the businesses that deploy them and their customers, as well as on the IT investments expected to deploy these solutions. We narrowed the solutions to those that are already being adopted by some leading businesses and are expected to be adopted by the majority of businesses in the 2009-2011 timeframe to balance the needs to be forward-looking yet accurate. We analyzed these solutions and identified the top usage model categories that are common to these solutions. These usage models differ significantly in their behaviors from the current solutions. Here are some examples of the differences:

- Competitive pressures to perform many operations in real-time (e.g., supply chain management, oil well analysis, medical imaging analysis, financial wealth management, and so on).
- Intellectual property, privacy, and regulatory compliance concerns, combined with globalization, drive a tremendous increase in the need for securing the platforms as well as the data that are being exchanged.
- XML, managed runtime environments, server virtualization, and autonomic management of the platforms are adopted increasingly to reduce the costs of developing and maintaining the solutions.

The above differences lead to new solution behaviors that are not adequately captured by the current industry-standard benchmarks. For example, the same database installation is increasingly required to support OLTP, decision support, and data synchronization concurrently.

We discussed several examples of real-world solutions that provide insights into these behaviors and their platform requirements. The current generation enterprise

platforms from Intel and its partners already meet some of these emerging requirements. Various technologies to support parallelism (e.g., Hyper Threading and EPIC), improved data center I/O technologies, large on-chip caches, and the technologies to improve the trustworthiness of the platforms are examples. Intel's current directions towards multi-core chips, LaGrande technology, larger on-chip caches, and faster access to larger amounts of data in memory are aligned with the solution requirements trends we have discussed. These platform technologies, the increasing ability to manage the platforms and the solution stacks running on them, and scale-out virtualization will make enterprise grids a viable deployment model for a larger cross section of solutions. The future platforms could potentially offer even greater value to these solutions through features such as support for accelerating XML processing and cryptographic operations on the chip, native support for higher-bandwidth-lower-latency I/O, and higher platform trust.

ACKNOWLEDGMENTS

The authors acknowledge contributions from Anne Bartlett, Dave Dempsey, Heather Dixon, and Joel Munter, their colleagues at Intel Corporation. Thanks are also due to the valuable review and suggestions from Mark Chang, Jackson He, and Joel Munter.

REFERENCES

- [1] *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*, Geoffrey A. Moore, July 1999, Harper Collins Publishers, New York, NY.
- [2] "SAP Position Paper," *W3C Workshop on Binary Interchange of XML Information Item Sets*, September 2003, Santa Clara, CA.
<http://www.w3.org/2003/08/binary-interchange-workshop/21a-W3CWorkshop-SAPPositionPaper.pdf>*
- [3] Borkar, Dubey, et al., "Platform 2015: Intel Processor and Platform Evolution for the Next Decade, 2005."
ftp://download.intel.com/technology/computing/archin/nov/platform2015/download/Platform_2015.pdf.
- [4] Grun, Paul, "The Changing Nature of Data Center I/O," 2003,
<http://www.intel.com/technology/pciexpress/devnet/docs/datacenterio.pdf>.
- [5] Meinschein, Robert, "Trusted Computing Group Helping Intel Secure the PC," *Technology@Intel Magazine*, January 2004.
<http://www.intel.com/technology/magazine/standards/st01041.pdf>.
- [6] "LaGrande Technology,"
<http://www.intel.com/technology/security/>.

AUTHORS' BIOGRAPHIES

Krishnamurthy Srinivasan is an architect in the Digital Enterprise Group in Intel Corporation. His current interests include understanding the platform requirements at various levels of the emerging enterprise solutions through performance characterization. Earlier, he played a key role in the evaluation and adoption of several software technologies in Intel's Planning & Logistics and Corporate IT groups. He made significant technical contributions to the development of Intel's Third-Generation e-Business vision. He managed Intel's Web service technology development and participated in industry-standards groups. During his sabbatical in 2002, he taught at the Indian Institute of Information Technology, Bangalore. He has a Ph.D. degree in Engineering from the Georgia Institute of Technology. His e-mail is Krishnamurthy.Srinivasan at intel.com.

Raj Ramanujan is a senior principal engineer in the Digital Enterprise Group in Intel Corporation and directs the platform initiatives and pathfinding activities. Raj is well recognized for his expertise and leadership in platform architecture definition. He has extensive experience in component (processor and chipset) micro-architecture complemented with experience in detailed performance analysis of architecture/micro-architecture alternatives. His e-mail is Raj.K.Ramanujan at intel.com..

Michael Amirfathi is a security architect in the Information Services and Technology Group in Intel Corporation. His current interests include understanding the enterprise security needs for protection of digital information and developing enterprise rights management solutions and services. He has extensive experience in architecting and developing enterprise applications for Aerospace, Finance, and High Technology industries. He has an M.S. degree in Engineering from Utah State University. His e-mail is Michael.Amirfathi at intel.com.

Enrique Castro-Leon is currently an enterprise architect and technology strategist for Intel Solution Services with 22 years at Intel working in OS design and architecture, software engineering, high-performance computing, platform definition, and business development. He has taught at the Oregon Graduate Institute, Portland State University, and he has authored over 30 papers. He holds Ph.D. and M.S. degrees in Electrical Engineering and Computer Science from Purdue University. He is also the founder of The Neighborhood Learning Center, a non-profit educational organization. His e-mail is Enrique.G.Castro-Leon at intel.com.

Copyright © Intel Corporation 2005. This publication was downloaded from <http://developer.intel.com/>.

Legal notices at <http://www.intel.com/sites/corporate/tradmarx.htm>

For further information visit:

developer.intel.com/technology/itj/index.htm