



Architecting the Era of Tera

Defining the need for a computing architecture to perform multimodal object recognition and synthesis over massive data sets

**Research &
Development
at Intel**

February, 2004

Author:

Intel Research and Development

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life saving, life sustaining applications.

Intel may make changes to specifications and product descriptions at any time, without notice.

Copyright © Intel Corporation 2004

* Other names and brands may be claimed as the property of others.

Architecting the Era of Tera

February 2004

The Digital Transformation

The industry is seeing a convergence of digital technologies around consumer electronics (CE), communications and computing platforms. Today's "connect anytime/anywhere" society is fueling this digital transformation. The amount of data in existence is doubling every 24 months — a sort of data equivalent of Moore's Law. Worldwide, data growth is increasing so quickly it's now measured by the exabyte — 10^{18} bytes, or a billion billion bytes. This escalation in digital technologies is changing the way the world works and plays.

History demonstrates that computing capabilities do not always advance linearly; rather there have been several "leaps" in computing capabilities that have had profound impacts. In the 1980s we saw the first PC, which brought computing from the realm of large corporations and academia and made it accessible to computing enthusiasts. With the development of the integrated floating point processor in the early 1990s, PCs moved from displaying glowing text on a black screen to color graphical user interfaces (GUIs). With the almost simultaneous development of graphical Web browsers, computing became appealing to the masses, and the resultant explosion in personal computing continues today across the globe.

In 1995, the introduction of MMX™ technology heralded the first media-specific architectural constructs. This "digitization of everything" helped fuel the growth of the Internet, and brought multimedia to the desktop. In 2004, wireless computing and the concept of the digital home are gaining traction, with the home PC acting as a "hub" used by families to serve up digital music, photos, and videos on the wirelessly connected CE device of choice. Continued multimedia architectural enhancements and new technologies such as Intel® Centrino™ Mobile Technology are enabling these new usages.

Disruptive technologies are seldom recognized until they cause a complete realignment of an entire business sector or industry, at which point everyone acknowledges what seems to be a logical and natural progression. Once the new technology becomes the status quo, innovation skeptics always seem to emerge, proclaiming we now have all the processing power we'll ever need. Yet history has taught us that new technologies quickly appear to take advantage of advances in processor power.

Intel believes the industry is again on the cusp of a tremendous opportunity.

Defining the Era of Tera

The digital transformation is serving as the catalyst for the rapidly increasing number of technological innovations. But without another leap forward in computing capabilities, the opportunity to harvest and take advantage of this wealth of data will be missed. The magnitude of this opportunity cannot be overstated. Digital data is pervasive and has been growing at close to 30 percent a year for the last several years. How much digital vs. nondigital data is out there? The picture on the right shows the amount of data currently in existence¹.



¹ How Much Information? 2003: <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/>

Architecting the Era of Tera

February 2004

As can be seen in the graphic, the amount of digital data dwarfs the nondigital data, and virtually all new data created is being recorded digitally. And there is no reason to believe the rate of digital data growth will abate.

What are the implications of this explosion of digital data? Imagine:

- Being able to quickly search through the digital images you have and sort by categories like pictures of your daughter in braces, or every picture with your pet dog—all quickly gleaned from the tens of thousands of digital photos or videos you have recorded.
- Your local doctor having a complete history of tests, x-rays and medical information that has been compiled over your lifetime, and being able to determine if a speck on a recent x-ray is significant or not based on data such as blood tests, calcium levels, and the deterioration of your skeletal system.
- The amateur filmmaker deeply analyzing every “classic” movie of a given era and scanning terabytes of video streams to be able to determine the seemingly abstract characteristics that made these films so appealing.

In short, we are looking at digital immersion that goes beyond the simple creation and consumption of digital media. We are looking at the opportunity that every minute, everywhere, everyone will be connected, enhanced and supported within this digital universe.

Defining Tera Era Workloads

To develop architectures capable of delivering tera-level computing, it is necessary to determine the classes of computing capabilities required. At Intel we classify these processing capabilities, or workloads, into three fundamental types: recognition, mining and synthesis or simply RMS. This model comprehends any type of processing capabilities for both existing and future computing workloads.

By understanding the types of workloads that are required to support tera-level computing, it becomes possible to define and develop architectures that will satisfy these workloads.

Recognition

Recognition: The ability to recognize patterns and models of interest to a specific application requirement.

Recognition is a capability that is required to recognize a pattern or a model in the large amounts of data in a specific application. In large data sets it is possible that thousands or even millions of patterns or models may be present. But not all would be interesting to a user or a specific application.

For example, in a September 11-like preterrorist situation, an incredible amount of data may be present in hundreds, if not thousands, of separate databases. These databases could include such information as airline reservations, credit card and banking transactions, gas purchases, and so forth. These data sources, along with FBI and CIA databases, could be analyzed to look for suspicious behaviors based on very subtle patterns or models of data that can characterize a specific behavior. Recognizing such a data pattern or model could help detect a terrorist threat before an attack occurs.

Recognition workloads are a cornerstone of tera-level computing, as the ability to find information within vast, massive data sets will be critical.

Architecting the Era of Tera

February 2004

Mining

Mining: The ability to examine or scan large amounts of real-world data for patterns or models of interest.

In mining, intelligent methods are used to inference useful information and relationships from large amounts of data. Consider the terrorist example used in the earlier section; recognizing a specific model of data alone is not useful to predict and avoid such a catastrophe. The systems need to infer and extract useful information from large data sets. This is most relevant when predicting a behavior based on a collection of well-defined models of data. Reorganization and mining are two capabilities that are both closely dependent on and complimentary to each other.

Synthesis

Synthesis: The ability to synthesize large data sets or a virtual world based on the patterns or models of interest.

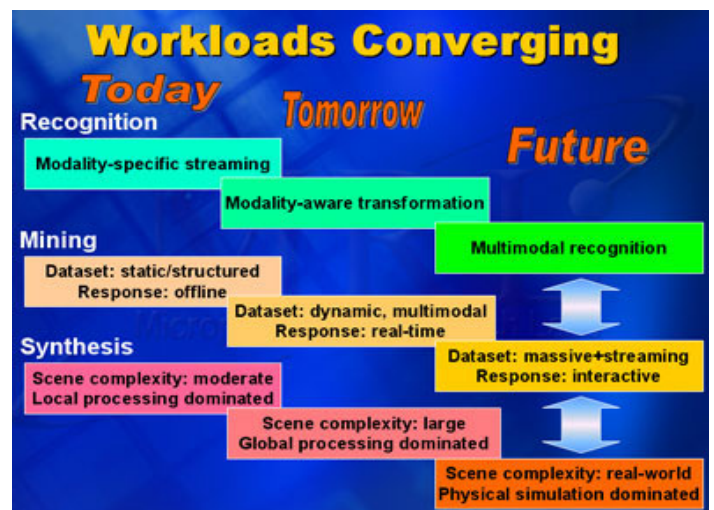
Synthesis can be thought of as creation. It can be the creation of a “virtual world,” or the creation of a summary or conclusion about data that has been analyzed. Working in conjunction with recognition and mining, synthesizing workloads over massive data sets will bring capabilities to computing platforms that are not possible today. Ray tracing is one example of a tera-level application. A billion polygons are required for photo-realistic graphics. Today’s supercomputers can do ray tracing, but it takes hours or even days to deliver one scene. It will take teraflops of performance to deliver real-time photorealistic animation.

In the case of the terrorist example discussed previously, as soon as the system predicts a specific kind of attack is underway, the system will synthesize an appropriate response and proactively alert the appropriate state and federal authorities about a possibility of such an event.

Workload Convergence

Recognition and Mining require enormous algorithmic processing power for pattern and model recognition, inference and extracting useful information from data, as well as high I/O bandwidth. Synthesis requires the processing of several kinds of algorithms that are dependent on a specific application of interest. For example, mining or recognition of large data sets might require extensive use of a genetic algorithm or something similar, while photorealistic rendering of images require processing of ray-tracing algorithms in real time, which will in turn require the processing of billions of polygons per second.

Today, architects are constrained due to available performance, and are forced to use very different algorithms that fit a given performance budget for each of the workloads; R, M, S. By performance budget, we mean a combination of resources including number of transistors on the die, power and heat requirements, and other factors that force architectural choices. These choices result in architectures optimized for specific classes of workload. Enterprise applications need huge I/O bandwidth and fast integer manipulations. State-of-the-art rendering requires graphics engines that are dedicated to pixel manipulations. But these workloads,



Architecting the Era of Tera

February 2004

or modal-specific architectures, are limiting. You can't render graphics on an enterprise server, and you can't perform database queries on the latest dedicated graphics engine.

As we examine RMS workloads and predict the needs of tera-level computing, we believe these workloads—recognition, mining, and synthesis—will require similar platform capabilities that are independent of the applications. The applications can be anything: statistical computing, collaborative filtering, physical simulation, behavioral modeling, Internet searching or the real-time rendering of photorealistic images, but they have common characteristics:

1. Teraflops of processing capabilities
2. High I/O bandwidth
3. The ability to efficiently execute or adapt to a specific type of workload

With tera-levels of performance, it becomes possible to bring these workloads together on one architectural platform, using common kernels. No longer would it be necessary to optimize architectures for one workload type at the expense of the others. Tera level computing platforms will have a single architecture to satisfy all R, M S workloads.

Enabling the Era of Tera

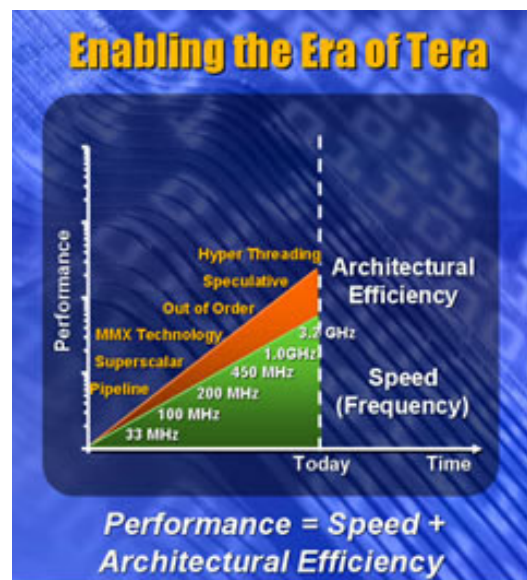
The task appears daunting: develop the computing architecture that will deliver 10x, or even 100x the capabilities of today's platforms. What will enable the next leap in computing capabilities?

Intel is developing a billion-transistor processor. Moore's Law predicts the transistor count or "building blocks" that are used by computer architects. But Moore's Law is not a performance law, it is a transistor count law that simply says that every 24 months or so the number of transistors on an integrated circuit doubles. This exponential growth and ever-shrinking transistor size result in increased performance and decreased cost. In less than 25 years the PC industry has gone from 5-MHz to 4-GHz processors—a thousandfold increase in the frequency of the chip. Coupled with architectural innovations such as superscalar, MMX technology, speculative execution and Hyper-Threading Technology, we've seen tremendous leaps in computing performance capabilities.

In summary, the past 25 years have seen processor-speed and architectural innovations as the drivers for improvement. However, this equation needs to change for the next 25 years.

Challenges

As the graph shows, speed improvements have been dominated by frequency increases rather than by architectural improvements. But the question today is what will this graph look like in the future? Moore's Law is one of the most well known in relation to computing, but is not the only one. Other, less-friendly laws of physics are confronting the industry. As clock frequencies increase and the feature size of the transistors decreases, obstacles are increasing in key areas.



Architecting the Era of Tera

February 2004

- **Power:** Power density is increasing at a rate that implies that tens of thousands of watts per centimeter (w/cm^2) will be needed to scale the performance of Pentium® processor architecture over the next several years. But that would produce more heat than the surface of the sun. The Power Exponent is setting hard limits to frequency increases.
- **Memory Latency:** Memory speeds are not increasing as quickly as logic speeds. During the i486™ CPU days, the requirements were 6-to-8 clocks per cycle to access memory. Today's Pentium processors require 224 clocks, about a 20x increase. These wasted clock cycles can nullify the benefits of frequency increases in the processor.
- **RC Delay:** RC (resistance-capacitance) delays on chip are become increasingly challenging, as well. As feature size decreases, the delay due to RC is increasing. In 65nm and smaller nodes, the delay caused by a one millimeter RC delay is actually longer than a clock cycle. Intel chips are typically in the 10-to-12 millimeter range, taking 15 clock cycles of delay to go from one corner of the die to the other, again negating many of the benefits of frequency gains.
- **Scalar Performance:** Experiments with frequency increases of various architectures such as superscalar, CISC (complex instruction set computing), and RISC (reduced instruction set computing) are not encouraging. As frequency increases, instructions per clock actually trend down, illustrating the limitations of concurrency at the instruction level.

From these observations, we can conclude that increases in performance will need to come primarily from architectural innovations, as the benefits of frequency increases become more and more limited. Monolithic architectures are reaching their practical limits to deliver orders of magnitude performance increases. A fresh approach in platform architecture is needed to deliver tera-level computing.

New Architectural Paradigm for the Era of Tera

In the past, mini- and mainframe computers have provided many of the architectural ideas currently seen in personal computers. Today, we're again examining other architectures for ways to meet these new challenges. High-performance computers (HPC) are delivering teraflop performance, although in very limited niche markets and at costs prohibitive to all but select government and academic institutions. The industry challenge is to make this level of processing available on platforms as accessible as today's PC.

Concurrency at Multiple Levels

Architectural innovation is nothing new to Intel. For several decades Intel has worked to improve performance through architectural innovations such as Intel® Super-Pipelined RISC Technology, superscalar, MMX technology, out-of-order execution engines, speculative execution and now Hyper-Threading Technology.

The key lesson from high-performance computing is the idea of multiple levels of concurrency and execution units. Instead of one execution unit, having four, eight, 64 or in some cases hundreds of execution units, or a multicore platform is the only way to achieve tera-level computing capabilities.

Rather than big, full-chip implementations, multicore architectures localize the implementations in each core and effective relationships with the "nth" level—second and third levels of cache. This creates enormous challenges in platform design. Having multiple cores and multiple levels of cache will scale the performance exponentially, but the issues related to memory latency, RC interconnect delay, and power are still issues—

Architecting the Era of Tera

February 2004

so platform-level innovations that address these issues are needed. This is an architecture that will comprehend changes from the circuit, through the microprocessor(s), platform and entire software stack.

In addition, as seen by the SPECint experiments², concurrency at the microprocessor level alone is not sufficient. A massively multicore architecture in which each core has multiple threads of execution with minimal memory latency, RC interconnect delay, and controlled thermal activity is needed to deliver teraflop performance.

The three attributes that will define this new architecture are scalability, adaptability, and programmability.

Scalability

Scalability 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 increasing demands of the RMS workloads.

There are two ways to scale performance. Historically, the industry has “scaled up”, by increasing the capabilities and speed of single processing cores. There is also “scaling out”, which implies adding multiple cores and threads of execution to increase performance. The best known types of “scaling out” architectures are today’s High Performance Computers which have hundreds, if not thousands of cores.

Scalability Example

An example of "scaling up" can be found in the helper thread technology. Helper threads implement a form of user-level switch-on-event multithreading on a conventional processor without requiring explicit OS or hardware support. Helper threads can improve single thread performance by performing judicious data prefetching while the main thread is stalled waiting for a cache miss to be serviced.

In today's platforms, processors are often idle. In fact for server workloads, processors can spend almost half of the total execution time waiting for memory. Given that a memory access can incur over 200 clock cycles, the challenge and opportunity is to use this waiting time in an effective way. Helper threads are an answer. Helper threads can be executed by a processor during the waiting time to prefetch data. Helper threads make use of otherwise idling processor resources to do prefetching in order to prevent subsequent cache misses. Experiments in Intel's labs showed that helper threads can eliminate significant amount of cache misses, up to 30%, and can improve performance of memory intensive workloads on the order of 10-15%

This is a simple example of scalability within a single execution unit. Future architectures will need to massively exploit scalability at multiple levels; multiple threads across multiple execution units.

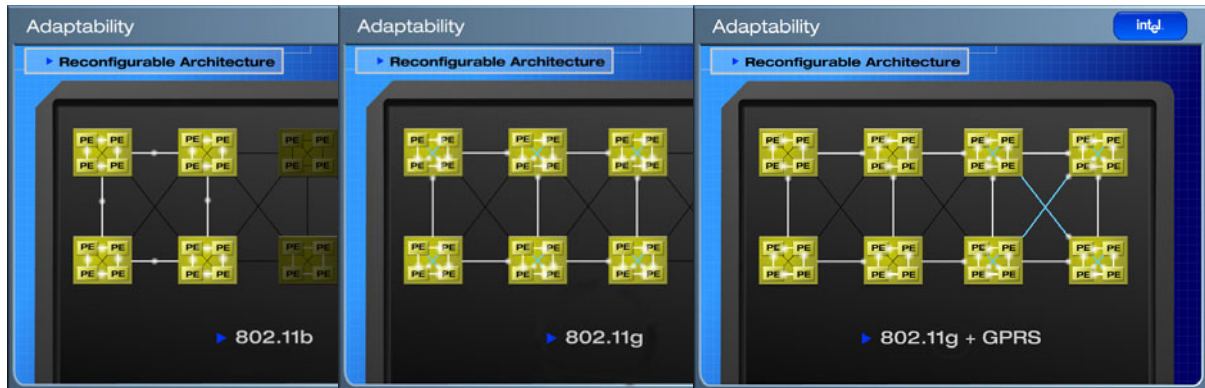
Adaptability

An adaptable platform proactively adjusts to the intrinsic workload type and the application requirements. As the tera-era level—applications and RMS workloads of the future converge in the platform, the platform must be adaptable to any type of RMS workload. That will be true convergence. Multicore architectures not only provide scalability, but also the foundation for adaptability. While we’re not there yet, the following example offers an overview of such adaptability through special purpose processing cores called processing elements. In this example, we show an architecture that adapts to different protocols, such as 802.11* a, b, g, Bluetooth*, and GPRS (general packet radio service).

² Standard Performance Evaluation Corporation: <http://www.spec.org/osg/cpu2000/CINT2000/index.html>

Architecting the Era of Tera

February 2004



This graphic illustrates multiple processing elements, each of which can be considered a processing core. These processing elements can each be recruited to perform a specific radio algorithm function, such as programmable logic array (PLA) circuits, Viterbi decoders, memory space, and other functions that would be appropriate for wireless applications. In general, these processing elements can be a digital signal processor (DSP) or an application-specific integrated circuit (ASIC), as per the processing needs.

With this kind of architecture, the platform can dynamically configure so that it operates for a workload like 802.11b by meshing a few processing elements. In another configuration, the platform can reconfigure itself to support GPRS or 802.11g or Bluetooth by interconnecting different sets of processing elements. This type of architecture can support multiple workloads like 802.11a, GPRS, and Bluetooth simultaneously. This is the power of the multicore microarchitecture. The system can dynamically reconfigure and move across different workloads, and can support multiple workload types simultaneously in the microarchitecture itself.

Programmability

The challenge of bringing high performance computing to the masses has been in defining parallelizable applications, and the needs of software environment to understand the underlying architecture. A programmable system will have workload characteristics like concurrency, data structures, and synchronization, and communications requirements communicated to the hardware. Simultaneously, architectural characteristics—like allocation of cores and threads according to the resource requirements of the workloads—are communicated back up to the applications.

The idea behind programmability is to enable the platform and the workload to recognize resources and application or workload characteristics in order to adapt the platform to the specific task at hand. Intel has already started down this path with compilers such as those developed for Itanium® processors, and we have started to see this in certain high-level language constructs, such as DOACROSS, COBEGIN, and COEND. However, much more needs to be done in order to take advantage of the new architectural features in these computing platforms.



Architecting the Era of Tera

February 2004

Conclusion

The industry is on the cusp of another leap in computing capabilities that will dramatically impact virtually everything we touch, our individual lives, and our society. This digital transformation offers another opportunity to make profound improvements for the world. The types of applications, usage models, and problems that tera-level computing can tackle are revolutionary. The tera era will see changes as dramatic as those brought about by the printing press, the automobile, and the Internet.

This is a huge challenge for the entire industry. Intel is leading the charge by developing the technologies and architectures that will enable tera-era computing, but we want you to join us in making this vision a reality. What we've seen is the tip of the iceberg, and the best is yet to come.