Volume 12 Issue 03 Published October 2008 ISSN 1535-864X DOI: 10.1535/itj.1203

Intel® Technology Journal

The Original 45nm Intel Core™ Microarchitecture

Intel Technology Journal Q3'08 (Volume 12, Issue 3) focuses on Intel® Processors Based on the Original 45nm Intel Core™ Microarchitecture: The First Tick in Intel's new Architecture and Silicon "Tick-Tock" Cadence

More information, including current and past issues of Intel Technology Journal, can be found at:
http://developer.intel.com/technology/itj/index.htm

Volume 12 Issue 03 Published October 2008 ISSN 1535-864X DOI: 10.1535/itj.1203

Intel® Technology Journal

The Original 45nm Intel Core™ Microarchitecture

Articles

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

Richard Bowles, *Publisher*
David King, *Managing Editor*

Intel® processors based on original 45nm Intel Core™ microarchitecture is the focus of this *Intel Technology Journal* (Vol. 12, Issue 3). This family of processors was originally referred to by the codename Penryn. Improvements in the Penryn processor family are numerous and lead to benefits for mobile, desktop, and server platforms. Penryn processors are the first to exploit the advantages of Intel's 45nm process technology.

The issue marks a transition in the Journal's management. After a dozen years of success under the leadership of Lin Chao, the Journal is now going to be a part of Intel Press, where it will join with reference books that are *by engineers, for engineers*. For now, the mission of the Journal remains the same: to provide a Web-based journal to publish state-of-the-art perspectives written by Intel engineers. Lin Chao is rightfully proud of the Journal she created. My Intel Press team and I intend to maintain and enhance the Journal's reputation.

Overall performance improvements for the original 45nm Intel Core Microarchitecture is the topic of the first article. Taking a broad perspective, the article explains improvements in SSE4.1 instructions, larger caches, faster divide techniques, and better load balancing across cache boundaries. Advances are reported in dynamic acceleration technology, silicon-level support for virtualization, and deep power down technology to improve performance per watt.

The content architect for this issue of the Journal is Varghese George, who is a member of the Mobility Group at Intel Corporation and co-author of an article that drills down on power management enhancements for processors in the Penryn family. Improvement in performance per watt is a primary design criterion at Intel, and elements of the Penryn architecture enable progress on this front.

The second paper drills down on changes in the Penryn processor microarchitecture aimed at speeding up super scalar processing. Named SSE4.1, new shuffle procedures join with additional SSE instructions to improve performance for graphics and video applications. Extending Intel processors to match emerging media requirements is a longstanding tradition, and this paper marks further progress.

Greater miniaturization leads to design challenges for packaging, which is the focus of the next article. These engineers explain how they designed a solution that balanced requirements for compatibility with predecessor design, cost-efficient manufacturing for Intel and for OEMs, and for thorough validation. Along the way, these members of a global Intel team shifted work following the sun to create a 24 hour work cycle.

The trend to further miniaturization is discussed in a paper about the design challenges in building a Wi-Fi Wireless communications daughter board in a smaller form factor. The paper stands as an example of Intel's focus on platforms and not just processors, and the need to keep platform components well integrated and balanced. And, as is typical, the new design is not 10-percent smaller, but rather just half the size of its predecessor.

In the last two articles of this Journal, Intel engineers explain the impact of the Penryn processor family on mobile systems and desktop systems. Miniaturization, performance, power consumption, and packaging issues all come together to describe how the Penryn processor family will influence the next generation of Intel platforms and, in turn, the next generation of computer systems.

THIS PAGE INTENTIONALLY LEFT BLANK

Foreword

Ofri Wechsler,

Intel Fellow, Manager, MG CPU Architecture

A few years ago Intel Corporation decided to base our entire IA microprocessor product line on the Intel® Core™ microarchitecture. The essence of the decision was to utilize the Core technology that was originally designed for mobile computing and to enhance Core in ways that would allow it to span both the desktop and server markets. In conjunction with the converged core decision, we have also outlined a new development model for Intel® IA microprocessors which we named the "Tick-Tock" model.

Improvements in compaction (Tick) would be followed with improvements in the microarchitecture (Tock). The fundamental objective of the Tick-Tock model was to allow Intel to take advantage of the converged core in its processor development and to synchronize and maximize the utilization and output of our development teams. Our thought was that this model will enable Intel to produce significant, predictable microprocessor improvements year after year.

The new Intel® Core™ processors (code-named Penryn) represent the first Tick for the Tick-Tock model. Based on historical track records, one might have expected Penryn processor improvements to be primarily caused by the shift to our new 45nm process technology (the "Tick"). That is, this first implementation would be a process technology lead vehicle with only moderate improvements attributed to changes other than process technology.

Surprisingly, this is far from being the case. The Penryn development team internalized the Tick-Tock strategy and was able to deliver an enormous number of improvements above and beyond a traditional "compaction" project. The Penryn processor development teams have set a very high bar for future Intel Tick processors.

The new Penryn processor family is marching in the footsteps of its predecessor, codename Merom, and continuing to improve the computation efficiency. The novel new hardware divider and super-shuffle units, the SSE4.1 instruction set, as well as many more microarchitectural enhancements of the Penryn processor are all aimed at the same goal: deliver more and more performance to the end user within the same or even lower power envelopes.

In addition to the performance and performance efficiency improvements that the Penryn processor family provides, it is also demonstrating a revolution in power and thermal management. Penryn processors introduce the novel Deep Power Down state, which allows the processor to draw minimal current when the processor is idling. The technological foundation that was put in the Penryn processor will allow future Intel processors to eliminate the idle power component completely from the energy equation as we move into more advanced power delivery schemes. And finally, with Penryn processors, we are introducing Intel® Dynamic Acceleration Technology that allows power and thermal budgets to move dynamically within the dual- and quad-core complexes and to boost single-thread performance even further in a restricted power envelope.

I am very proud of the newest and youngest member of the Intel Core microarchitecture family. The Penryn processor implementation teams have demonstrated for the first time what a Tick processor should look like and have expanded Intel's unquestioned leadership across the mobile, desktop and server market segments.

THIS PAGE INTENTIONALLY LEFT BLANK

Technical Reviewers for Q3 2008 ITJ

Subramani Bhamidipati, Digital Enterprise Group
Martin G. Dixon, Digital Enterprise Group
Stephen A. Fischer, Mobility Group
Benny Getz, Mobility Group
Steve Ghasemi, Digital Enterprise Group
Steve Gunther, Digital Enterprise Group
Sanjeev Jahagirdar, Mobility Group
Jason Ku, Mobility Group
Rob Milstrey, Mobility Group
Rajiv Mongia, Mobility Group
Asim Nisar, Mobility Group
Gunjan Pandya, Digital Enterprise Group
Shmuel Ravid, Mobility Group
Ronak Singhal, Digital Enterprise Group
John Wallace, Mobility Group

THIS PAGE INTENTIONALLY LEFT BLANK

Original 45-nm Intel® Core™ 2 Processor Performance

Asim Nisar, Mobility Group, Intel Corporation
Mongkol Ekpanyapong, Mobility Group, Intel Corporation
Antonio C Valles, Software Solution Group, Intel Corporation
Kuppuswamy Sivakumar, Server Platform Marketing Group, Intel Corporation

Index words: PenrynΔ, 45-nm Core 2 processor, Performance, SSE4.1, EDAT

Citations for this paper: Asim Nisar, Mongkol Ekpanyapong, Antonio C Valles, Kuppuswamy Sivakumar "Original 45nm Intel® Core™2 Processor Performance" Intel Technology Journal. http://www.intel.com/technology/itj/2008/v12i3/1-paper/1-abstract.htm (October 2008).

ABSTRACT

The 45nm Intel® Core™2 family of processors, codename PenrynΔ, improves upon the performance of Intel Core 2 processors through new microarchitecture features, a larger cache, new instructions, and enhanced power- and thermal-management schemes. This paper presents measured performance data that show the microarchitectural benefits of the Penryn family of processors on key applications and benchmarks. In addition, this paper showcases performance improvements achieved by new SSE4 instructions on a variety of media, imaging, and 3D workloads. The Penryn family of processors also introduced new power- and thermal-management schemes. This paper discusses performance improvements achieved by these enhanced thermal-management features in thermally limited platforms such as mobile thin and light and small form-factor computers.

INTRODUCTION

Performance is an integral part of product definition and success. Intel sets very aggressive performance targets to deliver products with compelling performance to the end user. While considerable effort is placed on functional validation of Intel® processors, Intel also employs significant time and effort to ensure that the processor performance meets expectations at every stage of the product development cycle from concept to silicon arrival to product launch. All design decisions are weighed against performance impact, and appropriate tradeoffs are made. As a result of this extensive effort, Intel delivered a product with record-breaking performance on a wide range of client and server applications.

In this paper, we present information on performance delivered by products based on the 45nm Intel® Core™ 2 family of processors, codename PenrynΔ. Please see [1] for a detailed architectural description of some of the new microarchitectural features. We begin with an overview of major performance features and then provide an in-depth discussion of measured performance improvements on a wide range of mobile and desktop products. We conclude by presenting performance and energy-efficiency improvements achieved on server platforms built with Penryn processors.

PENRYN MICROARCHITECTURE ENHANCEMENTS

The PenrynΔ family of processors is the next generation of Intel® processors based on the Intel® Core™ 2 microarchitecture, implemented on Intel's 45nm, Hi-k metal gate process technology. Frequency improvements, within existing power and thermal envelopes, over previous-generation processors, a larger L2 cache, microarchitectural enhancements, and improvements in power- and thermal-management schemes deliver improved performance per watt and energy efficiency for a broad range of client and server applications. The Penryn family of processors also added 47 new SSE4 instructions that can improve the performance of audio, video, image-editing applications, video encoders, 3-D applications, and games.

Microarchitecture enhancements that improve performance in the Penryn family of processors include the following:

- **Larger Cache**: Penryn processors include up to a 50 percent larger L2 cache with a higher degree of associativity that further improves the hit rate, maximizing its utilization. Dual-core Penryn processors feature up to a 6-MB L2 cache and quad-core processors up to a 12-MB L2 cache.

- **Faster Divider:** Penryn processors provide faster divider performance, roughly doubling the divider speed over previous generations through the inclusion of a new, faster divide technique called Radix 16.

- **Super Shuffle Engine:** Shuffles (the repositioning of bits) is a common operation in image- and video-editing applications. By implementing a full-width, single-pass, 128-bit-wide shuffle unit, a processor from the Penryn family of processors can perform full-width shuffles in a single cycle and is 3 times faster than previous-generation processors. The Super Shuffle Engine improves the performance of Intel Streaming Single Instruction Multiple Data (SIMD) Extensions (SSE), Streaming SIMD Extensions 2 (SSE2), Supplemental Streaming SIMD Extensions 3 (SSSE3), and Streaming SIMD Extensions 4 (SSE4) instructions, and this will benefit a wide range of applications including imaging and video applications, games, 3D modeling, and high-performance computing.

- **Inclusion Filter**: An Inclusion Filter was added in the Penryn family of processors to enhance the existing inclusion logic that was limiting server performance.

- **Renamed RSB:** The Renamed Return Stack Buffer (RRSB) increases return prediction accuracy and improves performance.

- **CLI STI Performance Tuning:** In the Penryn family of processors, the Clear Interrupt Flag (CLI) and Set Interrupt Flag (STI) macroinstructions were optimized to perform an execution pipeline serialization only when a new IF value is consumed and only if the new value is not yet updated, instead of post-serializing on every CLI or STI. This improves throughput of CLI-STI pairs by 2.5 times over previous-generation technology.

- **Enhanced Intel® Dynamic Acceleration Technology (EDAT):** EDAT is a power-management feature added to mobile processors that improves energy efficiency by dynamically increasing the performance of active core(s) when not all cores are utilized.

- **Enhanced Intel® Virtualization Technology:** Virtualization partitions or compartmentalizes a single computer so that it can run separate operating systems and software. This virtual partitioning better leverages multi-core processing power, increases efficiency, and cuts costs by letting a single machine act as many virtual 'mini' computers. The Penryn family of processors speeds up virtual machine transition (entry exit) times by an average of 25 percent to 75 percent. This is all done through microarchitecture improvements and requires no virtual machine software changes.

NEW INSTRUCTIONS (SSE4.1)

While many of the microarchitecture enhancements in the PenrynΔ family of processors can be utilized without recompilation, media-related kernels will achieve the maximum performance and power-efficiency gains by recompiling with the Intel compiler and or manually optimizing code, using the new SSE4.1 instructions introduced in the Penryn family of processors.

Intel works closely with industry partners including independent software vendors (ISVs) to understand their performance needs and to improve their applications' performance. The Penryn family of processors' new instructions, SSE4, are a customer-driven response to improve performance on audio-, video-, and image-editing applications, video encoders, 3-D applications, and games. In this section we discuss performance results achieved by using the SSE4 instructions.

Intel® HD Boost technology

Intel HD Boost, the combination of SSE4 instructions and the Penryn family of processors' Super Shuffle Engine, can provide large speedups on a wide range of applications. The following instructions in particular can provide significant benefits to video, imaging, and audio applications.

- There are twelve new integer format conversions that can perform a conversion such as Byte-> Double-Word in one cycle with one instruction.

- The new MPSADBW instruction performs eight sums of absolute differences (SAD) in one instruction. This is twice what the SSE2 PSAD instruction can do.

- The new PHMINPOSUW instruction can be used to perform a horizontal minimum search to locate a minimum unsigned word in an XMM register or a __m128 data type.

The MPSADBW and PHMINPOSUW SSE4 instructions can be used to significantly improve motion vector search algorithms (also known as block matching) used in motion estimation for video applications. An Intel whitepaper [2] showcases how to use these two instructions for block matching. The whitepaper reports a 1.6× to 3.8× performance improvement (see (Figure 1)).

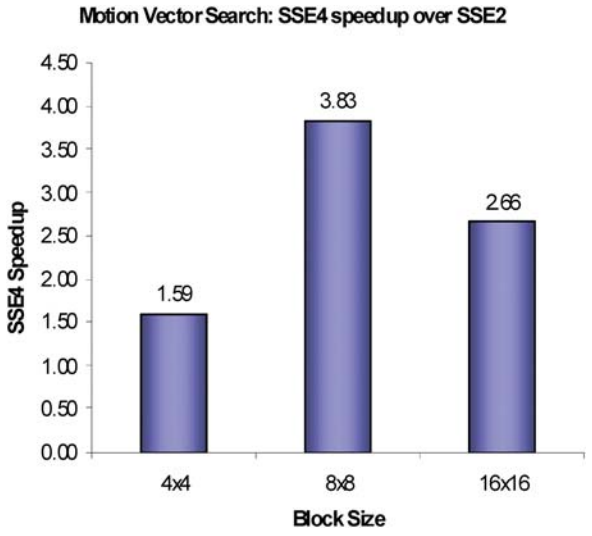

Figure 1: ***SSE4.1 function-level speedups to motion vector search, also known as block matching, used in motion estimation.***

The integer format conversions are commonly used in imaging and video applications. For example, they can be used when converting RGBA from four bytes to four floats prior to computation on a pixel. One SSE4 convert instruction can do the same thing as four SIMD instructions did previously, as shown.

```
SSE2:
pmovd xmm0, m32
pxor xmm7, xmm7
punpcklbw xmm0, xmm7
punpcklwd xmm0, xmm7
cvtdq2ps xmm0, xmm0

SSE4:
pmovzxbd xmm0, m32
cvtdq2ps xmm0, xmm0
```

Conditional moves, blends, early outs

Branches have always been one of the limitations of SIMD code. SSE4 provides new instructions (six Blend instructions plus a PTEST instruction) that can be used to replace either some branches or existing lengthy SIMD code written to get around branches.

The Blend instructions can be used to replace conditional move flows. For example, the PBLENDVB instruction can replace the PAND PANDN POR instructions commonly used in conditional moves where masks are created from a comparison instruction. Another SSE4 instruction, PTEST, can be used as an early out. It is able to compare the entire 128-bit register in one pass. This instruction can be used for conditions that are meant to be infrequent such as divide-by-zero exceptions. One of the benefits of these new instructions is that they provide the compiler more vectorization opportunities; that is, they provide more opportunities to optimize the high-level code by compiling it to use the SIMD instructions.

However, the real benefit of the Blend and PTEST instructions is when multiple branches in a loop can be replaced with multiple Blend and PTEST instructions. The Mandlebrot [3] code shown in Figure 2 is an example that demonstrates how multiple branches can be replaced with multiple PTEST and Blend instructions. In the SSE4 implementation (Figure 3) notice the use of two PTEST instructions:

```
if(_mm_test_all_ones(_mm_castps_si128(vmask)))
if(_mm_test_all_zeros(_mm_castps_si128(vmask),
   _mm_castps_si128(vmask)))
```

and 3 Blend instructions:

```
sx = _mm_blendv_ps(x + sx*sx − sy*sy,sx,vmask);
sy = _mm_blendv_ps(y +_F_TWO_*old_sx*sy,
                   sy,vmask);
iter = I32vec4(_mm_blendv_epi8(iter+_I_ONE_,
                iter,_mm_castps_si128(vmask)));
```

```
void mandelbrot_C()
{
  int i,j;
  float x,y;
  for (i=0,x=-1.8f;i<DIMX;i++,x+=X_STEP) {
      for (j=0,y=-0.2f;j<DIMY/2;j++,y+=Y_STEP) {
          float sx,sy;
          int iter = 0;
          sx = x;
          sy = y;
          while (iter < 256)
          {
              if (sx*sx + sy*sy >= 4.0f)
                break;

              float old_sx = sx;
              sx = x + sx*sx - sy*sy;
              sy = y + 2*old_sx*sy;
              iter++;
          }
          map_C[i][j] = iter;
      }
  }
}
```

Figure 2: ***C implementation of Mandlebrot.***

```
__declspec(align(16)) float _INIT_Y_4[4] = {0,Y_STEP,2*Y_STEP,3*Y_STEP};
F32vec4 _F_STEP_Y(4*Y_STEP);
I32vec4 _I_ONE_ = _mm_set1_epi32(1);
F32vec4 _F_FOUR_(4.0f);
F32vec4 _F_TWO_(2.0f);

void mandelbrot_F32vec4() {
  int i,j;
  F32vec4 x,y;

  for (i=0,x=F32vec4(-1.8f);i<DIMX;i++,x+=F32vec4(X_STEP)) {
    for (j=0,y=F32vec4(-0.2f)+*(F32vec4*)_INIT_Y_4;j<DIMY/2;j+=4,y+=_F_STEP_Y) {
      F32vec4 sx,sy;
      I32vec4 iter = _mm_setzero_si128();
      int scalar_iter = 0;
      sx = x;
      sy = y;
      while (scalar_iter < 256) {
        int mask = 0;
        __m128 vmask = _mm_cmpnlt_ps(sx*sx + sy*sy,_F_FOUR_);
        if (_mm_test_all_ones(_mm_castps_si128(vmask)))
          break;

        F32vec4 old_sx = sx;
        if (_mm_test_all_zeros(_mm_castps_si128(vmask),_mm_castps_si128(vmask))) {
          sx = x + sx*sx - sy*sy;
          sy = y + _F_TWO_*old_sx*sy;
          iter +=  _I_ONE_;
        }
        else {
          sx = _mm_blendv_ps(x + sx*sx - sy*sy,sx,vmask);
          sy = _mm_blendv_ps(y + _F_TWO_*old_sx*sy,sy,vmask);
          iter = I32vec4(_mm_blendv_epi8(iter + _I_ONE_, iter,_mm_castps_si128(vmask)));
        }
        scalar_iter++;
      }
      _mm_storeu_si128((__m128i*)&map_SSE4[i][j],iter);
    }
  }
}
```

Figure 3: ***SSE4 (using F32VEC4) implementation of Mandlebrot.***

By using the new SSE4 instructions on the Mandlebrot code, the Mandlebrot performance improves by 2.8 times over the C implementation.

Graphics building blocks

SSE4 instructions can be used to speed up graphical applications such as games. The DPPS DPPD instruction can be used to speed up collision detection and common vector matrix operations such as vector normalization. A detailed example of collision detection and usage guidelines of the DPPS DPPD instructions is discussed in [1]. The example showcases a 1.5x speedup in collision detection by using the DPPS instruction and the EXTRACTPS instruction.

A common problem in graphics applications is 'Data Swizzling' or converting from an Array-of-Structures (AOS) data layout implementation to a more SIMD-friendly Structures-of-Array (SOA) data layout in order to use SIMD. Users have to weigh the cost of these conversions before deciding if it is worth using SIMD. By using the INSERTPS instruction, the data-swizzling operation on the next-generation, Penryn microprocessors now take fifteen cycles per four vertices, down from 23 cycles on the Intel 65-nm Core 2 Duo microprocessors, codename Merom. [5] (see Figure 4).

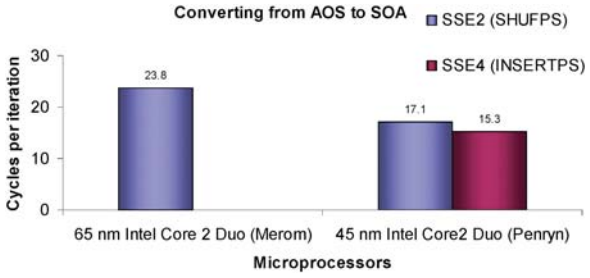

Figure 4: ***Data swizzling improvements per four vertices (one iteration converts four vertices).***

Another potential game improvement is the streaming load instruction MOVNTDQA. This instruction provides a fast method to execute a 16-byte aligned load from Write Combining (WC) memory, such as graphics memory, with a non-temporal hint such that the cache is not polluted.

This instruction can provide a $5\times$ to $7\times$ [6] memory throughput performance increase.

Intel tools

Intel's Integrated Performance Primitives (IPPs), Version 6.0 has over a thousand functions optimized with SSE4. The average speedup across all SSE4-optimized IPP functions vs. SSSE3-optimized IPP functions is $1.12\times$. Figure 5 shows which IPP categories have been optimized with SSE4 and their SSE4 speedup over SSSE3.

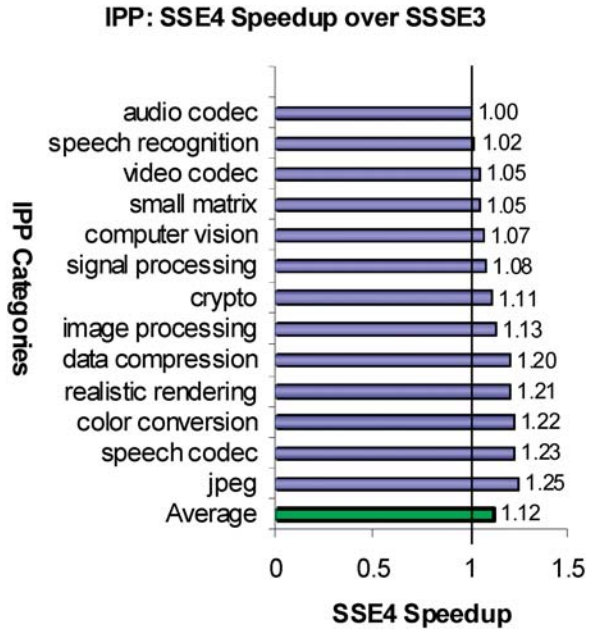

Figure 5: ***Intel Performance Primitive categories and their SSE4 speedups over SSSE3 versions.***

SSE4 instructions also enhance the compiler's ability to vectorize certain loops. Vectorization is when the compiler optimizes a loop to use SIMD instructions including SSE4 instructions. The Intel Compiler, Version 10.0 and later, can be used with the QxS compiler flag to generate SSE4-optimized code specifically for the Penryn family of processors.

SSE4 instructions combined with the Super Shuffle Engine can significantly improve the performance of imaging, video, audio, multimedia, and high-performance computing applications. Users can add these instructions to their applications via assembly code or use Intel tools such as the Intel Compiler 10.0 and IPPs. For detailed information on the SSE4 instructions, including throughput, latency, and optimization guidelines, please see the Intel 64 and IA-32 Architectures Optimization Reference Manual [5] and the instruction manuals [7,8].

INTEL® CORE™ PROCESSOR

Desktop and Mobile

The PenrynΔ family of processors, including dual- and quad-core desktop processors and a dual- and quad-core mobile processor are branded as 'Intel Core processors.'

Desktop and mobile systems built with 45nm Intel® processors, based on Penryn Core architecture, give gamers, researchers, and serious multitaskers a significant performance boost over previous-generation processors. In this section of the paper we present measured performance data on next-generation Intel Core™ 2 Extreme processors, the new addition to Intel's high-end desktop product line-up. We compare this processor with previous-generation Intel Core 2 Extreme processors on key client benchmarks and real-world applications.

Microarchitectural performance

Improvements

Figure 6 shows a comparison between the Intel Core 2 Extreme QX6850 processor (3.00 GHz, 1333 MHz FSB, 8 MB L2) and the next-generation 45nm Intel Core 2 Extreme QX9650 (3.00 GHz, 1333 MHz FSB, 12 MB L2) on an Asus* P5EX38 board at the same frequency and platform configuration. 45nm quad-core performance is up to 6 percent faster than previous-generation technology on SPEC 2006.

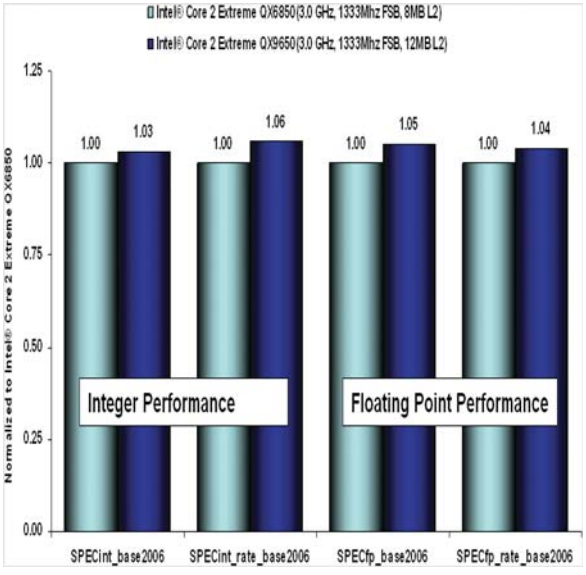

Figure 6: ***Quad-core performance comparison for SPEC CPU2006 at same frequency (estimated SPEC CPU2006 as measured on pre-production systems).***

Video and audio encoding are becoming increasingly important in the world of personal computing. Home-editing of videos and sound recordings are among the popular applications as is standard archiving of DVD material. As shown in Figure 7, the 45nm Intel Core 2 Extreme QX9650 provides a significant boost over previous-generation processors at the same frequency and platform configuration for some of the media-encoding applications. For example, Premiere* Pro CS3 software from Adobe is used to create high-quality visual and editorial effects; it allows users to add color correction, lighting, and other effects such as

audio filters and more, with fast, flexible, built-in tools. As shown in Figure 7, the new Qx9650 is 20 percent faster than the Qx6850 in rendering 210 frames to the disk using this Adobe software. Fathom* is an advanced encoding platform product from Inlet Technologies that is used by media companies to encode content for streaming over the Internet or broadcasting over the air. As shown in Figure 7, Intel measures a 23-percent improvement with new 45nm processors for Fathom to transcode 1080i YV12 high-definition video (HDV) to a 1080i VC1 format. Intel measures a 40-percent improvement for Qx9650 over previous-generation technology for a Pegasys* TMPGenc XPress 4.4 encoder to convert original Variable Bit Rate encoded, 76 second, 29.97fps, 1440×1080 video clips into HDV format MPEG video with 1440×1080 resolution, 29.97 fps, and 25 000 Kbs Constant Bit Rate encoding. Another example is VirtualDub* software, which is a video capture processing utility that uses the DivX* 6.7 software for encoding movies. Virtualdub* 1.7.1 and later with DivX 6.7 are optimized for SSE4 instructions and provide a very noticeable 60-percent performance gain over previous-generation processors that use encoding in SSE2 to convert to the higher-compression DivX format.

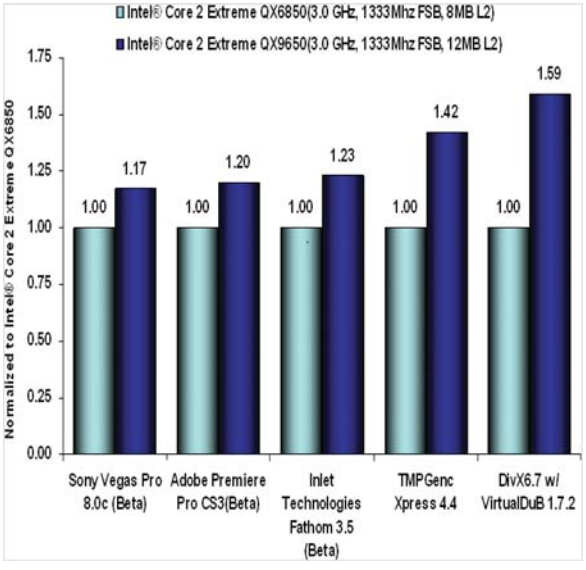

Figure 7: ***Quad-core performance comparison for video encoders at same frequency.***

Figure 8 shows a similar comparison for some of the popular games. The new 45nm Intel® Core™ 2 Quad processor is roughly 10 percent faster than previous-generation processors at 1024×768. Even when looking at just the two quad-core processors that run at the same FSB and clock speeds, the Intel 45nm Core 2 processors have a clear lead over previous-generation processors. The larger cache, new microarchitectural features, the Penryn high-definition boost, the Super Shuffle unit, and the SSE4 instructions that were discussed in the previous sections all contribute to the increased performance.

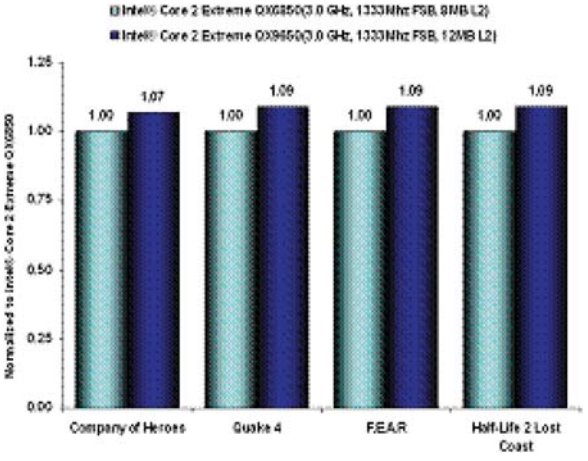

Figure 8: ***Quad-core performance comparison for games at same frequency.***

Frequency and platform improvements

Figures 9 and 10 compare the Intel Core 2 Extreme QX6850 processor (2.93 GHz, 1066 MHz FSB, 8 MB L2) on an Intel 975BX2 board with DDR2 800 RAM with an Intel Core 2 Extreme QX9770 processor (3.20 GHz, 1600 MHz FSB, 12 MB L2 Penryn). Please see Table 1 for a detailed system configuration. This is a more realistic comparison as it takes into account core enhancements, frequency improvements achieved with new 45nm technology, and other platform improvements that were added to support core enhancements. A new 45nm Intel Core 2 Quad-based platform provides double-digit gains on compute-intensive workloads such as SPEC 2006 and the Sysmark 07 Preview that reflect usage patterns of business users in the areas of video creation, E-learning, 3D modeling, and office productivity. Intel measures a 17 percent improvement for Cinebench's multi-threaded rendering test and roughly a 20 percent improvement for Quake 4* and Half Life 2*. A video-encoding application such as DivX and TMPGenc* see a 50 percent to 80 percent gain. The increased frequency and 1600-MHz FSB improves the system bus and memory bandwidth and are the significant contributors to the performance difference. The enhancements in the Penryn family of processors are setting milestones in desktop computing performance.

Table 1: Detailed system configuration for the results shown in Figures 9 and 10.

Processors	Intel® Core™ 2 Extreme QX6850 8 MB L2, 3.0 GHz, 1333 MHz FSB	Intel® Core™ 2 Extreme QX9770 12 MB L2, 3.2 GHz, 1600 MHz FSB
Memory	Deluxe Dual channel DS Corsair 2 GB (2 × 1 GB) DDR3-1333 9-9-9-24	Dual channel DS Corsair CM3X1024-1600 C7DHXIN XMP 2 GB (2 × 1 GB) DDR3-1600 7-7-7-20
Graphics card	1 × G8800 GTZPCle graphics	
Motherboard	Asus* P5E3 X38 Deluxe board	
Hard disk	Seagate 320 GB NCQ SATA	
BIOS	Beta 0504, INF:8.4.0.1016, Graphic: NV163.69	

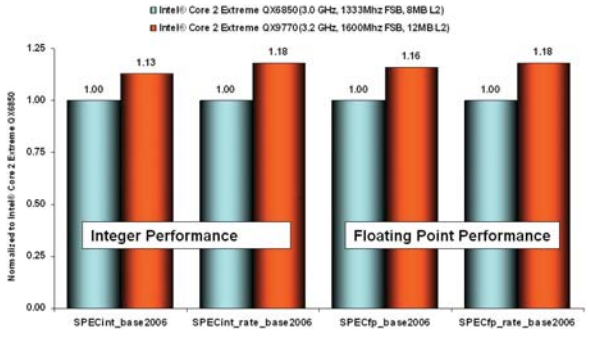

Figure 9: ***QX9770 (45 nm) comparison with QX6850 (65 nm) or SPEC 2006.***

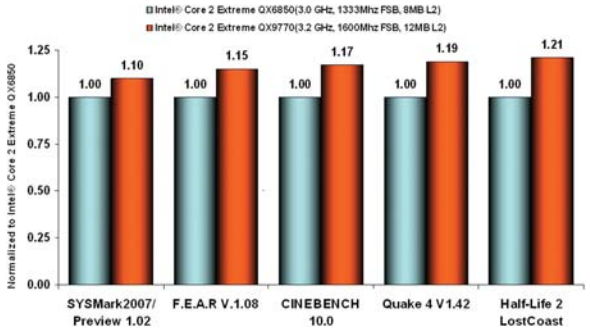

Figure 10: ***QX9770 comparison with QX6850 for games and general purpose applications.***

Enhanced Intel® Dynamic Acceleration Technology

Performance data presented in Figures 6–10 are based on desktop system measurements, but mobile platforms built with 45nm cores will see similar performance improvements over previous-generation platforms. Mobile processors, however, operate at lower frequencies because of tighter power and thermal limitations. In this section we illustrate how 45nm enhancements in Intel® Dynamic Acceleration Technology (DAT) improve mobile platform performance. DAT is a power-management feature that can improve system performance by increasing the frequency of active core(s) when at least half of the cores in a multi-core processor are inactive and thermal headroom is available. DAT was introduced in the 65-nm Intel Core 2 mobile processors, but in the Penryn family of processors we further enhanced DAT performance and energy efficiency by reducing the number of transitions in and out of DAT, reducing transition overhead in high-interrupt-rate scenarios. In the Penryn family of processors, we also extended DAT support to quad-core mobile processors. Architectural implementations and more details about these enhancements in Intel's Enhanced Dynamic Acceleration Technology (EDAT) are discussed in [4]. Single-threaded applications running on a dual-core or quad-core processor based on the Penryn family of processors, or two single applications (or a two-threaded application) on a quad-core processor, can take advantage of EDAT.

Figures 11-12 illustrate EDAT performance on an Intel Core 2 Quad processor for SPEC 2006, games, and multimedia applications on pre-production mobile platforms with 2 GB of DDR3 memory, a 1066 MHz FSB, a 120 GB hard disk, and a baseline frequency of 2.4 GHz.

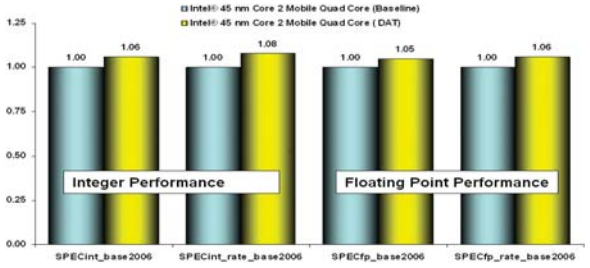

Figure 11: ***EDAT performance improvements (estimated SPEC 2006 as measured on pre-production hardware).***

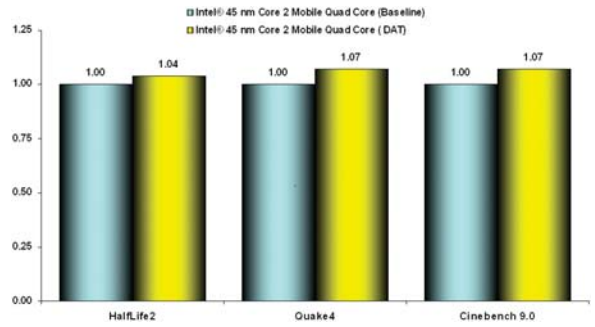

Figure 12: EDAT performance improvements (multi-media and games as measured on pre-production hardware)

To fit QC processors into the mobile Thermal Design Power (TDP) envelope, Intel had to reduce the processor's operating frequency, which in turns gives a wider dynamic range in terms of thermal headroom to operate at EDAT frequencies for one- and two-core operations. The performance gains for EDAT are solely from frequency scaling. The amount of frequency improvement varies from product to product. Single-threaded applications, or two single-threaded applications, or an application with two worker threads with good frequency scaling running on a quad-core processor, can see up to a 10 percent performance boost from EDAT.

INTEL XEON® PROCESSORS

New servers, workstations, and high performance computing (HPC) systems are built with new quad-core Intel Xeon processors 5400 series that are based on the 45nm PenrynΔ core technology. Intel's 45nm technology packs 820 million transistors into the Intel Xeon processor 5400 series. The chip is smaller than the previous-generation Intel Xeon processor 5300 series (214 mm^2 vs. 286 mm^2), which had 582 million transistors. More transistors on new 45nm technology means more capability, performance, and energy efficiency.

Increased performance

The Intel Xeon 5400 series, based on technology from the Penryn family of processors, featuring a larger 12-MB L2 cache, delivered a strong performance gain to the already stable and shipping server platform based on the Intel 5000 series chipset. A drop-in into the existing platform, the 5400 series, added up to a 21 percent performance increase over the previous-generation, quad-core Intel Xeon processor 5300 series for mainstream server benchmarks at the highest frequency level (comparing Xeon X5460 at 3.16 GHz to Xeon X5365 at 3 GHz). Figure 13 shows the comparison on a range of server benchmarks. Figure 14 shows performance comparisons to the previous generation at the same clock frequency (3 GHz) achieving up to a 19 percent performance increase and highlighting the benefits of the improvements in the Penryn family of processors.

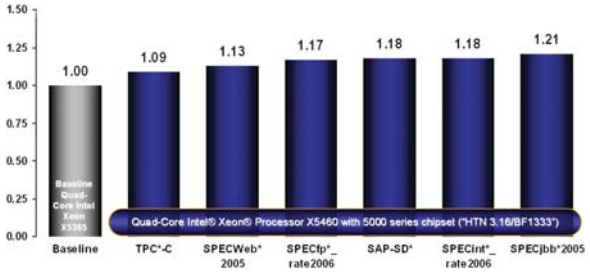

Figure 13: ***Comparison of server benchmarks with those of the previous generation.***

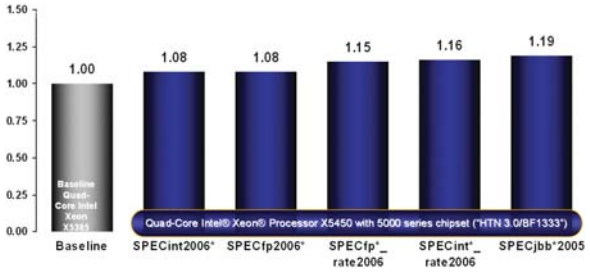

Figure 14: ***Comparison of server benchmarks to those of the previous generation at the same frequency on the same platform.***

The Penryn family of processors' architecture presents several compiler optimization opportunities. These opportunities include tuning for the new ISA, a larger cache, and hardware pre-fetching. The overall gain for the SPEC CPU2006 benchmark suite is shown in Figure 14 as 15–16 percent on the 'rate' benchmark. But gains across the individual components could be as high as 57 percent. Figure 15 shows some of the highlights across the Integer and Floating-point component workloads. All the results shown are on the peak result metric (SPECint_rate2006 and SPECfp_rate2006).

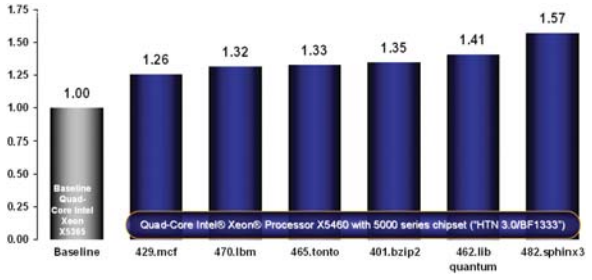

Figure 15: ***Gains on specific SPEC CPU2006 'rate' components at same frequency.***

To leverage the full potential of the 45nm micro-architecture, a new platform targeting the HPC market segment was launched with the new Intel 5400 chipset that could run at the faster FSB speed of 1600 MHz. The additional bandwidth delivered by the platform is critical for the HPC segment. Figures 16a and b show results on key HPC workloads on segments such as manufacturing, financial services, energy, weather and climate modeling, electronic design automation (EDA), and life sciences. It is important to note that the gains achieved by these applications were due to a combination of microarchitectural enhancements and the increased platform bandwidth.

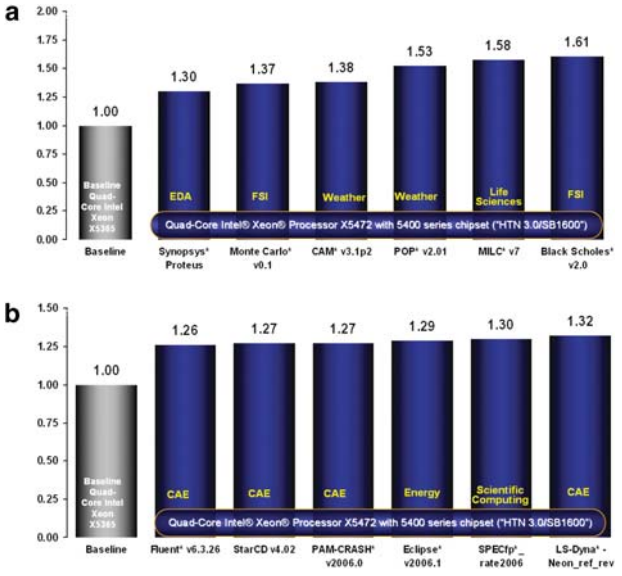

Figure 16: ***(a) HPC benchmarks. (b) HPC benchmarks.***

Improved energy efficiency

Servers based on the quad-core Intel Xeon processor 5400 series also maximize data centers performance and density through improved energy efficiency. As shown in Figure 17, these processors can deliver up to 38 percent more performance per watt in the same platforms and at the same system power level. The platform power in this chart is based on measured average power value at the steady-state window of the benchmark run. For this comparison we used the most energy-efficient mainstream processor SKU from each of the processor families. In this case, that would mean the Intel Xeon E5450 running at 3 GHz compared to the Intel Xeon E5345 running at 2.33 GHz, both at an 80 W TDP rating. The 45-nm, Hi-k-based processor also lowers the idle power significantly. Results on the new industry standard SPECpower*s_ssj2008 benchmark, which is the first comprehensive benchmark to measure energy-efficient performance across a load-line including idle power, highlights the energy-efficient performance of the Penryn family of processors. Table 2 shows the top-10 list for this benchmark, all occupied by platforms based on the Penryn family of processors' architecture.

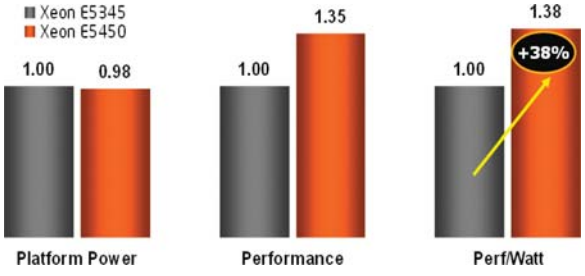

Figure 17: ***Energy efficiency-SPECjbb2005 Benchmark.***

Table 2: Top ten results on the first Industry Standard benchmark for energy efficiency–SPECPower*_ssj2008 (as of June 10, 2008).

Rank	Sponsor	Overall ssj_ops/ Watt	Platform	# of Sockets	Processor(s)	Processor Microarchitecture
1	FSC	1124	TX150 S6	1	Intel® Xeon® X3360	Intel® "Penryn"
2	IBM	1054	X3200 M2	1	Intel® Xeon® X3360	Intel® "Penryn"
3	FSC	1018	TX150 S6	1	Intel® Xeon® X3360	Intel® "Penryn"
4	HP	930	DL180 G5	2	Intel® Xeon® L5420	Intel® "Penryn"
5	IBM	926	X3350	1	Intel® Xeon® X3360	Intel® "Penryn"
6	IBM	913	X3250 M2	1	Intel® Xeon® X3350	Intel® "Penryn"
7	Inspur	910	NF290D2	2	Intel® Xeon® L5420	Intel® "Penryn"
8	IBM	854	X3450	2	Intel® Xeon® E5462	Intel® "Penryn"
9	Dell	800	PE R300	1	Intel® Xeon® L5410	Intel® "Penryn"
10	HP	778	DL180 G5	2	Intel® Xeon® E5450	Intel® "Penryn"

SPECpower results from http://www.spec.org/power_ssj2008/results/power_ssj2008.htmlas of June 10, 2008.

Enhanced virtualization

Virtualization partitions or compartmentalizes a single computer so that it can run separate operating systems and software. These partitions can better leverage multi-core processing power, increase efficiency, and cut costs, by letting a single machine act as many virtual 'mini' computers. Consolidating applications onto fewer systems not only results in better multi-core utilization but improves performance density. In the Penryn family of processors, virtual machine transition (entry exit) times show an improvement of between 25 percent and 75 percent. Based on virtualization benchmark results on different VMMs, the Penryn family of processors provide up to a 20 percent performance gain when compared to previous-generation platforms. Figure 18 shows comparisons of different benchmarks such as VMmark and vConsolidate running various VMMs such as VMware ESX Server, Parallels Virtuozzo, and Virtual Iron 4.0.

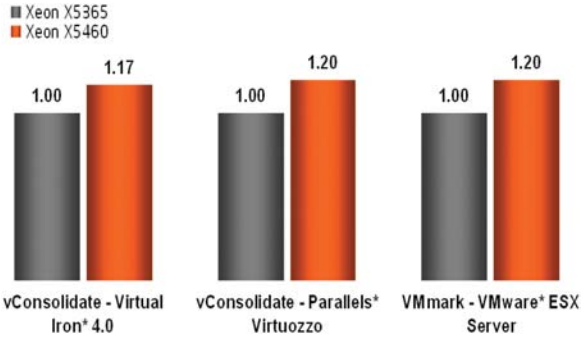

Figure 18: ***Virtualization performance.***

CONCLUSION

The PenrynΔ family of processors brings record levels of performance to the end user through its larger cache, new microarchitectural features, new instructions, and enhanced power- and thermal-management schemes. These processors, manufactured on Intel's 45nm, Hi-k metal gate process technology, not only provide significant performance improvements over the previous-generation processors, but they also provide building blocks for software to be created to take advantage of this power and deliver new levels of functionality to end users.

ACKNOWLEDGEMENTS

The authors thank all the researchers, architects, designers, validators, and software engineers who took the PenrynΔ family of processors from a vision to a real product. Special thanks go to Mike Fard, Jeff Reilly, Ronen Zohar, Daniel Shea, and Eric Palmer who provided performance data and or examples used in this paper.

REFERENCES

[1] Coke, Jim. "Improvements in the Intel® Core™ 2 PenrynΔ Processor Family Architecture and Microarchitecture." Intel Technology Journal, Volume. 12, Issue. 3, 2008.

[2] Kuah, K. "Motion Estimation with Intel® Streaming SIMD Extensions 4 (Intel® SSE4)," at http://softwarecommunity.intel.com/articles/eng/1246.htm.

[3] Mandelbrot set. http://en.wikipedia.org/wiki/Mandelbrot_set.

[4] Jahagirdar, S. "Power-Management Enhancements in 45 nm Intel®Core™ Microarchitecture." *Intel Technology Journal*, Volume 12, Issue 3, 2008.

[5] *Intel® 64 and IA-32 Architectures Optimization Reference Manual*. http://www.intel.com/products/processor/manuals/.

[6] Jha, A. and Yee, D. "Increasing Memory Throughput With Intel® Streaming SIMD Extensions 4 (Intel(4) SSE4) Streaming Load." At http://softwarecommunity.intel.com/articles/eng/1248.htm.

[7] *Intel® 64 and IA-32 Architectures Software Developer's Manual, Volume 2A: Instruction Set Reference, A-M*. At http://www.intel.com/products/processor/manuals/.

[8] *Intel® 64 and IA-32 Architectures Software Developer's Manual, Volume 2B: Instruction Set Reference, N-Z*. At http://www.intel.com/products/processor/manuals/.

AUTHORS' BIOGRAPHIES

Asim Nisar is a Senior Architect with Intel's Mobile Microprocessor Group in California, USA. He led the 45nm Intel® Core™ 2 performance modeling, analysis, performance projections, pre-silicon and post-silicon performance validation activities. Prior to this, he was involved in defining and enhancing several microarchitectural features. He has an M.S. degree from the Georgia Institute of Technology, USA. His e-mail is asim.nisar at intel.com.

Mongkol Ekpanyapong is a Senior Architect with Intel's Mobility Microprocessor Group in California, USA. He received his B.E. degree from Chulalongkorn University, Thailand, his M.E. degree from the Asian Institute of Technology, Thailand, his M.S. and Ph.D. degrees from the Georgia Institute of Technology, USA. Mongkol was involved in performance modeling and validation for the Intel® Core™ 2 Duo processor architectures. His current focus is on performance modeling for the next-generation Intel microprocessor. His e-mail is mongkol.ekpanyapong at intel.com.

Antonio C Valles is a Senior Software Engineer in Intel, Chandler, AZ focusing on broad and in-depth pre-silicon and early-silicon tests of Intel microprocessors and chipsets. Antonio has created multiple internal pre-silicon and post-silicon tools and kernels for performance analysis and coordinates development of tuning guidelines for the processors. He received his B.S. degree in Electrical Engineering from Arizona State University in 1997. His email is antonio.c.valles at intel.com.

Kuppuswamy Sivakumar (Siva) is a Marketing Manager for Intel's Server Platforms Group in California, USA. He received his B.E. degree from the University of Madras, India, his M.S. degree from the University of Kentucky, Lexington, and his MBA from the University of California, Berkeley. Siva manages performance marketing activities for Intel® Xeon® server products. His e-mail is kuppuswamy.sivakumar at intel.com.

LEED - Leadership in Energy & Environmental Design (LEED®)

Power Management Enhancements in the 45nm Intel® Core™ Microarchitecture

Jose Allarey, Mobility Group, Intel Corporation
Varghese George, Mobility Group, Intel Corporation
Sanjeev Jahagirdar, Mobility Group, Intel Corporation

Index words: power management, CPU Sleep, C states, idle power, frequency

Citations for this paper: Jose Allarey, Varghese George, Sanjeev Jahagirdar "Original 45nm Intel® Core™ 2 Processor Performance" Intel Technology Journal. http://www,intel.com/technology/itj/2008/v12i3/2-paper/1-abstract.htm (October 2008).

ABSTRACT

Intel® processors based on the original 45nm Intel Core™ microarchitecture, originally referred to by the codename Penryn, improved the energy efficiency and performance per watt of the Intel Core microarchitecture. This paper discusses the new technologies introduced in the Penryn family of processors that enabled lower idle power and higher performance levels.

The Penryn family of processors builds on the power-management capabilities of the Intel processors based on 65nm Intel Core microarchitecture, originally referred to by the codename Merom, and takes them to the next level of idle power reduction and multi-core Enhanced Dynamic Acceleration Technology performance. The Penryn family of processors took very aggressive goals for idle power reduction in mobile, desktop, and server platforms. All features have been demonstrated to be fully functional on silicon and meet or beat the expectations of power reduction and increased performance. In fact, most of these features have already been enabled on the products that are currently shipping. Furthermore, the Penryn family of processors introduced new processor sleep states and Dynamic Acceleration mode on Intel Core 2 Quad processors for the first time in Intel to enable the acceptance of quad core as a mainstream product.

INTRODUCTION

The Penryn family of processors, implemented in a 45nm high-k metal gate silicon process technology, is designed to fit a wide range of power envelopes and market segments. Energy efficiency (energy consumed for doing a unit of work) is significantly improved due to process power scaling and innovative architectural power-management features. Power scaling enables better performance and higher power efficiency in most workloads, and the power-management features primarily enable lower idle power that leads to an overall reduction of platform energy consumption. Lower idle power helps improve battery life in mobile platforms, allows platforms based on the Penryn family of processors to meet or exceed Energy Star and other regulatory requirements for idle power consumption in desktop PCs, and lowers electricity and cooling costs for servers. The details of the 45nm high-k metal gate process technology and its power benefits are discussed in the last issue of the Intel Technology Journal [1,2] and hence are not covered here in detail. In this paper we focus on the architectural innovations in the power-management domain of the Penryn family of processors.

The Penryn family of processors builds upon the power-management capabilities of the core microarchitecture. The Advanced Configuration and Power Interface (ACPI) specification [3] describes the processor sleep states and performance states in detail. When the processor is executing instructions, it is in the C0 active state. The C1, C2 states, etc. are successively lower-power processor sleep states in which no instructions are being executed. P states are processor performance states defined by the processor frequency. The Penryn family of processors supports the Core Microarchitecture C states, P states, and thermal monitoring functions. In addition, the Penryn family of processors introduces several new key features and extends some of the existing mobile platform capabilities to desktop systems. Two key features introduced in the Penryn family are as follows:

- Deep Power Down (DPD), a new idle power state.

- Enhanced Dynamic Acceleration Technology (EDAT), a feature to increase Single Threaded (ST) performance by using the power headroom of the idle core. A simpler version of this feature is also available in the later versions of the Intel® Core™ 2 Duo 65nm processor.
- Power Management features extended to other segments are as follows:
- The Deeper Sleep state is now available in desktop platforms and in quad-core products.
- A version of the Deep Sleep state (called CC3 or Core-C3) state is now available in server platforms.
- The Enhanced Dynamic Acceleration Technology (EDAT) and Deeper Sleep state technology were extended to the mobile Quad-Core Extreme Edition product.

CHALLENGES

The Penryn family of processors was developed as a common core for the mobile, desktop client, workstation, and DP MP server platforms. The processor Thermal Design Power (TDP) in these segments ranges from 25 watts to 130 watts. Each platform has some common and some independent goals in power and thermal management. The challenge to the core development team is to anticipate the various requirements and design the support for all these platforms into the common core. For example, generic improvements such as reducing peak power and idle power consumption are applicable to all platforms. On the other hand, reducing energy usage during run time is a primary concern for the mobile and server platforms more than the desktop platform. Additionally, the mobile platform has tighter constraints on the peak power dissipation due to its form factor. Server platforms demand higher peak performance and are also capable of supplying and dissipating the power to support the peak demand conditions. The desktop platform usage mode has a lot of idle-ON time and standby time. Hence this platform has a specific requirement to reduce the idle and standby energy usage so that the end user will have a more energy-efficient, always-on always-available experience.

IDLE POWER IMPROVEMENTS

DPD technology

In mobile systems, battery life is an extremely important consideration. This is driving the need for low ''average power'' consumption in mobile processors. Designing high-performance mobile processors that consume approximately 30–40 watts during normal operation but have extremely low-power consumption during idle is challenging. The leakage in the millions of transistors in a processor design adds up to several watts if not several tens of watts. Consuming several watts of power while idle degrades the battery life significantly.

In the days of the Intel 486™ processor there was only one type of idle state—the Autohalt state. Most of the clocks to the processor were stopped in this state; active power was cut down significantly, and this brought the total processor power consumption down because leakage was not an issue. Since then, however, leakage has gotten worse with every new process generation, and more aggressive power-management states (C states) have been added to processors to combat this issue while idle.

A brief outline of the various C states follows as an introduction to the DPD technique.

The processor running state is called the ''C0 state'' in Advanced Configuration Power Interface (ACPI) [3] nomenclature. The processor is not executing any instructions in C states other than C0. A higher-numbered C state generally consumes lower power at the expense of higher exit latencies than a lower-numbered one.

In the C3 state, the processor Phased Locked Loop (PLL) is shut down to turn off all the clocks in the chip. This, however, does not lower the leakage, since voltage is not changed. In the C4 state, the voltage applied to the processor in C3 is lowered to reduce leakage. Here, the voltage is lowered only to the point where state can be retained in both the core and the caches. Intermediate states such as C1E, which achieve lower leakage with Vcc reduction yet maintain the advantages of a cache coherent state with low exit latency, have been implemented in recent processors.

The mobile product of the Merom processor family has added a state called Enhanced Deeper Sleep state (referred to as the C5 state), in which Vcc is reduced even further, that is, below the cache retention voltages. In C5, the Vcc to the core is just high enough that the processor core retains its state. At these voltage levels, leakage per transistor is low; however, given that the processors have millions of such leaky devices, it still adds up to a significant power loss.

In DPD the critical state of the processor is saved in a dedicated SRAM on-chip that is powered by the I O power supply for the chip (VccP), and then the core voltage is reduced to a very low level via the Voltage Regulator Module (VRM). Figure 1 shows the processor-VRM connectivity. At this point, it is

equivalent to the Vcc core being powered off, i.e., not consuming any power. Upon a break event such as an interrupt, the processor signals the VRM to ramp the Vcc back up, relock its PLLs, and turn clocks on. It then does an internal RESET to clear the states, restore the state of the processor from the dedicated SRAM, and open up the L2 cache. It then continues program execution from where it left off in the execution stream. All of these steps are completed in 150–200 us, all in the processor hardware, thereby making it transparent to the operating system and the existing power-management software infrastructure.

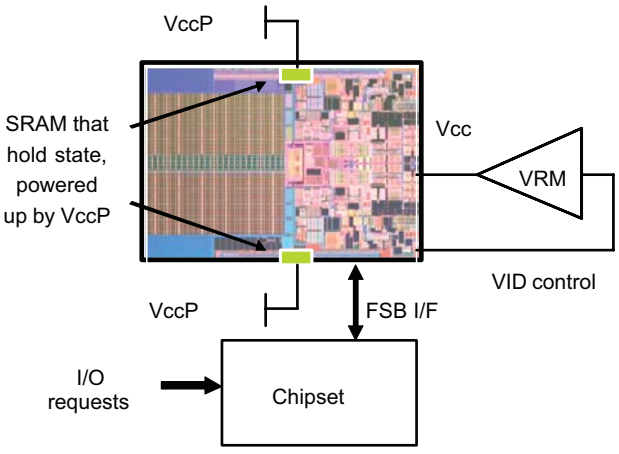

Figure 1: ***Voltage control for processor.***

Because the entry and exit are managed entirely by the processor and do not require software assistance, DPD can be enabled under existing operating systems and platforms. DPD achieves a breakthrough for idle power since it is agnostic to Vcc min and provides a minimal power consumption state with a short exit latency. Preliminary results on silicon show that the DPD feature reduces processor average power by 27 percent to 44 percent as measured by the Mobile Mark 05 battery life benchmark.

Quad-Core idle power improvements

Until a few years ago, quad-core processors were used only in server platforms. Now they are available as mainstream desktop processors. In the Penryn family of processors, the Intel Core 2 Extreme processor offers four cores for the mobile platforms. To provide users with a battery life of greater than 2.5 h for DVD playback, it was necessary to reduce the idle power of the quad-core processor. Previous-generation quad-core processors supported only the Autohalt state and hence were not optimal for the mobile market segment.

Quad-core processors are implemented in a multi-chip package (MCP) configuration. The two dual-core processor dice, referred to as the ''Master'' and ''Secondary'' sites, have independent PLLs, so the dice can run at different frequencies but share a common voltage plane. The Master die controls the voltage sent to the VRM and the voltage supplied to both dice. The secondary site coordinates its voltage requirements with the master die. The Penryn family of processors expands this coordination functionality for Deeper Sleep state support by using the same interface to communicate information regarding the sleep states of the cores. Each core pair on a die will resolve the correct idle state to enter. They will wait for the core pair on the other die to be ready to enter an idle state. At that point the Master die determines the resolved idle state and puts the package and the platform into that state.

Server idle power improvements (CC3)

The server versions of the Penryn family of processors are targeted for single-socket and multi-socket workstations and servers. In multi-socket platforms, a memory access from any core generates snoops to all other cores and sockets. Activity related to snoops of processor caches burns about 30 percent of active core power. If a core is in the idle state, it has to be woken up so that its caches can be queried for the data being requested by the snoop. After the snoop data has been returned to the requesting core, the woken core can return to the idle state. The wakeup and reentry negatively affects the idle residence and energy usage. By avoiding snoops into idle cores, this power can be saved.

In processors that predate the Penryn family of processors, the idle cores were put into Core C1 (CC1), which is a snoop-able state.

In the Penryn family of processors, idle cores can be put into Core C3 (CC3), which is a non-snoop-able state. The first-level caches are flushed into the L2 cache before putting cores into CC3. This prevents cross-core snoops and therefore the additional power burnt for snoops. Figure 2 shows how snoops are routed based on the cores' CC states. The additional latency to enter CC3 is insignificant (less than 1 us). This allows CC3 to be used as a replacement for CC1 with no affect on software and operating systems. In cases in which the additional latency cannot be tolerated, the CC3 state can be exposed by the ACPI interface as C2 [3]. This lets the Operating System Power Management (OSPM) layer pick either CC1 or CC3, based on the latency and policy settings.

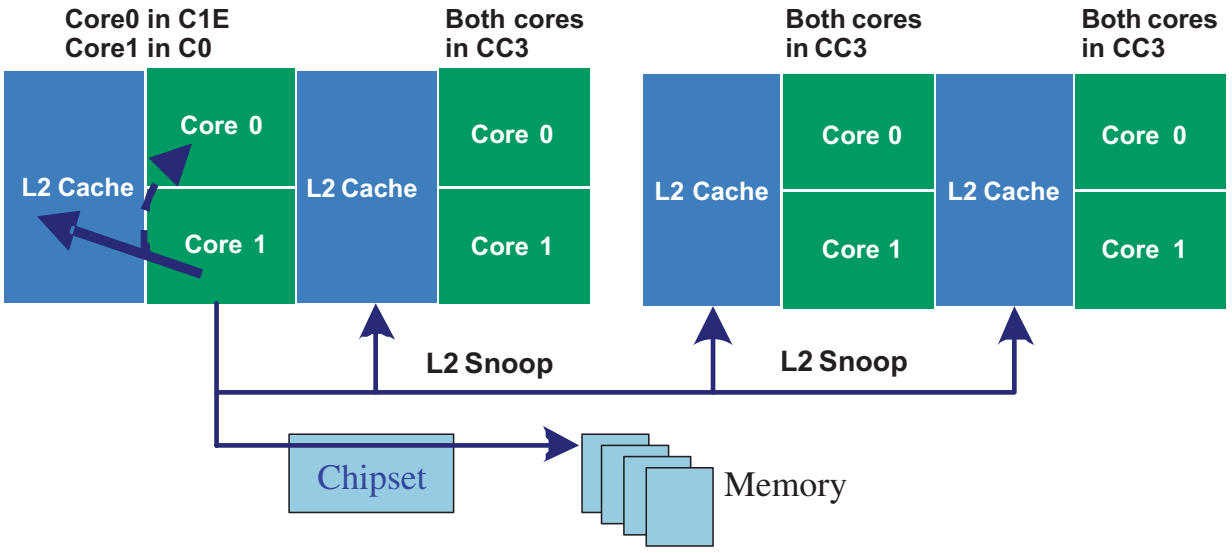

Figure 2: ***CC3 and cache snoops.***

The estimated savings of CC3 relative to CC1 are directly proportional to the number of snoops that occur during the sleep state. In an idle scenario, the snoops in the system are very low. In a fully-loaded scenario, the residence in CC1 CC3 is very low. Hence, the savings in idle and fully-loaded scenarios is less than when the system is lightly loaded. On average, the expected power savings from this feature is about 10 percent.

Even though we have been discussing this as an idle power improvement, it is more than that. In today's typical server platform there are two, four, or more cores. Chances are one or more of these cores is in Idle, except for periods when the servers are under peak load conditions. The idle cores will enter CC3 and reduce the system power.

The other important aspect for server power savings is the fact that there are hundreds or thousands of servers in a server farm. A 10 percent reduction in energy usage not only translates into reduced energy costs but also increases the performance watt cubic feet ratio of the server farm.

Desktop idle power improvements

Battery-life requirements spurred the innovations and enhancements to the sleep states in the mobile platform. Non-mobile platforms do not have a battery life requirement; instead, the focus there is to reduce overall energy usage. Energy Star requirements in the United States and similar requirements elsewhere in the world specify the maximum energy usage under various idle conditions for compliance. The energy usage in Idle is directly determined by the deepest idle power sleep state. When the processor is in a deep sleep state, the activity on the platform is very low. Hence, the total power and energy savings on the platform will be equal to the sum of the savings in the processor, the chipset, the VRM, et cetera [5].

The desktop processors that predate the Penryn family of processors used the Autohalt and Stop Grant states for the processor. The Penryn family of processors adds support for Deeper Sleep state in desktop platforms to lower the processor and platform idle power. The processor communicates to the platform (via the Graphics Memory Controller Hub (GMCH) and I O Controller Hub (ICH)) that it is in the Deeper Sleep state. This allows the GMCH and ICH to power off portions that are required to service the processor requests. This enables further reduction of the platform power when the processor is idle.

The Deeper Sleep state reduces processor power consumption by reducing the voltage of the processor to a lower level. At this level, the processor cannot execute instructions, but its state is retained. When in a sleep state, the power is determined by the amount of leakage in the processor. Leakage is a strong function of the voltage, and hence it is lower in the sleep states in which voltage is reduced. The idle power in the Deeper Sleep state was estimated to be significantly lower than the power in Autohalt or Stop Grant state due to the associated voltage reduction. The target power in the Deeper Sleep state is about 50 percent lower than the power in the Autohalt and Stop Grant states.

The Penryn family of processors' desktop platform also implemented one more platform power-saving

feature along with the Deeper Sleep state. The VRM losses can amount to more than 10 percent of the platform power when in idle state. Most VRMs are optimized to work at medium to highly-loaded conditions. This means that they are less efficient when they are lightly loaded as in the case of the processor being in Deeper Sleep state. The Penryn processor family communicates to the VRM when it enters the Deeper Sleep state, and it lets the VRM shut down all but one phase. This reduces the conversion losses in the other phases and also boosts the efficiency of the single active phase.

Desktop platforms use both the quad-core and dual-core processors of the Penryn family. Support for the Deeper Sleep state is available in both the dual- and quad-core configuration. The challenges of extending Deeper Sleep state support to the quad-core configuration are described in the previous section.

Enabling deeper sleep state on desktops

The Deeper Sleep state was exposed in the ACPI tables [3] as the C3 state. The Deeper Sleep state has a longer latency to memory traffic and interrupt response compared to the Autohalt and Stop Grant state as a result of its having to ramp up the voltage to the processor before responding to memory snoops and interrupts. There was a concern that this increased exit latency could have two undesirable consequences. Firstly, devices that were not designed to tolerate the latency to memory accesses could have buffer overruns and fail. Secondly, specific applications could suffer performance degradation due to the increased interrupt latency. To address the concerns, the exit latency was tuned down to the minimum value possible in the platform timers via a BIOS configuration. Various peripherals such as USB and Firewire (IEEE 1394) were tested for memory access latency increases. Benchmarks such as Sysmark were tested to understand the implications of increased interrupt latency. The results showed that the device functionality was unaffected, and the benchmark score differences were within the normal run-to-run variations. Performance benchmarks such as SPEC* do not have to be tested with Deeper Sleep because the processor is never idle during those benchmark runs.

The conclusion was that introducing Deeper Sleep state to the desktop platform did not have any noticeable adverse impact.

Power constrained performance in multi-core processors

EDAT takes advantage of the power headroom of the idle core to boost the performance of the active core while running an ST application. The EDAT frequency, which is pre-programmed in the chip, is chosen such that the total power still remains within the specified TDP as illustrated in Figure 3. In the Penryn mobile platforms, this optimization provides a 10 percent frequency boost for ST applications. The Penryn family of processors also implements a hysteresis mechanism that allows it to tolerate short wake-up intervals of the idle core without having to exit the EDAT frequency. This helps minimize performance loss in high-interrupt-rate workloads.

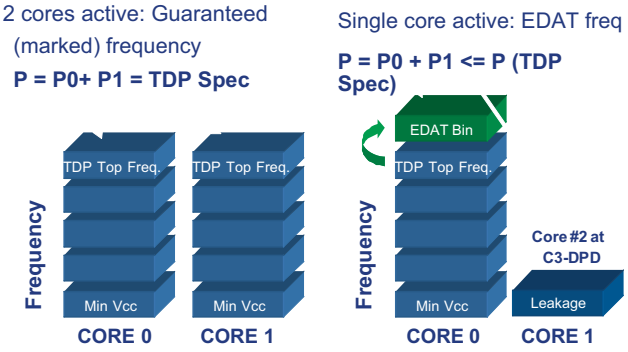

Figure 3: ***EDAT power budget reallocation.***

Multi-core architectures with two, four, or more cores are the standard for power-efficient computing. However, software is lagging in terms of being able to employ all available cores, either because applications have limited threading potential or because tools are not readily available to (re-)write or (re-)build code into threaded applications. As the number of cores within a package grows, many platforms, mobile in particular, become thermally constrained because they are designed to a power target primarily based on form factor. For multi-core designs, this power target usually assumes a workload that pushes the TDP envelope of all the cores. This implies that workloads that do not utilize all the cores underutilize the thermal capacity of the platform. This phenomenon will only grow more prevalent as the computer industry heads towards enabling smaller and cooler form factors.

The Penryn family of processors implemented EDAT in dual-core and dual EDAT in quad-core configurations to intelligently utilize the power headroom from idle cores and opportunistically boost the performance of applications that do not utilize all available cores, without exceeding the system thermal design constraints.

EDAT principles

EDAT operation is based on the following assumptions:

List Item (Number Type: 1,2,3,…)

1. The EDAT frequency (f_{EDAT}) for dual-core and quad-core processors is one frequency bin—typically 267 MHz or 333 MHz—over the maximum TDP-limited (guaranteed) frequency.
2. In dual-core processors, running one core at EDAT frequencies while the other core is idle will not exceed the TDP limit of the processor. Similarly, in quad-core processors, running one or two cores at EDAT frequencies while the other cores are idle will not exceed the TDP limit of the processor.
3. The EDAT frequency is requested by software via the legacy SpeedStep interface that sets the processor's frequency and voltage (F V) operating point.
4. The processor will transition to f_{EDAT} only if the appropriate number of processors is idle and if the operating system requests the highest performance state (P-state).
5. If there are not enough idle cores available to meet the EDAT criteria when the operating system requests the highest P-state, then the processor will run at the guaranteed operating point.

Assumptions 1 and 2 allow worst-case power assumptions to be applied to running cores during EDAT operation without having to worry about violating TDP limits.

Assumptions 3 and 4 ensure that EDAT will not be activated independently by hardware, without an operating system request, to enter a high-performance state; therefore these assumptions prevent the processor from consuming high power while lightly loaded. These assumptions also ensure that the processor will deliver at least the guaranteed performance level regardless of whether EDAT can be activated.

(Figure 4) depicts how the SpeedStep mechanism normally takes operating system P-state requests (i.e., F V operating point) and compares them against a fixed guaranteed F V operating point before determining the processor's "resolved" F V operating point. If the requested operating point is above the guaranteed operating point, the request is clipped to the guaranteed operating point. It also shows how EDAT dynamically changes the F V limit between the guaranteed and EDAT limit. The additional logic lets the processor run at the higher EDAT frequency based on how many cores are active, and a hysteresis mechanism ensures the hardware does not switch too often between the guaranteed and EDAT operating points.

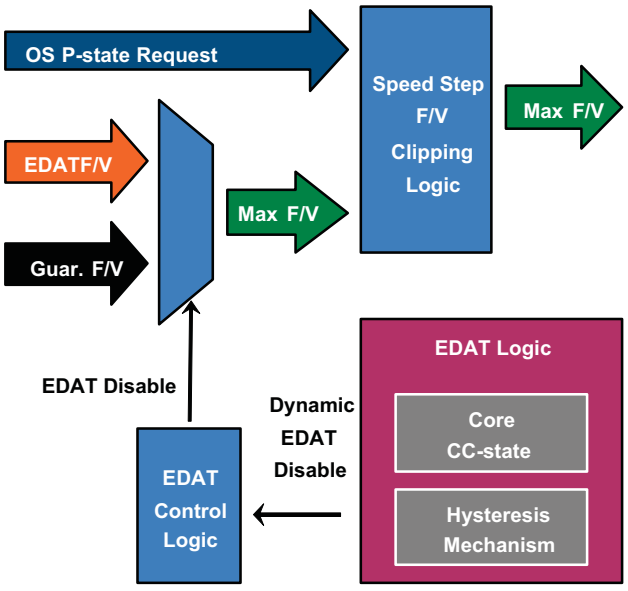

Figure 4: ***EDAT control mechanism.***

Idle cores and wake-up rates

The number of idle cores is one of the inputs to the EDAT decision logic. For the Penryn family of processors, cores are considered idle when their CC state is CC3, CC4, or CC6. In these states, cores consume only leakage power, and their clock distribution networks are shut off, which frees up TDP headroom for any active cores to use to run at f_{EDAT}. Furthermore, dedicated caches/buffers are flushed in these states, which allows any active cores to operate without having to wake up the idle core for cache snooping.

The hysteresis mechanism is the second input into the EDAT decision logic. It permits the processor to run at f_{EDAT} even if the number of idle cores does not meet the minimum requirement for a limited time, and it ensures that the processor stays within power and thermal constraints. This addresses a performance glass jaw that could be encountered if EDAT is simply disabled as soon as idle cores wake up.

Without the hysteresis mechanism, it is possible for idle cores to be frequently woken up by break events but remain active just long enough to service them before going idle again. Service times can be as short as 10 us to 20 us for Timer Interrupt Service Routines. This can cause rapid transitions in and out of the EDAT operating point and result in performance degradation due to SpeedStep transition overhead and from running the processor at the guaranteed frequency while servicing the break event. A common example of this is when multimedia applications set the timer interrupt rate to 1000 interrupts per second.

Having two cores active for 10 us to 20 us at a time is not likely to cause any thermally significant change on the platform since thermal time constants are in milliseconds. This means that EDAT does not have to be deactivated every time idle cores wake up. The hysteresis mechanism detects these situations and avoids unnecessary SpeedStep transitions.

An example of how the hysteresis mechanism works in a dual-core scenario is illustrated in Figure 5.

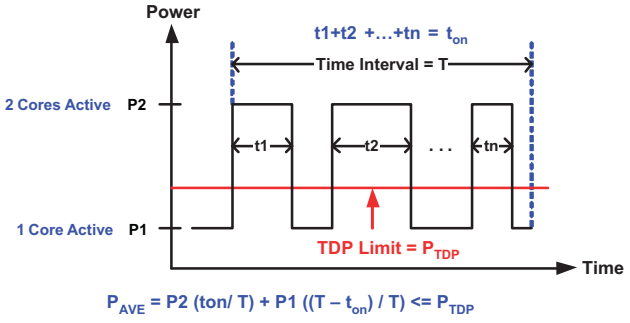

Figure 5: ***Hysteresis mechanism and average TDP.***

- P1 is the worst-case power usage when one core is active at f_{EDAT}, and P2 is the worst-case power usage when two cores are active at f_{EDAT}. PTDP can be arbitrarily chosen to be some small percentage, less than 5 percent, higher than the TDP of a top-bin, non-EDAT processor. In order to allow two cores to run at f_{EDAT}, platform voltage regulators need to be able to supply the current for two cores active at that frequency for short durations. The cost of supplying the larger current for short durations is usually small and is expected to add an insignificant cost to the platform.
- The time interval, T, is in milliseconds and is based on platform thermal time constants. Using this definition of T, along with the power data above, it is possible to compute the time that two processors can be at f_{EDAT}, ton, such that the average power, PAVE, over the interval, T, is less than or equal to PTDP.
- If two processors are active at f_{EDAT} for more than ton in any interval, T, the hysteresis mechanism "expires" and EDAT is disabled. The hysteresis mechanism also gates re-entering f_{EDAT} to avoid pathological cases in which a core immediately goes into CC3 right after exiting f_{EDAT}, the processor re-enters f_{EDAT}, and the idle core comes out of CC3. In this example, the total time that two cores are active at f_{EDAT} could exceed ton across the two f_{EDAT} periods.

EDAT extension into quad-core architecture

EDAT support in Penryn quad-core processors—dubbed Dual EDAT—was achieved in the same way that we enabled Deeper Sleep State in quad-core processors. The quad-core power-management coordination functionality was extended for Dual EDAT to

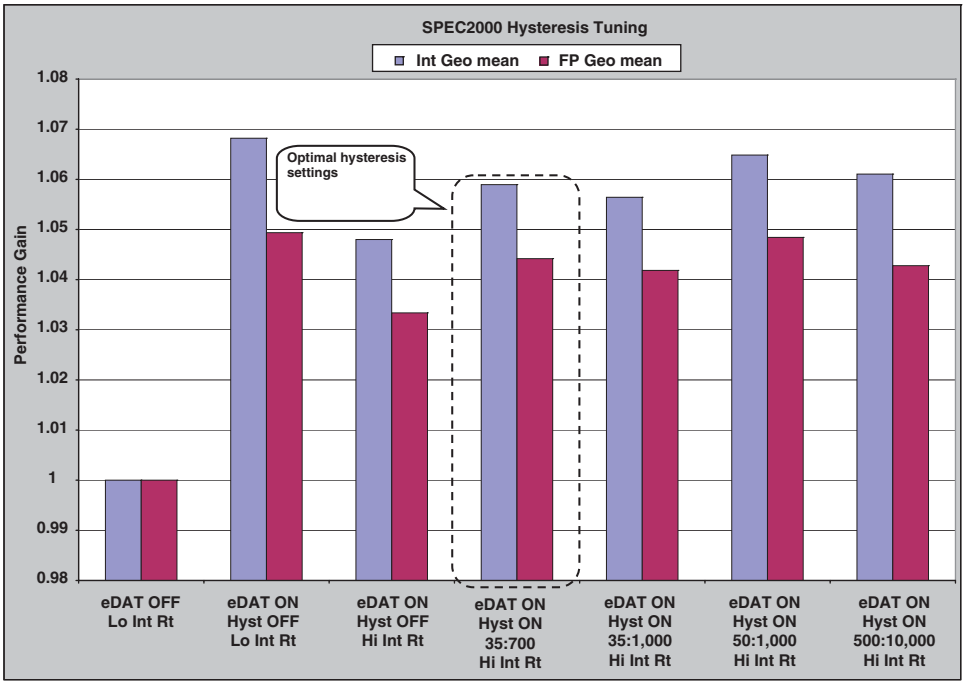

Figure 6: ***EDAT hysteresis mechanism tuning (Estimated SPEC* CPU2000 as measured on pre-production systems).***

share core idleness information and if EDAT was enabled by BIOS and or software. Each site locally resolved whether or not to allow the processor to run at EDAT frequencies, and it was expected that, in steady-state, they would both either allow or disallow the processor to run at f_{EDAT} frequencies.

RESULTS

DPD

Average power on the Penryn mobile processor was benchmarked using the MobileMark® 2005 (MM05) battery life benchmark. The benefit of DPD is shown in Figure 7. These measurements were done on multiple Penryn processor parts with varying leakage power behavior. DPD reduced the average power by approximately 40 percent or more on most of the parts tested. This testing was done on Intel Customer Reference Board platforms using a fresh build of Windows XP* and the MM05 benchmark according to the guidelines specified by BAPCO* [6]. The processor power was measured by sensing the processor voltage and current during the MM05 run and averaging it.

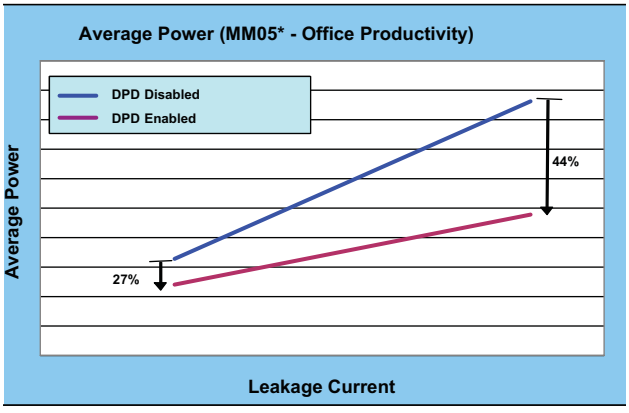

Figure 7: ***DPD power savings.***

Deeper sleep in mobile quad core processors

The same measurements as for DPD (listed above) were done for mobile quad-core processors. The average power on MM05 under Windows XP* was within 15–30 percent of the dual-core processors.

CC3 in servers

CC3 power-saving measurements were done on the Intel Customer Reference Board platform with the Seaburg chipset and ESB2. The dual-processor CPUs based on Intel Xeon® technology, from the Penryn family of processors, at 3.2 GHz with a 1600 MHz Front Side Bus (FSB) had two processors installed in the system for the tests. The server workloads were applied by SPECpower_ssj2008. The power saving was compared to the case in which only C1E was enabled and CC3 was not enabled. These configuration changes were made using the Intel BIOS options to enable and disable CC3. The savings range from 0 percent at complete idle to a savings of 10–20 percent at medium loads. The savings decrease at maximum loaded conditions as the idle time in C1E and CC3 decreases. This is shown in Figure 8.

Deeper sleep in desktop platforms

The measurements for idle power improvements were measured on the Intel Desktop Customer Reference board with dual-core and quad-core desktop processors from the Penryn family of processors. These measurements were done under operating system idle conditions. The idle power decrease from Autohalt Stop Grant to Deeper Sleep state was in the range of 40 percent–60 percent.

EDAT

Performance measurements on the latest Penryn family of processors mobile quad-core processors showed that Dual EDAT gave the expected performance gains on SPECInt2000, SPECFP2000, and their two-threaded SPEC rate counterparts. (Refer to "45nm Intel Core 2 Processor Silicon Performance" in this issue of the Intel Technology Journal, where Figures 11 and 12 show that Dual EDAT yielded a healthy 5 percent to 8 percent performance increase on these workloads for an 11 percent frequency boost over the guaranteed frequency [4]). Note that Dual EDAT has an idle core limit of two or more; hence, running SPEC rate with three or more copies does not result in any performance improvement.

Like most of the power and performance features, we tuned the EDAT hysteresis mechanism post-silicon to give the best performance on processor-intensive benchmarks such as SPEC2000. Figure 6 indicates that it recovered over half of lost performance gains due to high interrupt rates.

For EDAT hysteresis measurements, we used the Mobile Customer Reference board configuration with a Penryn 2.40 (HFM) 2.60(EDAT) GHz, an 800 MHz Bus, and 2×2 GB FB DIMM at 667 MHz. The operating system was Windows XP SP2*. The Workload was SPEC2000 measured by enabling and disabling the EDAT hysteresis mechanism with a timer interrupt rate at 1,000 interrupts second.

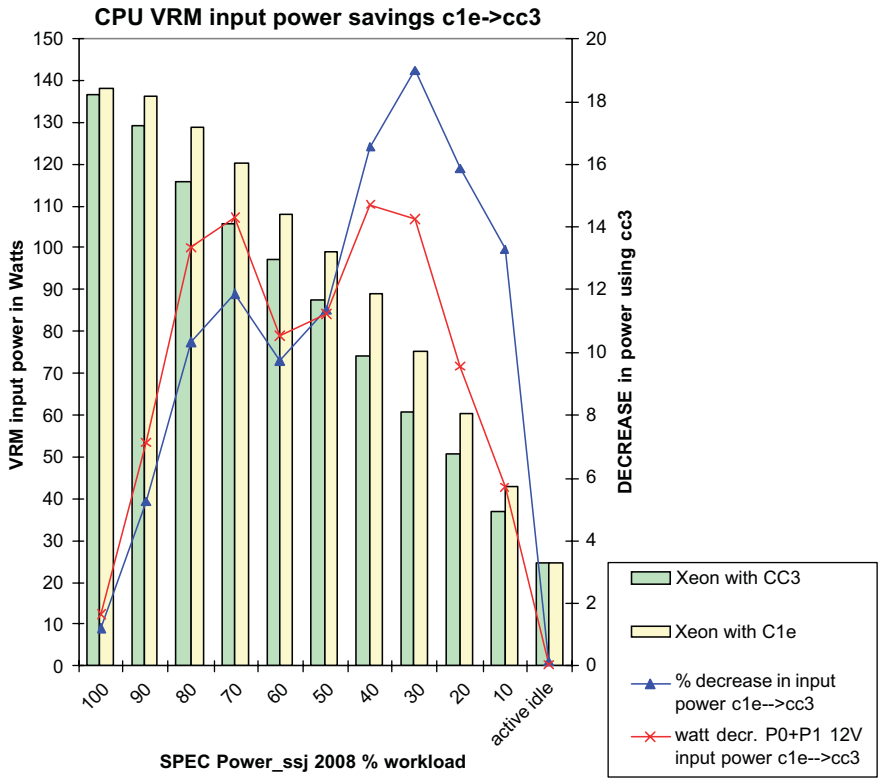

Figure 8: CC3 Power Savings vs. CPU Load

CONCLUSION

All the idle power improvement features introduced in the Penryn family of processors, namely DPD, CC3 in servers, Deeper Sleep in the desktop platform, and quad-core processors, met or exceeded the pre-silicon estimations. To the end user this means an increase in battery life on the mobile platforms, or a decrease in electricity usage in the desktop and server platforms. All these power improvements were achieved without any loss in functionality and negligible performance impact on real-world applications and benchmarks.

EDAT and Dual EDAT provide a means of efficiently utilizing the power envelope of multi-core platforms such that performance is not always constrained by the worst-case TDP scenario. This technology will be employed by this and the next-generation Intel processors.

ACKNOWLEDGEMENTS

Eric Heit and Suresh Doraiswamy were key contributors to the definition and implementation of these power-management features on the Penryn family of processors. We consulted with Steve Fischer and Rob Milstrey who provided guidance throughout the definition, implementation, and productization of these features. Alon Naveh and Efraim Rotem provided help and guidance on the 65nm Intel Core microarchitecture power and thermal management. Asim Nisar provided performance characterization for the EDAT and Dual-EDAT features. Ivan Herrera, Evgeny Vaisman, and Skip Lindsay were the key validators of the power-management features. Paul Zagacki played a key role in the definition and enabling for Deep Sleep on the desktop platform. Prabhu Rajamani, John Trelford, and Vamsi Jakkampudi helped with silicon data collection and characterization. Susumu Arai, Barnes Cooper, and Anil Aggarwal provided consultation on BIOS, operating systems, and OSPM aspects.

REFERENCES

[1] http://www.intel.com/technology/itj/.

[2] Intel Technology Journal, Vol. 12, No. 3 2008 at http://www.intel.com/technology/itj/2008/v12i2/.htm.

[3] ACPI Specification at www.acpi.info.

[4] N. *et al.* 45nm Intel Core 2 Processor Silicon Performance. Intel Technology Journal 2008 Vol. 12, *No.* 3.

[5] Paul Z. *et al.* Desktop Platform Power Improvements. Intel Technology Journal, Vol. 12, No. 3, 2008.

[6] BAPCO MobileMark 2005 at http://www.bapco.com/products/mobilemark2005/.

AUTHORS' BIOGRAPHIES

Jose Allarey joined Intel in 1998 and is currently working as a Computer Architect in the Mobility Group at the Folsom Design Center. His areas of specialization are Power and Thermal Management. He has six patents pending in this area. He received a B.S. degree in Electrical Engineering from the University of the Philippines in 1992 and an M.S. degree in Computer Architecture from Purdue University in 1998. His email is jose.p.allarey at intel.com.

Varghese George is a Principal Engineer in Intel's Mobility Group. He currently leads the MG-US Architecture team. For the Penryn family of processors, he led the Power Thermal Management Architecture team. Varghese joined Intel in 1993 and has worked on various processor products in areas of microarchitecture, multi-processor performance analysis, and power management. His focus in the last few projects has been on power-management techniques on-chip, and he has been instrumental in architecting many key power-management features into Intel processors. He holds more than 15 patents in various domains and has more pending. Varghese holds a B.S. degree in Electrical Engineering from the University of Mysore in India and an M.S. degree in Computer Engineering from the University of Maryland, College Park. His email is varghese.george at intel.com.

Sanjeev Jahagirdar joined Intel in 1996 and is currently working as a Computer Architect in the Mobility Group at the Folsom Design Center. His areas of specialization are power and thermal management. He holds seven patents in this area and has more pending. He received a B.S. degree in Electrical Engineering from the College of Engineering, Poona, India; in 1992 and an M.S. degree in Computer Architecture from Arizona State University in 1996. His email is sanjeev.jahagirdar at intel.com.

Glossary

C1	Autohalt state
C1E	Enhanced Autohalt state
C3	Deep Sleep state
C3E	Enhanced Deep Sleep state
C4	Deeper Sleep state
C4E	Enhanced Deeper Sleep state
C5	Enhanced Deeper Sleep state
C6	Deep Power Down state
CC	Core level C state. For example, CC3 is a core-level Deep Sleep state

Improvements in the Intel® Core™ 2 Penryn Processor Family Architecture and Microarchitecture

James Coke, Mobile Microprocessor Group, Intel Corporation
Harikrishna Baliga, Mobile Microprocessor Group, Intel Corporation
Niranjan Cooray, Mobile Microprocessor Group, Intel Corporation
Edward Gamsaragan, Mobile Microprocessor Group, Intel Corporation
Peter Smith, Mobile Microprocessor Group, Intel Corporation
Ki Yoon, Mobile Microprocessor Group, Intel Corporation
James Abel, Software Solutions Group, Intel Corporation
Antonio Valles, Software Solutions Group, Intel Corporation

Index words: SSE4.1, super-shuffle, radix-16, MOVNTDQA, streaming reads, CLI, STI, return stack buffer, super shuffle, SMC detection, Inclusion filter

Citation for this paper: Harikrishna Baliga, Niranjan Cooray, Edward Gamsaragan, Peter Smith, Ki Yoon, James Abel, Antonio Valles "Original 45mn Intel® Core™2 Processor Performance" Intel Technology Journal. http://www.intel.com/technology/itj/2008/v12i3/3-paper/1-abstract.htm (October 2008).

ABSTRACT

Intel® Corporation continuously seeks to improve the performance of each Intel Architecture microprocessor generation through architectural initiatives as well as process and circuit improvements. The predecessor to the Penryn family of processors, the 65nm Intel Core microarchitecture, codename Merom, led the competition in performance. This paper illustrates architecture techniques used by Intel in the family of processors to maintain this leadership position.

The new SSE ISA improvements (dubbed SSE4.1) are discussed, and we look at how the Penryn family of processors was able to utilize the Merom SSE enhancements to both enable SSE4.1 and improve legacy instructions. The instruction set is also examined to determine how instructions were targeted to improve various super-scalar workloads.

The paper explains how in the Penryn family of processors, the divide instructions are updated from Radix-4 to Radix-16. To minimize the hardware investment, integer divides are handled as floating point divides, so conversion techniques between integer and floating point are also discussed.

There were many other changes to improve the performance of the family of processors including improved data forwarding from stores to loads, removal of serialization from Set Interrupt Flag Clear Interrupt Flag (STI CLI), enhanced Self-modifying Code (SMC) detection, and "renaming" of the Return Stack Buffer.

INTRODUCTION

The family of processors is the latest production version of the Intel® Core™ 2 and Intel Xeon® product lines implemented on Intel's 45nm Hi-k silicon process. The Penryn microarchitecture is based on the 65nm Intel® Core™ 2 microarchitecture (codename Merom). A significant component of the Penryn value proposition was the addition of architectural and microarchitectural performance enhancements above the expected conversion to 45nm, so there was a strong desire to improve the performance of the architecture over that of its predecessor, the 65nm Intel Core™ 2 microarchitecture. There are many techniques to improve performance on a microprocessor, and each brings its own value to the final result. In this paper, we examine various architectural and microarchitectural changes that were used to attain the goal of improved performance. The majority of the performance improvements achieve a performance benefit on existing binaries, while the SSE4.1 changes require software changes to enable the added performance. A detailed discussion of the tradeoffs leading to these changes and the performance evaluation are documented in "45nm

Core 2 Silicon Performance Enhancements'' [1]. In this paper we focus on actual changes that led to a successful result.

SUPER SHUFFLE

One shuffler is better than two

Merom architecture dramatically improved SSE performance through a simple but highly effective method—doubling the width of SSE execution from 64 bits to 128 bits by instantiating two 64-bit execution units side by side and adding two small shuffle units to communicate between the 64-bit halves that handled only quad words. While increasing the execution width to 128 bits dramatically improved performance, the 64-bit ''wall'' between the execution halves was left in place.

The 64-bit wall caused some legacy instructions to be slower than would be expected. For example, the legacy instruction SHUFPS followed these steps in the Merom architecture:

1. Gathered all input DEST bits to [63:0] side of the ''wall.''
2. Gathered all SRC bits to the [127:64] side of the ''wall.''
3. Shuffled according to the immediate instruction.

SHUFPS output bits [127:64] always come from the SRC, and output bits [63:0] always come from the DEST, so we had to move SRC bits [63:0] to the upper half of the SSE execution unit. SHUFPS output bits [63:0] always come from the DEST input, so we had to move DEST bits [127:64] to the lower half of the SSE unit.

Some of the performance penalty of having three operations is covered by the Merom architecture having shuffle units on two ports as well as another SSE shuffler that handled only 64-bit data sizes on a third port.

The family of processor's solution to this problem is to merge the two shuffle units into a single Super-Shuffle unit that does not have a 64-bit wall. The Super-Shuffle is significantly more costly in terms of routing, but the added area cost is covered by merging the area from the two old shuffle units into the Super-Shuffle.

In the Penryn family of processors, the SHUFPS shuffling algorithm becomes:

1. SHUFPS!

While the Merom algorithm had a nominal throughput of 1, it used three operations. The implementation also has a throughput of 1, but it uses only one operation, leaving more execution bandwidth for other instructions. In the architecture we reduced the SHUFPS latency from 2 to 1.

We made similar improvements to SSE instructions PACKxSxx, PUNPCKxxx, PHADD*, PSHUFB, PALIGNR, PINSRW, PEXTRW, UNPCKLPS, UNPCKHPS, HPADDPS, HSUBPS, and PSHUFD.

INTRODUCTION TO SSE4.1

Many of the SSE4.1 instructions were created by noticing patterns in kernels that were commonly repeated using multiple instructions and that could be readily converted to a single instruction in hardware. Some of the instructions fill in gaps in the existing instruction set such as the new PMINxx, PMAXxx, PEXTRx, PINSRx, PACKUSDW, PCMPEQQ, and PMULLD. PMULDQ is the signed version of PMULUDQ. Please refer to the Software Developer's Manual for details [4].

The Super Shuffle breaking the 64-bit wall is a key enabler of many SSE4.1 instructions' performance improvement. The PMOVSX, PMOVZX, PEXTRx, PINSRx, and INSERTPS all require the new Super Shuffle to realize their full potential. Figure 1 shows PMOVZXDQ moving data across the 64-bit wall without a special operation to move data from the low 64 bits to the high 64 bits. The MPSADBW and PTEST instructions also depend on crossing the 64-bit boundary albeit in other execution blocks.

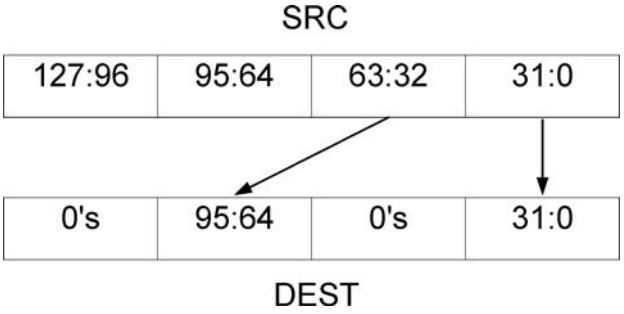

Figure 1: ***PMOVZXDQ crosses 64-bit boundary on the family of processors without a special 64-bit operation.***

SSE4.1 instructions DPPS, DPPD, and INSERTPS solve the problem of requiring additional instructions to selectively zero portions of the register. This zeroing effectively compresses two instructions into one for INSERTPS and four instructions into one for DPPS and DPPD.

As shown in Figure 2, DPPD and DPPS are the first floating-point SSE instructions to have multiple floating-point operations.

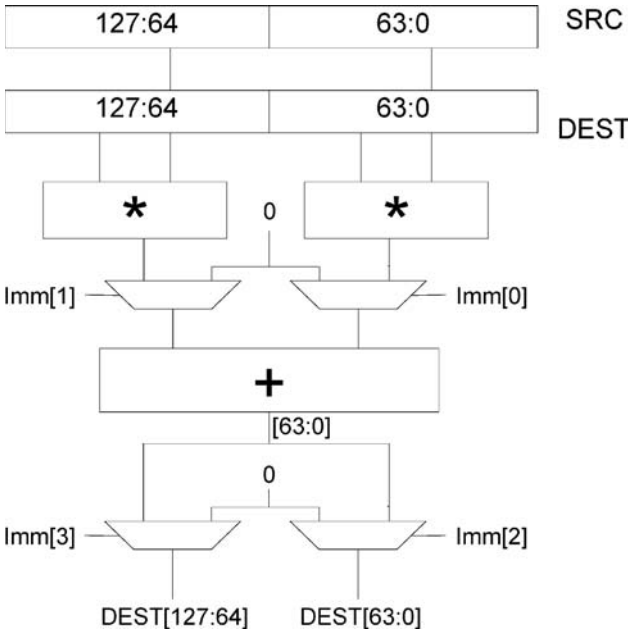

Figure 2: ***DPPD provides zeroing after both floating-point operations.***

Two new rounding instructions, ROUNDPS and ROUNDPD, provide rounding of floating-point values to integers and return the values in floating-point format. The rounding control is selectable from either the immediate instruction or MXCSR.RC. The user also has control over suppressing Precision Exceptions based on an immediate bit.

Prior to SSE4.1, SSE instructions always operated on subsets of 128 bits. PTEST is the first SSE operation to operate on the entire 128 bits as a single entity. PTEST is very useful for detecting all 0s and all 1s and reporting the result in the flags to simplify decisions.

MPSADBW performs a series of eight 4 X 4 SAD (Sum of Absolute Differences) operations across an 8-byte window of the destination. The starting point for the SRC and DEST windows is selectable using the immediate instruction.

PHMINPOSUW forms a very useful counterpart to MPSADBW because it finds the minimum word and returns both the value and position of the minimum word.

INSERTPS is a very generalized insertion between XMM registers. It allows any packed single quantity to be selected from the source and inserted into any position in the destination. The control to select the packed single position from the source and the position to insert the packed single in the destination are controlled by the instruction immediate. INSERTPS also allows selected packed single positions in the destination to be zeroed, also under control of the immediate.

EXTRACTPS sounds as if it should be the complement of INSERTPS, but it isn't. EXTRACTPS extracts to a General Purpose (GP) Register instead of to another XMM register, which makes EXTRACTPS very similar to PEXTRD. The only difference between EXTRACTPS and PEXTRD is the handling of REX.W. EXTRACTPS will zero extend the 32-bit value, whereas PEXTRD will become the 64-bit instruction PEXTRQ.

The BLENDxx and BLENDVxx instructions are a per-element select of the source or the destination register. Control of the select comes from the immediate for BLEND instructions, and for BLENDV instructions the control comes from the element sign bits of a third XMM register that must be XMM0.

STREAMING READS

Previous generations of Intel architecture processors supported a fast write mechanism from the processor to memory (such as to video and graphics memory) via streaming non-temporal writes. This greatly improved the write bandwidth from the processor to memory. However, up to now, Intel architecture was lacking a fast memory read mechanism for memory regions that are typically mapped as uncacheable with weak ordering—typically graphics video memory. The fast cacheable memory read mechanism cannot be utilized in this case because we do not want these data to be cached in the processor caches. In addition, we also do not want this type of data to expel useful data from the processor caches. The SSE4 instructions in the processor introduced a new streaming read IA instruction, MOVNTDQA, to fill this void. This new instruction, which is introduced on the second production stepping of the architecture, performs very high-bandwidth reads from weakly ordered, uncacheable (USWC) memory regions, typically used for graphics memory, without any pollution of the processor caches. This gives the programmers the ability to utilize the fast execution units inside the processor to operate on graphics-type data, which until now was not desirable due to very slow read bandwidth by the processor.

By allowing fast non-coherent transfers across PCIe or access to UMA graphics directly, streaming reads help increase the performance of analog and uncompressed high-definition video capture (20–30 percent of these workloads involve readback). It also makes hardware accelerated transcode (encode followed by decode) and video motion estimation feasible, with the fast

readback mechanism after HW accelerated decode in the northbridge.

The semantics of the MOVNTDQA instruction is to load an aligned 16-byte quantity. It is a demand load operation with a streaming hint. When this instruction is used to load 16 bytes from a memory region that is mapped as USWC, the processor automatically converts the load operation to a "streaming" load operation. By treating the load as a streaming load operation, the processor automatically converts the 16-byte load to a full cache line (64-byte) load operation and uses the maximum Front Side Bus (FSB) bandwidth to transfer the data from memory. For a 333-MHz FSB (1.333-GHz data transfer rate) we could achieve a 10.6-GB/s data transfer rate from USWC memory using MOVNTDQA loads, which is the same maximum bandwidth achievable by cacheable loads. This is compared to the maximum data transfer rate of 1.3 GB s for loading from USWC memory using non-streaming load instructions, assuming the FSB is 100 percent utilized.

When the processor treats a load operation as a streaming type (via MOVNTDQA), the entire USWC cache line (aligned 64 B) that contains the address of the load is loaded into an internal processor buffer, and the requested 16 bytes of data are served. The use of a temporary buffer for streaming along with a read-once policy helps maintain the uncacheable semantics of the USWC memory type. As shown in Figure 3, this internal buffer is not drained at the completion of the requested 16-byte load but is kept alive so that subsequent NT loads (MOVNTDQA) can be serviced from the same buffer rather than initiating new memory transactions. Thus, a program issuing four MOVNTDQA loads will be satisfied by a single buffer and a single memory transaction. A program that is designed to loop on four MOVNTDQA loads (such as operating on a block of memory, loading one cacheline at a time, and operating on it) can achieve data-read bandwidths up to the maximum FSB bandwidth. Once the entire contents of the temporary buffer are consumed (by four MOVNTDQA load operations), the buffer is automatically deallocated. Since the processor contains a limited number of internal temporary buffers, care must be taken while programming to not overflow or underutilize these resources.

Here is an example usage of MOVNTDQA instructions to efficiently utilize the streaming read buffers. Note that eax addresses are aligned to a line boundary.

```
MOVNTDQA xmm0, [eax]
MOVNTDQA xmm1, [eax+16]
MOVNTDQA xmm2, [eax+32]
MOVNTDQA xmm3, [eax+48]
PAVGB xmm0, xmm1
PAVGB xmm2, xmm3
PAVGB xmm0, xmm2                <Code 1>
```

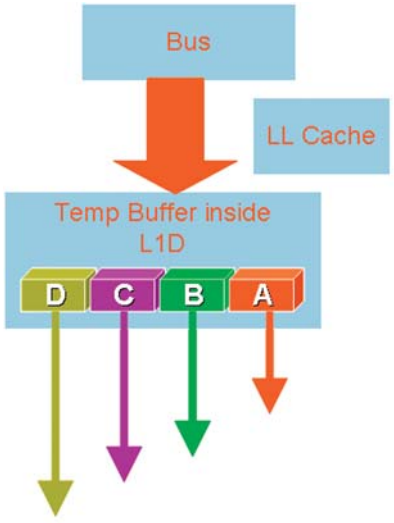

Figure 3: ***MOVNTDQA allows use of the full temp buffer before starting a new bus cycle.***

SSE4.1 EXAMPLES

SSE4 instructions were created to provide speedup on various types of applications. Two instructions in particular, MPSADBW and PHMINPOSUW, in combination with the Super Shuffler, can provide large performance improvements in block-matching algorithms commonly used in motion estimation. A detailed discussion on the block-matching performance (a 1.6x–3.8x function-level speedup) and how these instructions provide the performance improvements is documented in Penryn Silicon Performance [1].

Two other SSE4 examples are discussed in this section. First, we briefly showcase the measured performance of streaming loads [2] to conclude the discussion in the previous section. Then we discuss the DPPS DPPD instruction and usage models where DPPS DPPD will improve performance. We provide an example that uses the DPPS instruction to showcase how it and another SSE4.1 instruction (EXTRACTPS) can be used to speed up collision detection performance.

Streaming loads

In this section we continue the discussion from the previous section by briefly examining streaming loads measured results and optimization guidelines [2]. To maximize streaming load throughput, users need to utilize the streaming load buffers of two cores at the same time. That is, two software threads executing on

two different cores perform streaming loads from separate USWC parts at the same time and copy the data into separate WB cacheable memory buffers (see Figure 4). The WB buffers have to be small enough to fit in the first-level cache to minimize resource contentions, and the four streaming loads making up one cacheline (64 bytes) need to be done close together.

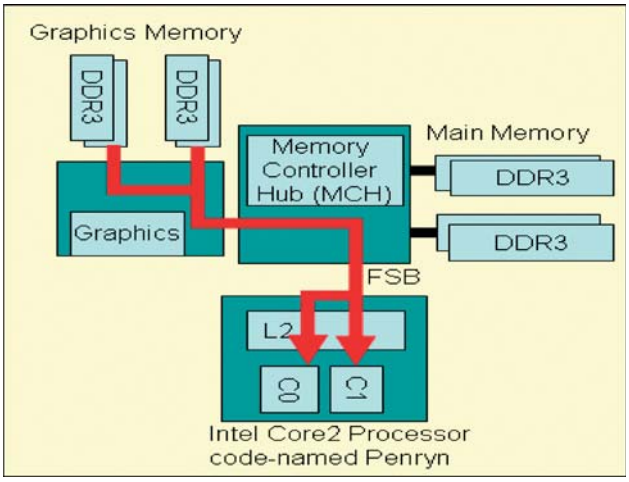

Figure 4: Example of streaming load: accessing graphic card memory and utilizing two threads to maximize memory throughput

Tests were conducted on a 45nm Intel Core 2 desktop processor (E7200) with a 1067-MT sec FSB.

The theoretical memory throughput (cacheable and uncacheable memory) can be calculated as follows:

Theoretical memory throughput

```
= FSB Transfer/sec * bytes/transfer
= 1067 MT/sec * 8 B/T
= 8.53 GB/sec
```

The single-threaded streaming load implementation that utilized one core's streaming load buffers was measured to provide approximately 50 percent of the theoretical memory throughput. The dual-threaded streaming load implementation that utilized the streaming load buffers of two cores was measured to provide approximately 90 percent of the theoretical memory throughput. Utilizing two core's streaming load buffers is the recommended way to get the highest memory throughput out of streaming loads.

Single-Precision Floating-Point Dot Product

The Dot Product of Packed Single Precision Floating-Point Values (DPPS) instruction and the DPPD instruction for Double Precision Floating-Point numbers can provide performance benefits in games, multimedia, and high-performance computing applications. This instruction has a high latency due to multiple numbers of operations being done at once. Thus, this instruction provides the most benefit in situations in which the Array of Structures (AOS) data layout is being used as opposed to the Structure of Arrays (SOA) data layout [3]. The AOS layout is usually not Single Instruction Multiple Data (SIMD)—friendly except for the horizontal instructions such as DPPS and HADDPS. Users can use these horizontal instructions to avoid the heavy data swizzling [3] costs in converting to the SOA data layout. An SSE3 implementation of a dot product of Vector Length 4 in the AOS format can be implemented by using the HADDPS instruction as shown:

```
void dot_product_vlength4_SSE3
(float *src,float *dst,int Count)
{
__asm {
      mov esi, dword ptr [src]
      mov edi, dword ptr [dst]
      mov ecx, Count
start:
      //a3, a2, a1, a0
      movaps xmm0, [esi]
      //a3*b3,a2*b2,a1*b1,a0*b0
      mulps xmm0, [esi + 16]
      //a3*b3+a2*b2,a1*b1+a0*b0,
      //a3*b3+a2*b2,a1*b1+a0*b0
      haddps xmm0, xmm0
      movaps xmm1, xmm0
      psrlq xmm0, 32
      addss xmm0, xmm1
      movss [edi],xmm0
      add esi, 32
      add edi, 4
      sub ecx, 1
      jnz start
   }
}          <Code 2>
```

Notice that the SSE3 implementation of the dot product requires the MULPS + HADDPS + MOVAPS + PSRLQ + ADDSS instructions. The SSE4 implementation replaces all of these instructions with one: DPPS (Code 3).

```
void dot_product_vlength4_SSE4
(float *src,float *dst,int Count)
{
__asm {
      mov esi, dword ptr [src]
      mov edi, dword ptr [dst]
      mov ecx, Count
start:
      movaps xmm0, [esi]
      dpps xmm0, [esi + 16]
      movss [edi],xmm0
      add esi, 32
      add edi, 4
      sub ecx, 1
      jnz start
  }
}            < Code 3>
```

Table 1 shows the measured performance of the two different dot product implementations in AOS data layout as compared to the C implementation. The DPPS instructions can provide performance speedups on multiple vector matrix operations that require a dot product such as vector normalization [3] and collision detection.

Table 1: Performance of dot product implementations.

Implementation	Cycles/Loop	Speedup over C
C	9.8	1.0 ×
SSE3	7.8	1.26 ×
SSE4	5.7	1.72 ×

COLLISION DETECTION

The dot product can be used for collision detection in games. This is another example of using the DPPS instruction in an AOS layout. In this example, the DPPS instruction is used to speed up collision detection of two spheres. To explain collision detection, consider two circles. The circles are said to have collided if the sum of radii is greater than or equal to the distance between the centers (Figure 5). This same equation applies to spheres, except in that case there is a z-axis that contributes to the distance formula.

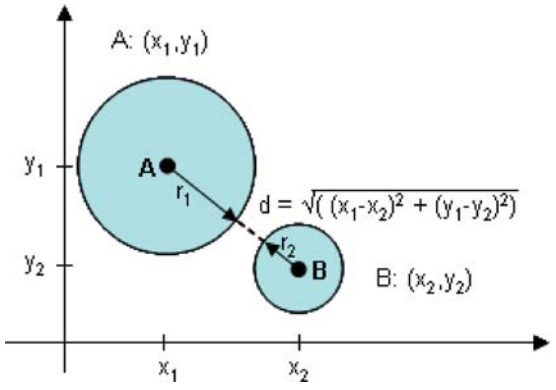

Figure 5: ***Circles/spheres collide if the sum of the radii is greater than or equal to the distance between the two centers.***

Two spheres collide if

The distance between two centers <= sum of radii

$$\text{sqrt}((x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2) <= r1 + r2$$

$$(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2 <= (r1 + r2)^2$$

Dot Product (point A, point B) <= $(r1 + r2)^2$

$$A * B <= (r1 + r2)^2 \quad (1)$$

As an example, imagine a fast-moving, hot fire particle about to incinerate other objects/particles. Collision detection can be used to find out which of these objects comes in contact with the fire particle and will need to be set on fire (see Figure 6).

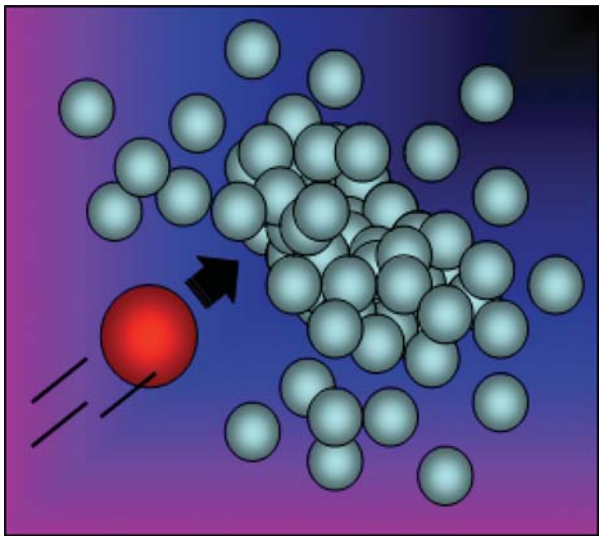

Figure 6: ***Collision detection example: one fast-moving fire particle is about to slam into and incinerate other objects/particles.***

The C code implementation (Code 4) uses Eq. 1 to detect sphere collision between the fire particle and a thousand other particles.

```
struct VEC3
{
float x,y,z;
float dot() const
{return x*x + y*y + z*z;}
};
VEC3 operator -
(const VEC3 a, const VEC3 b)
{
VEC3 res;
res.x = a.x - b.x;
res.y = a.y - b.y;
res.z = a.z - b.z;
return res;
}
struct SPHERE
{
VEC3 center;
float rad;
};
void sphere_collision_C
(SPHERE *sp1,SPHERE *sp,int *coll)
{
for(int i=0; i<gNumSpheres;i++)
{
float center_distance_squared =
(sp1[i].center-sp->center).dot();
float radii_sum_squared =
(sp1[i].rad + sp->rad)*
(sp1[i].rad + sp->rad);
if(center_distance_squared <
radii_sum_squared)
{
//which spheres collided
coll[i]++;
    }
  }
}
```

<Code 4>

The collision detection code can be optimized with SSE4.1 instructions. One single DPPS instruction can be used to make the three different calculations in Eq. 1: the dot product of AB: $(x1-x2)^2+(y1-y2)^2+(z1-z2)^2$, the radii sum: $(r1+r2)^2$, and the subtraction of the radii sum from the dot product.

$$res = A * B - (r1 + r2)^2$$

Collision occurred if res's sign-bit is set (Code 5).

```
int sphere_collision_intrinsics (SPHERE
*sp1,SPHERE *sp,int *coll)
{
int res;
__declspec(align(16))
static long _mask[4] =
{0,0,0,0x80000000};
__m128  s,s1,s2;
__m128 _mask128 =
*(__m128*)_mask;
s = _mm_load_ps((float *)sp);
//set sign bit to add: r1 - (-r2)
```

```
s = _mm_xor_ps(s,_mask128);
for(int i=0;i<gNumSpheres;i++)
{
s1 = _mm_load_ps((float *)sp1);
s1 = _mm_sub_ps(s1,s);
//set sign bit on radii sum
s2 = _mm_xor_ps(s1,_mask128);
//s1[0-31] =
//center_distance_sqaured -
//radii_sum_squared
s1 = _mm_dp_ps(s1,s2,0xff);
//get sign bit of subtraction
res = _mm_extract_ps(s1,0);
res >>= 31;
//coll if radii_sum_squared
// > center_distance_squared
res &= 1;
coll[i] += res;
}
}                          <Code 5>
```

To use a single DPPS instruction to do all of this, we had to use a few SIMD tricks. First, we laid out the data so x,y,z,r of the sphere could be loaded with one aligned load. Then we used a mask to modify the sign bit of one of the radii before the packed subtract. We did this so that the subtract operation actually causes an addition, (r_1-(-r_2)). Then we used the mask again to modify the sign bit of one of the radii sums before the DPPS instruction. This causes the radii sum squared to be subtracted from the dot product of A and B.

These are the contents of the __m128 variables before the DPPS instruction:

```
//sign bit set on upper 32-bytes of _m128
s2 = [-(r1+r2), z1-z2, y1-y2, x1-x2]
s1 = [r1+r2, z1-z2, y1-y2, x1-x2]
```

These are the contents of s1[0-31] after the DPPS instruction: If s1[0–31] is negative, then the radii sum squared is greater than the dot product of A and B, and a collision occurred.

Another SSE4.1 instruction, EXTRACTPS, is used to extract the single precision floating-point value from an XMM register to a GP register to check if the sign-bit is set. The EXTRACTPS instruction removes the branch and enables GP registers to be used to do data manipulations. Both the DPPS and EXTRACTPS instructions provided the 1.5x speedup over C as shown in Table 2.

Table 2: SSE4.1 Collision detection speedups.

Implementation	Cycles/Iteration	C Speedup
C	14.3	1.0 ×
SSE3	17.7	0.8 ×
SSE4	9.5	1.5 ×

The analogous SSE3 solution has to use MULPS + HADDPS + MOVAPS + PSRLQ + ADDSS instructions as shown in Code 2 for the dot product. It also has to move the result to a floating-point register to do the comparison similar to the C-code. The SSE3 implementation has too large of a latency to provide a speedup over C.

In this section we provided two SSE4.1 examples. We discussed the measured performance of the streaming loads and provided the recommended usage scenario. We also discussed the DPPS instruction and the recommended usage scenarios for this instruction as demonstrated in the collision detection example.

NEW RADIX-16 DIVIDER

The new Radix-16 floating-point divider with variable latency Radix-16 integer divide capability replaces the Merom Radix-4 floating point divide and Radix-2 square root and integer divide hardware. The preceding algorithm dated back to the Pentium® divide implementation.

Motivation and implementation

Divide hardware is costly both from die size and performance perspectives. Its large size makes it prohibitive to add multiple units on a single core. On the other hand, the long latency and low throughput of divides has a dramatic impact on CPU performance. The implementation provides a remedy for the latter by reducing the number of loop iterations for a single divide.

In the Sweeney, Robertson, and Tocher divide algorithm (SRT) [5–8], the divide operation is broken up into three parts: pre-processing, loop, and post-processing. The loop accounts for the predominant source of the latency and prevents subsequent micro

operations from utilizing the hardware in a pipelined manner. Specific implementations may choose to pipeline consecutive operations over the three parts (for example, a second operation's loop may be implemented to begin once the first enters post-processing). However, any de-pipelining pales in comparison to the loop latency's impact to divide throughput. The latency of different Radix implementations is shown in Table 3.

Table 3: Divide and Square Root latencies.

	Latency in cycles			
	Pre + post processing	R2 loop	R4 loop	R16 loop
Single precision	5 to 6	28	14	7
Double precision	5 to 6	57	29	15
Extended precision	5 to 6	68	34	17

The loop latency has a direct correlation to the number of quotient mantissa bits in any given precision. In Radix-2, one quotient bit is calculated in every cycle; thus, the number of cycles in the loop equals the number of mantissa bits. In Radix-4, two quotient bits are calculated every iteration. For Radix-16, four quotient bits are calculated. It can be seen why the enhanced divider had a profound impact on performance: divides were up to 1.75 times faster, and square roots were up to 3.3 times faster.

The new variable latency integer divide algorithm utilizes the underlying Radix-16 floating-point divider without the need to implement a different integer algorithm or build a separate integer divide unit. The same exact algorithm can be used on integer numbers after they undergo an integer normalization and shift amount recording, prior to the pre-processing performed by modified existing hardware.

In addition to improving performance by moving from Radix-2 integer divides on the Merom processor to Radix-16 on the family of processors, the integer divide operation can finish sooner than what is depicted in Table 3 depending on the specific data operands. Since the loop iteration count depends on the number of quotient bits produced, and given that integer operations produce an integer quotient and separate remainder, the integer divide algorithm stops the loop after the quotient is created and begins post-processing. However, due to other existing microarchitectural restrictions, the total divide micro operation latency is at least 11 cycles, excluding early out conditions (such as 0 div by n). Thus when there are 17 or more quotient bits produced but less than 29 bits (for r m32; less than 61 bits for r m64), then there is a further reduction in latency over the previous algorithm.

Challenges

Historically, implementing high-Radix fast dividers has been a design challenge. Finding the correct balance between implementing a high-Radix quotient and a fast-quotient selection logic (QSL) is a difficult task. In the family of processors, we addressed this by applying a new digit-redundant structure and an implicit bias bits concept to ordinary basic divide algorithms, such as Non-performed on a binary digit basis, without rippling the carry forward. Thus, each digit produces two outputs: the sum and the carry. After all of the redundant arithmetic is performed, completion adders are employed to roll in the carry bits in the final step.

As can be seen in Figure 7 the divider is essentially double pumped, producing two bits of quotient every phase to yield four bits per cycle. Contrast this with the previous Radix-4 design in Figure 8 in which two bits of quotient were produced per cycle. By using the new digit-redundant structures in conjunction with the implicit biasing for selecting the quotient, an efficient and fast way of selecting the quotient can be achieved with a small number of bits of the partial remainder and the divisor. This will allow for fast redundant implementations of the internal loop computation and the quotient selection logic. The simplified quotient selection logic that is based on only a few bits of the estimated value of the partial remainder will in turn allow a very fast implementation that enables a multiple of these QSL blocks to exist in the same cycle, allowing for very high Radix dividers. The paths in the main loop and QSL are equalized by overlapping them, and they were targeted to a delay of just MUX delay plus truncated adder/comparator delay on either path.

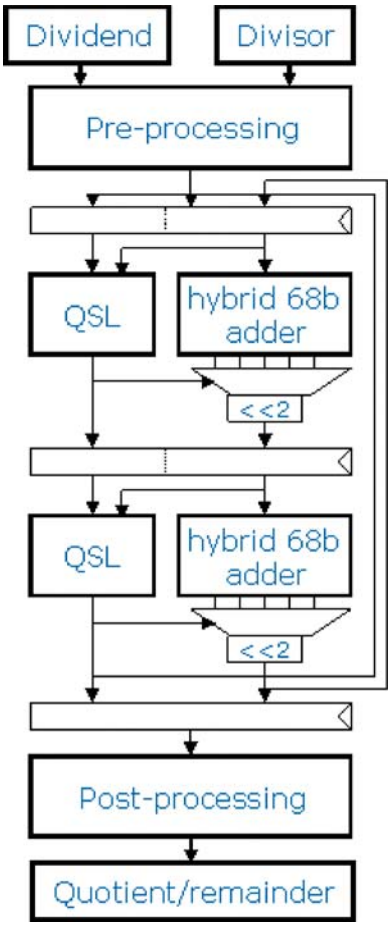

Figure 7: *New divider microarchitecture.*

CLI/STI PERFORMANCE TUNING

In the Software Developer's Manual, both the Clear Interrupt Flag (CLI) and the Set Interrupt Flag (STI) macro-instructions are said to be non-serializing. However, due to past microarchitectural simplifications, both instructions serialized the execution pipeline, leaving an unoptimized performance situation. The serialization was deemed necessary to ensure that an updated copy of the Interrupt Flag (IF) in System Flags was ready to be evaluated at the end of every macro-instruction. Furthermore, the IF masking that is done at the end of STI needs to occur only when the IF is transitioning from clear to set and needs an updated copy of IF at the beginning of the STI instruction.

It was determined that for multi-tasking operating systems, the IF can get frequently masked and unmasked during atomic operations to prevent other processes from obtaining the context in the middle of modification. As such, there can be a noticeable performance degradation due to the aforementioned CLI STI serialization. Pre-silicon performance simulations showed a 1.3 percent improvement on productivity workloads if this penalty was removed.

On the processor, instead of post-serializing on a CLI or STI, a serialization occurs only when the new IF value is consumed and only if the new value is not yet updated. Additionally, we added new dedicated hardware to the retirement logic to detect whether the IF transitions from clear to set during an STI, in order to

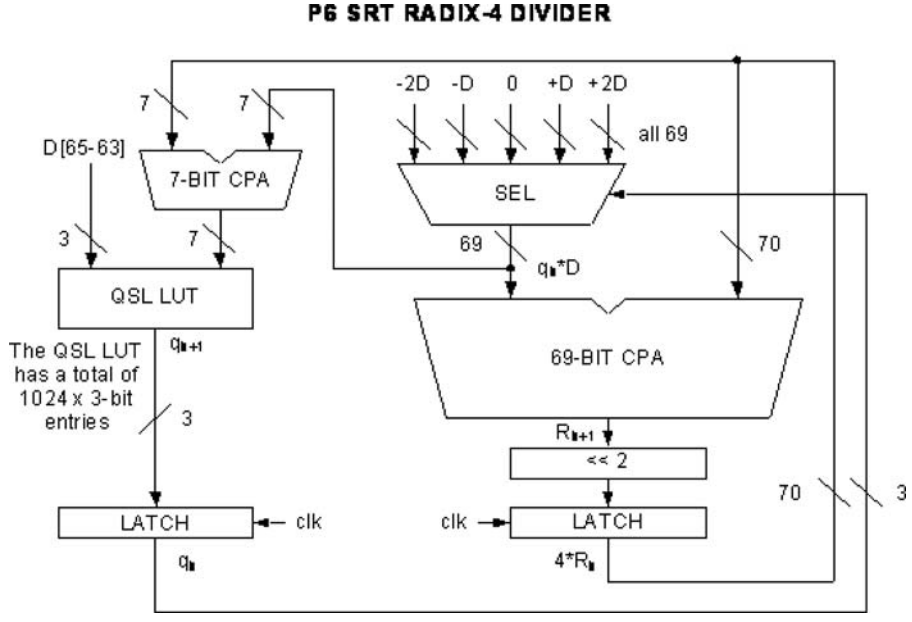

Figure 8: *Previous divider microarchitecture.*

avoid a new pre-serialization condition. The results are that the throughput of a CLI is 5 cycles, the throughput of an STI is 8 cycles, and a CLI-STI pair is 13 cycles, yielding a 2.5× improvement over the Merom architecture performance.

INCLUSION FILTER

The inclusion filter enables detection of instructions in the processor pipeline, mainly to support self-modifying code (SMC). To reduce design impact on this timing-sensitive area of the processor pipeline, detection techniques are architecturally minimized to provide pessimistic estimates. The goal is a logically optimized solution in which false SMC detection is sufficiently uncommon such that the resulting performance loss is negligible.

Motivation

As the processor pipeline capacity increases, the inclusion detection solution needed to be re-examined. The pipeline capacity includes instructions in all stages and structures between instruction fetch and retirement, which increased substantially when the issue width was increased for the Merom architecture from 3 to 4. The increased instruction capacity resulted in increased false SMC detection conditions during Merom silicon testing, which tended to have a more limiting impact on server performance. For example, Transaction Processing Performance Council Benchmark C performance increased by 2 percent with inclusion checking disabled. The Inclusion Filter in the Penryn processor significantly reduces false SMC detection by using an alternative technique to filter from the existing detection mechanism the most common false detection scenarios.

Solution

Most instructions in the pipeline will also naturally exist in the Instruction Cache (ICache), so the Inclusion Filter monitors ICache activity to algorithmically identify states in which this common property is guaranteed (Figure 9). Snoops in this state can then be filtered from the existing inclusion-detection mechanism, and this combination virtually eliminates false SMC detection.

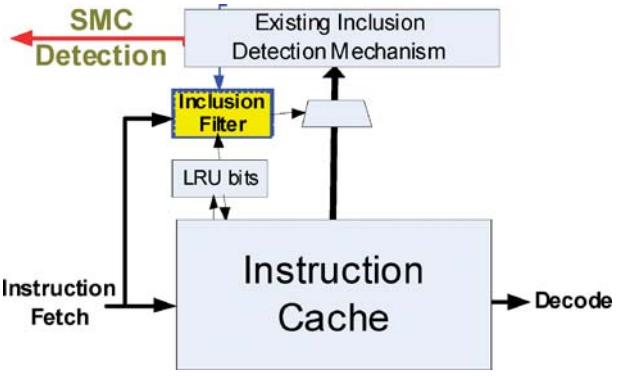

Figure 9: ***Inclusion Filter reduces false SMC detection.***

To increase confidence in this new microarchitectural solution, it was essential to minimize design complexity. To reduce logic risk and validation requirements, the Inclusion Filter has a single functional output. To avoid the risk of frequency degradation, it is logically separated from the existing ICache structure (which is ideal for separating logic vs. process-related debug) and uses only non-timing-critical signals, such as the ICache LRU bits.

For example, a property of the ICache pseudo-LRU algorithm is that for an X-way configuration, an accessed entry will not be evicted until at least log2 (x) different entries in the same set have been accessed. Therefore, for an 8-way cache, each set is allowed to filter at least three ICache evictions prior to resorting to the existing inclusion detection mechanism. Determining a "different entry" can be accomplished without additional storage by detecting a change of LRU bits. Monitoring changes to specific LRU bits and other control logic can increase the limit substantially using other similar properties.

When the Inclusion Filter is saturated and finally allows the existing mechanism to be used, it is more beneficial to have it return to its reset state than to continue filtering. From a reset, the average cycles needed for the Inclusion Filter to resort to the existing inclusion mechanism is 50 times greater than the cycles needed to ensure that a fetched instruction is no longer in the pipeline. Therefore, when the Inclusion Filter is finally saturated, it takes the opportunity to completely reset its state, but it disables filtering until it is certain that all instructions in the pipeline at the time of this reset have been retired.

In effect, a small window is opened during which the existing detection method is used, then it is closed for a very long time (98 percent closed on average). This translates to a 98 percent reduction in false SMC detection, and near optimal performance.

RENAMED RSB

A Renamed Return Stack Buffer (RRSB) was added to improve performance by increasing return prediction accuracy. The goal was to supplement the existing RSB by providing a recovery mechanism from a common source of RSB corruption.

Background

A single function (or procedure) can be called from multiple places within a program by using a "CALL" instruction. Exiting the function back to the calling program can be done with a "RET" (return) instruction. The CALL instruction is similar to a direct jump that also pushes the RET address onto the stack (in memory). The RET instruction is an indirect jump whose target address is popped from the stack.

The processor's Branch Prediction Unit (BPU) shares both its bimodal prediction resources to accurately predict the existence of CALL or RET instructions and also its Branch Target Buffer (BTB) to predict the target of a direct CALL. However, the target of a RET instruction is dependent on the CALL, so the Return Stack Buffer (RSB) is used.

All P6 microprocessors have implemented the RSB as a simple push pop stack structure. This "classic" RSB (CRSB) has the following basic behavior:

1. The BPU uses its Linear Instruction Pointer (LIP) to predict a CALL instruction.
2. The BPU "pushes" the CALL's Next Linear Instruction Pointer (NLIP) onto the CRSB stack.
3. The BPU predicts the target of the CALL from the BTB and redirects the instruction flow.
4. Later, the BPU predicts a RET instruction based on its LIP.
5. The BPU predicts the target of the RET from the CRSB and redirects the instruction flow.

CRSB corruption

Useful RET predictions in the CRSB are sometimes overwritten by bogus speculative updates. These bogus updates should be corrected after a branch misprediction to ensure accuracy. This requires saving the CRSB state for each potential misprediction and restoring that state after misprediction recovery. Practically, however, we can save only the CRSB Top-Of-Stack (TOS) pointer that is stored in the Branch Information Table (BIT). When the CRSB TOS is restored from the BIT, the contents may have been overwritten while traversing down the bogus path. For instance, if the bogus path has a RET followed by a CALL, a valid return address will be overwritten that will later result in a performance penalty. The TOS pointer will be restored, but the CRSB contents are corrupted. Figure 10 (top) describes this common CRSB corruption scenario.

Renamed RSB implementation

To address this corruption, we added the "Renamed RSB" (RRSB) to the Penryn family of processors. The RRSB is similar to the CRSB, but it incorporates an additional pointer (Alloc) and a linked-list structure for updating the TOS. Figure 10 (bottom) shows how the RRSB is able to recover from bogus updates. The pointers are updated as follows:

- The CALL NLIP is written to the Alloc entry. The TOS pointer is adjusted to point to the Alloc entry, and then the Alloc pointer is incremented (Column 3 in Figure 10). The Alloc pointer never decrements. The TOS linked-list is updated to retain the previous TOS.

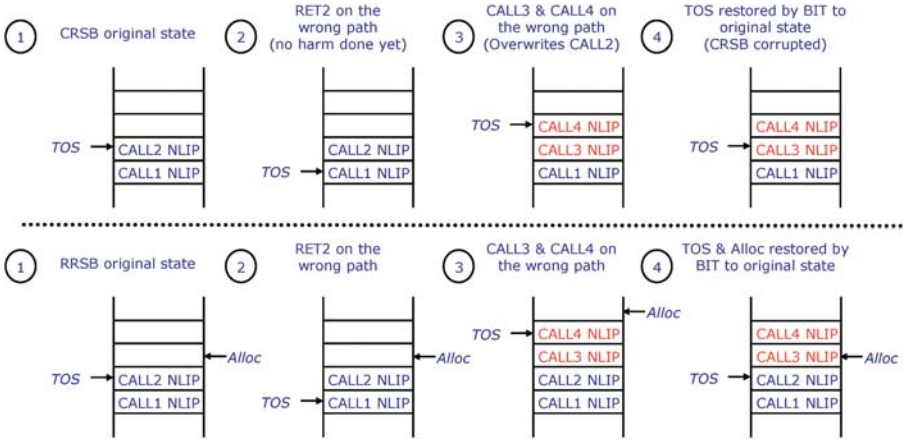

Figure 10: ***CRSB vs. RRSB.***

- The RET target is read from the TOS entry and uses the linked-list to adjust the TOS pointer to the previous TOS. The Alloc pointer is not updated on RET instructions.

The CALL NLIP is never overwritten and therefore retains entries that may be lost by the CRSB.

While the RRSB is more accurate on the speculative path, it overflows (wraps) more quickly since Alloc never decrements. Therefore, the return prediction defaults to the CRSB when RRSB detects the wrap condition. We added a 16-entry RRSB to the architecture that works in conjunction with the 16-entry CRSB as shown in Figure 11.

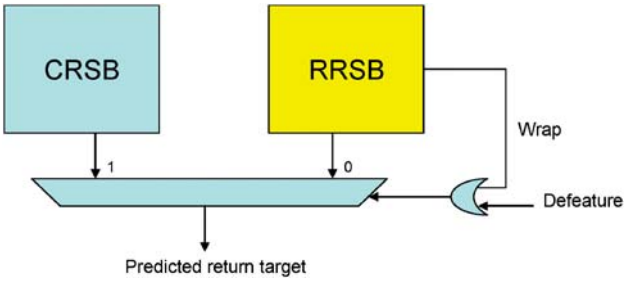

Figure 11: ***RRSB implementation.***

CONCLUSION

The improvements described in this paper should provide the reader a better understanding of the value architectural and microarchitectural enhancements have brought to the marketplace above and beyond a silicon process improvement. Each feature improves some aspect of performance so that in conjunction with the frequency improvement, the end user will realize real value in the product.

For a quantification of the performance improvement, please refer to the paper "Original 45nm Intel® Core™ 2 Processor Performance" in [1].

ACKNOWLEDGEMENTS

We thank the following contributors to our paper: Ronen Zohar for providing the collision detection example; Ashish Jha for providing the streaming load kernel white paper; Lihu Rappoport for his contribution to the Renamed RSB; and Mohammad Abdallah for his contributions to the development of the divider and SSE4.1.

REFERENCES

[1] Nisar A., Ekpanyapong M., Valles A., and Sivakumar K., 45nm Intel® Core™2 Processer Performance, Intel Technology Journal, Vol. 12, No. 3, 2008.

[2] Jha A. Yee D. Increasing Memory Throughput With Intel® Streaming SIMD Extensions 4 (Intel(4) SSE4) Streaming Load. http://softwarecommunity.intel.com/articles/eng/1248.htm.

[3] Intel® 64 and IA-32 Architectures Optimization Reference Manual. At http://www.intel.com/products/processor/manuals/.

[4] Intel® 64 and IA-32 Architectures Software Developer's Manual. At http://www.intel.com/products/processor/manuals/.

[5] Atkins D.E. 'Higher radix division using estimates of the divisor and partial remainders.' IEEE Transactions on Computers, 1968; C-17: 925–934.

[6] Parhami B. Tight upper bounds on the minimum precision required of the divisor and the partial remainder in high-radix division. IEEE Transactions on Computers, Vol. 52, No.: 11, November 2003, pp. 1509–1514.

[7] Wey C.-L., Wang C.-P. Design of a fast radix-4 SRT divider and its VLSI implementation. Computers and Digital Techniques, IEE Proceedings, Vol. 146, No. 4, July 1999, pp. 205–210.

[8] "Design issues in radix-4 SRT square root & divide unit" Burgess N., Hinds C. "Signals, Systems and Computers." *Conference Record of the Thirty-Fifth Asilomar Conference,*Vol. 2, November 4–7, 2001, pp. 1646–1650.

AUTHORS' BIOGRAPHIES

Jim Coke is a Staff Architect and Microcoder in Intel's Mobile Microprocessor Group in Folsom, CA. He received his B.S.E.E. degree from the University of Michigan and his M.S.C.E. degree from the National Technological University. Jim joined Intel in 1982 and has worked in product engineering, design, and architecture. Jim was the lead implementation architect for SSE4.1. His primary interests are microcode and microarchitecture. His e-mail is James.S.Coke at intel.com.

Hari Baliga is a Senior Staff Engineer in Intel's Mobile Microprocessor Group in Folsom, CA. He received his Bachelor of Engineering degree from Regional Engineering College, Surathkal in India and his M.S. degree from Arizona State University in Tempe. Hari joined Intel in 1996 and has worked on many microprocessors developed by the Folsom Design Center. His e-mail is harikrishna.baliga at intel.com.

Niran Cooray is a Senior Staff Architect in Intel's Mobile Microprocessor Group in Folsom, CA. He received his B.Sc. degree from the University of

Moratuwa in Sri Lanka and his M.S. degree from Northeastern University in Boston. Niran joined Intel in 1995 and has worked on many microprocessors developed by the Folsom Design Center. Niran worked as a Senior Design Leader on P6-based microprocessors before moving on to become a microarchitect for the Intel® 45nm Core™ 2 Duo processor. His email is Niranjan.L.Cooray at intel.com.

Edward Gamsaragan is a Staff Architect in Intel's Mobile Microprocessor Group in Folsom, CA, working there since 1995. For the Penryn family of processors, he was the microarchitect responsible for the out-of-order and execution clusters. His current focus is on next-generation memory technologies. Ed holds a B.S.E.E. degree from the University of California at Los Angeles. His e-mail is Edward.Gamsaragan at intel.com.

Peter Smith is a Senior Architect with Intel's Mobile Platform Group in Folsom, CA. For the Penryn family of processors, he was the architect responsible for the front-end and MSID clusters. His previous experience includes software design, system administration, circuit design, silicon debug, and performance analysis. His primary interests include probability theory, heuristics, and creative problem solving. Peter received his B.S. degree from the University of Wisconsin-Madison and joined Intel in 1996. His e-mail is Peter.J.Smith at intel.com.

Ki Yoon is a Senior Staff Architect with Intel's Mobile Platform Group in Folsom, CA focusing on microprocessor microcode and debug. Ki developed microcode and played a key role in system debug on the Intel Pentium® III and the Intel® Core™ 2 processor generations. Most recently, Ki was involved in the definition of the 45nm Intel Core 2 Duo processor architecture and Intel Virtualization Technology. He received his B.S. degree from the University of Texas at Austin in 1994. His e-mail is ki.w.yoon at intel.com.

James Abel is a Principal Engineer in Intel in Chandler, Arizona. James obtained a Bachelor's Degree in Electrical Engineering from Bradley University in Peoria, Illinois in 1983 and a Master's Degree in Computer Science from Arizona State University in 1991. His interests include computer architectures, performance analysis tools, digital signal processing, and multimedia algorithms. His email is James.C.Abel at intel.com.

Antonio Valles is a Senior Software Engineer in Intel in Chandler, Arizona focusing on broad and in-depth pre-Si and early-Si tests of Intel microprocessors and chipsets. Antonio has created multiple internal pre-Si and post-Si tools and kernels for performance analysis and coordinates the development of tuning guidelines for the processors. He received his Bachelor's Degree in Electrical Engineering from Arizona State University in 1997. His email is antonio.c.valles at intel.com.

Mobility Thin and Small Form-Factor Packaging for Intel® Processors Based on Original 45nm Intel Core™ Microarchitecture

Kyan Pedrami, Mobility Processor Design Team, Intel Corporation
Satish Prathaban, Mobility Platform Hardware Development Team, Intel Corporation

Index words: penryn, small form factor

Citations for this paper: Kyan Pedrami, Satish Prathaban "Original 45mn Intel® Core™2 Processor Performance" Intel Technoloy Journal. http://www.intel.com/technology/itj/2008/v12i3/4-paper/1-abstract.htm (October 2008)

ABSTRACT

The Intel 45nm processor, originally referred to by the codename Penryn, based on Intel Core™ microarchitecture, is the first processor that uses multicore-on-die design to maximize performance and minimize power consumption. This paper provides an overview of how the Penryn processor's mobility platform and package design teams delivered the Penryn silicon in smaller and thinner packages that enabled customers to design both smaller and thinner form-factor platforms. It also provides insight into the mechanical and electrical challenges of these families of thin, small, form-factor packages, challenges that were overcome without incurring significant performance degradation.

INTRODUCTION

The Intel® processor, originally referred to by the codename Penryn, is the first 45nm processor that is based on Intel Core™ microarchitecture that uses multicore-on-die design to maximize performance and minimize power consumption. It is also the first "all-green" Intel 45nm processor product that is lead and halide free. Historically, Intel mobility processors were packaged in a 35-mm × 35-mm substrate, in eight layers of package routing, with a z-height of 2.89 mm, on a socket, and with a pin pitch of 1.27 mm. This footprint and z-height were design targets for the Penryn mobility processor family. During the life cycle of the Penryn family of processors, two major packaging design requirements were introduced. First, a cost reduction target was identified for the Penryn family of processors. The team responded with innovative designs implementing both six- and four-layer packages, compared to the traditional eight-layer package. Second, a customer requirement was introduced for a smaller and thinner package option. The team was able to leverage the lower-layer-count proposal in order to achieve the new requirement profile. This new package design required a processor package z-height of less than 2 mm, with a footprint of 22mm × 22 mm, a pin pitch of 0.952 mm, and a pinmap, supporting the High Density Interconnect (HDI) board. The reduction of the package footprint size meant packing all the signals within the smaller package with negligible additional crosstalk. The challenges of such a design resides in processor power delivery, due to the lower number and placement of package capacitors in the available space constraints. In both cases, the design team was faced with the challenge of fitting a robust core power-delivery system, with negligible performance impact, into a package smaller than the traditional mobility packages.

The Penryn mobility processor with a 3-MB cache form-factor was the first processor in this family to be packaged into the lowest possible package stackup; and with a 6-MB cache, was the first mainstream Small Form Factor (SFF) mobility product to tape out a package in just nine weeks with pinmap redefinition and bottom-up package design. In this paper, we provide an overview of how the Penryn family of processors' mobility platform and package design teams delivered the Penryn silicon in smaller and thinner packages that enabled customers to design both smaller and thinner form-factor platforms. We provide insight into the mechanical and electrical challenges of these families of thin and SFF packages without incurring significant performance degradation. We also explain how design team members, located across multiple geographical

areas, synchronized their work for maximum productivity to achieve multiple package design breakthroughs in mobility package design.

OVERVIEW OF THE 3-MB PIN GRID ARRAY AND SMALL FORM FACTOR PACKAGES

In the context of mobility package design, the Pin Grid Array (PGA) package was traditionally designed first, and later on, the PGA design was converted directly to a Ball Grid Array (BGA) package by replacing the pins on the design with balls. Doing so enabled the back-end team to concentrate on one processor package form-factor validation and helped the package design team to focus on one design. As an additional benefit, investment in parts and validation tools to test two simultaneous designs was not necessary. The downside for the "one package design fits all" scenario was that the final implemented package contained all design requirements of both PGA and BGA packages combined, which results in very small cost optimization. Initially, this same package design strategy was planned for the Penryn mobility family of processors. This 35-mm × 35-mm socket has not been changed for three generations due to a backward compatibility requirement; however, in the case of the mobility processor, the pinmap was slightly modified to meet manufacturing and reliability requirements. This backward-compatibility feature meant the same 35-mm × 35-mm PGA package, socket, and pinmap could be used throughout multiple processor designs. This reuse helped in the verification of the new package on the old platform thus enabling customers to reuse their mechanical and thermal solutions from the previous platforms, an obvious reduction in design time and cost. Traditionally, the mobility processor packages were also designed with an eight-layer stackup with the top layer dedicated for Front Side Bus (FSB) routing. The original Penryn mobility processor package design stackup was eight layers. The four-package internal layers were used exclusively for processor and I O power delivery. The focus of the package design team when designing the Penryn family of processors' mobility package was to also optimize the package layer count, and if possible, optimize the package footprint.

With the introduction of the Penryn family of processors' cost-saving challenge, the team analyzed opportunities to reduce package cost. In 45nm design technology, it is feasible to reduce the overall number of package layers to six or even four, netting a significant cost reduction throughout the life of the product. For the 3-MB 35-mm × 35-mm package, the team pursued a four-layer PGA package design. The design practice of serial development of PGA and BGA designs was no longer followed, and packages were instead developed in parallel, making the two designs no longer dependent on one other. Although this approach now required validation of each form factor, this new approach facilitated the removal of BGA design elements from the PGA design. Coupled with the reduction in the number of layers, these innovations drove significant cost savings in the Penryn 3-MB processor's PGA package manufacturing cost. The design team was able to further leverage these advantages in response to a customer request for an SFF product. With the removal of PGA elements from the BGA package, the team was able to implement the design in a 22-mm × 22-mm footprint with a six-layer stackup. The socketless nature of the BGA package, coupled with fewer layers, enabled the overall z-height profile of the product to be reduced.

PACKAGE DESIGN—SMALL FORM FACTOR

In conjunction with a reduced number of layers in packages, many customers requested a reduced processor package size that could enable the design of a smaller platform form factor. We analyzed the external requirements and internal capabilities and concluded that package design size should be 22 mm × 22 mm for the Penryn BGA SFF package. Moreover, customers were also ready to use the HDI board, if the package size could be reduced. Taking advantage of the 22-mm × 22-mm package footprint, we chose a diagonal staggered pin pitch of 0.673 mm.

There were many roadblocks to clear during this phase of the design: solder joint reliability concerns and IO routing to support customers' Layer-1 and Layer-3 routing on platform.

Due to the solder joint reliability balls in the Penryn 22-mm × 22-mm package, the BGA package allowed only one column of signals for IO power delivery on the data side.

Another constraint was that the pin pitch of the package completely blocked the direct path for west-to-east power delivery. A novel method was introduced to feed the IO power from south to north by extending the pinmap down south, thereby allowing the power to enter from the south.

The pinmap was also adjusted so that it could support the HDI board. HDI is a type-4 board, with buried vias that help to reduce the package size. This is because the Plated Through Hole (PTH) does not extend up to the solder side, as via pad-to-pad spacing limits the pin pitch. In the HDI board design, motherboard Layers 1 and 3 are used

for signal routing. Most of the earlier platforms used eight layers, with Layers 3 and 6 being the routing layers, and Layers 7, 5, 2, and 4 being ground planes. In the Penryn family of processors' SFF HDI platform, one channel (odd bytes) of FSB was routed on the top layer (Layer 1) with Layer 2 as a reference (microstrip routing). The other channel (even bytes) was routed on Layer 3, with Layers 2 and 4 being ground reference planes. This method of routing enabled lower-layer board design but with the added cost of manafacturing HDI boards. This meant separate signal integrity (SI) analysis to validate both the microstrip and the stripline routing.

PLATFORM AND SUBSTRATE POWER-DELIVERY CHALLENGES AND IMPLEMENTATIONS

The standard voltage (SV) package of the Penryn family of processors is designed to enable the maximum core frequency. Frequency of operation is a function of the minimum voltage provided to the circuits in the processor; so the tighter the tolerance of the voltage at the processor, the higher the frequency that can be obtained at a given voltage.

In the PGA package, land-side capacitors act as the high-frequency decoupling solution for the package; however, due to BGA package construction, there is no cavity in which to place land-side capacitors. In both 35-mm^2 PGA 3-MB with four-layer stackup and 22-mm^2 BGA 6-MB or 3-MB with six-layer stackup packages, the surface layer has the FSB routed as the microstrip. Due to the space constraints and reduced package size, FSB lengths in the BGA package were not matched to PGA. The processor power delivery in both packages is from north to south. Unlike the PGA, in which the IO power delivery is from east to west, in the BGA, pin pitch does not allow the power feed from east or west, so the IO power delivery is from south to north.

Since the Penryn family of processors' 3-MB PGA processor had to follow the pinmap of the 35 mm × 35 mm package and the socket compatibility, the package design team concentrated only on reducing the layer count in the 3-MB Penryn family of processors.

After simulation and lab analysis on previous eight-layer stackup packages, the team determined that eight-layer packages were too robust for the Penryn family of processors' power-delivery requirements. First, two varieties of six-layer packages were evaluated. Due to BGA z-height requirements, the design required a thinner organic stiffener for package core material in the BGA compared to the PGA. In mobility packages, since the FSB was predominantly routed as microstrip on the outer top layer, only the power-delivery solution and signal-referencing layers needed to be resolved for packages with a reduced number of layers.

We realized from preproduction power-delivery analysis that removing two core layers from the original eight-layer package had minimal effect on processor performance; however, doing so substantially reduced package manufacturing costs.

Based on all preproduction findings, we decided to take a calculated risk and design the final Penryn family of processors' mobility SFF BGA package with optimized package layers, changing from the eight-layer original design to a six-layer package. This resulted in a huge savings in manufacturing costs, and moreover, as a result of fewer package layers, the total z-height was reduced, as per our customer's dictated request.

Preproduction, and post-package production data correlations, in conjunction with lab analysis of both Signal Integrity and power delivery of the Penryn family of processors' 6-MB PGA six-layer, showed that a four-layer package will be robust enough for the Penryn family of processors' 3-MB PGA mobility processor product.

In the PGA package, the decoupling capacitors are placed in the cavity directly below the die that provides the shortest path for processor discharge (Figure 1).

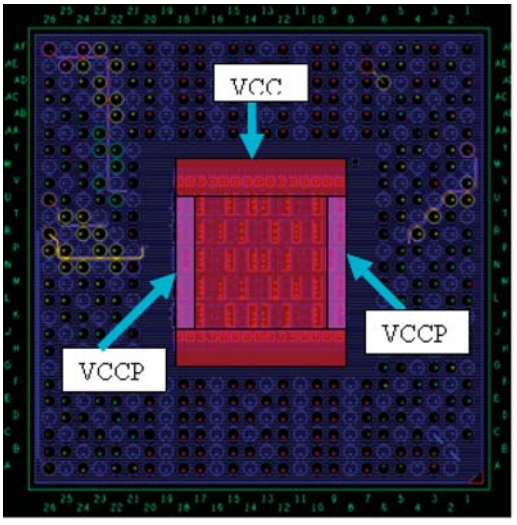

Figure 1: ***PGA land-side capacitor cavity.***

The load-line impedance is the target impedance for the power-delivery network from the voltage regulator to the processor voltage sensing points, and its impedance characteristic in the frequency domain can be extracted from the processor voltage sense points.

This characteristic can be divided into three major contributors: at low frequency (otherwise known as "third droop"), the voltage regulator and motherboard are the dominant contributors; at mid-frequency (otherwise known as "second droop"), the socket and package are the main contributors; and at high frequency (otherwise known as "first droop"), the processor itself is the main contributor.

We simulated the impedance profile for each of the package options and compared them with their previous-generation predecessor. At first droop the impedance was higher. We conducted many experiments on the previous packages by removing capacitors on the package and increasing the first droop impedance. The experiments showed that the impedance can be increased by a factor of 2 without impacting the frequency of operations.

The PGA decoupling solution uses 30×0306 and 30×0402 package land-side capacitors for the core power delivery. The board decoupling solution for core power delivery is 6×330 uF [ESR per capacitor = 9 mohms]. The mid-frequency capacitor on the board is either $12 \times 0805 \times 22$ uF in the cavity region or $16 \times 0805 \times 22$ uF on the bottom-side of the board. The IO FSB is on the east and the west of the processor with the data bus on the east and the address bus on the west of the die. The PGA package cavity could accommodate 5×0306 on either side for IO power delivery. Similarly, the decoupling solution for IO power delivery is a 6×0402 capacitor on the bottom-side of the board. Figures 2 and 3 show the location of the power-delivery capacitors on the PGA motherboard.

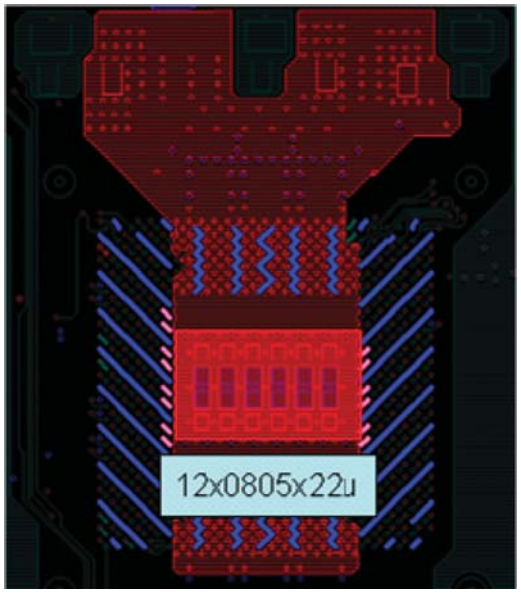

Figure 2: ***Motherboard top layer power delivery for PGA.***

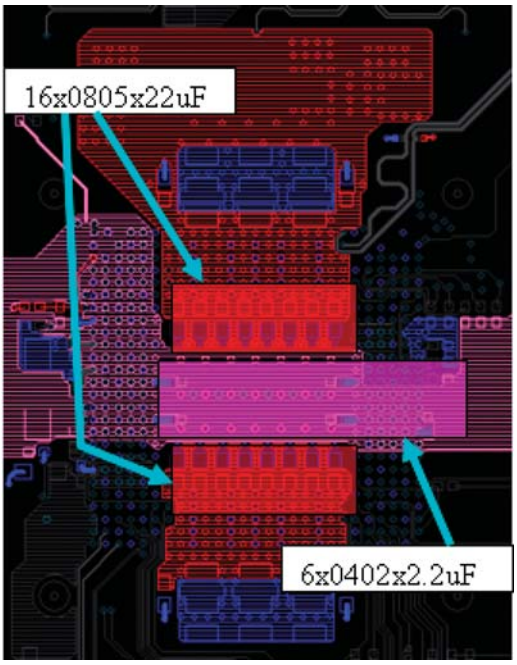

Figure 3: ***Motherboard bottom layer power delivery for PGA.***

The PGA package via pattern in the core area is chosen so that the loop inductance is low so as to reduce the first droop. The voltage regulator and motherboard load-line are kept at 2.1 mohms, the same as predecessor designs, for package backward and forward compatibility, so that customers can reuse their past designs.

The package power delivery for the SFF BGA package that lacks land-side capacitors consists of making the processor side capacitors accommodate both core and IO power delivery. The SFF BGA package contains 4×0402 capacitors on the north and 5×0402 capacitors on the south, providing the processor power-delivery solution for the package. The 4×0201 on the east and 4×0201 on the west provide the IO power-delivery solution for the package.

We designed the core power-delivery solution to meet a 4-mohm load-line with platform capacitors of $24 \times 0603 \times 10$ uF and $24 \times 0402 \times 1$ uF. The motherboard stackup is an eight-layer HDI stackup. We chose via patterns so as to reduce the loop inductance from the back-side capacitors of the board to the package. Our inability to put the land-side capacitors in the BGA with the pin pitch increases the loop inductance. The board IO power delivery required $6 \times 0402 \times 1$ uF on the east and the west (see Figure 4).

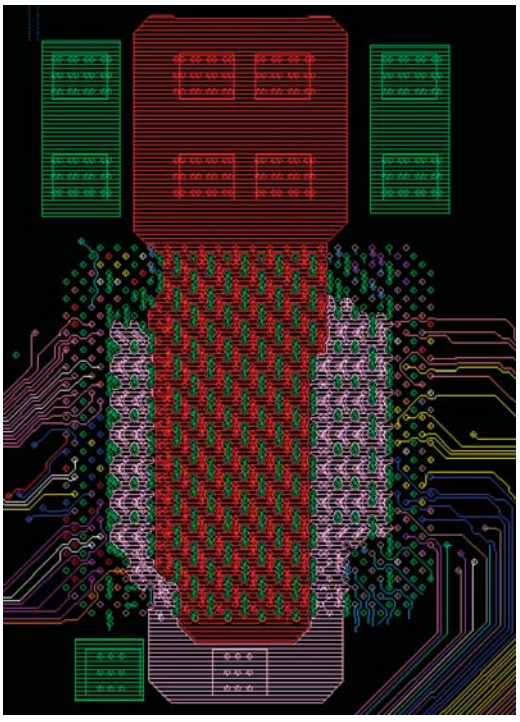

Figure 4: ***Platform power-delivery solution for the Penryn family of processors SFF BGA.***

In the end, the core power-delivery simulation results for both packages were closely correlated with the IFDIM [1] measurement in the validation cycle. Also, the simulation model was correlated to the Pico probe measurement on the package processor voltage sense points.

PACKAGE DESIGN TEAM'S CHALLENGE

The mobility package design team for the Penryn family of processors included only four team members: a design engineer, a requirement engineer, a platform power-delivery engineer, and a layout designer. This extremely small team faced the additional challenge of being geographically dispersed. Initially, team coordination was extremely challenging, especially since the dispersal of team members spanned multiple continents with all the inherent time differences. After aligning the work flow, this setup actually worked to the team's advantage and enabled the team to meet and beat multiple package design deliverables. For example, at the beginning of the typical cycle, team members located in the U.S. did what-if changes and analysis to the package design database. At the end of the U.S. workday, designers passed on the preliminary design database to the package layout team member in Malaysia. The Malaysian team member performed layout design rule cleanup and production-worthy implementation of changes and then, at the end of the Malaysia work day, passed on the design to the team member located in India. The Indian team member completed power-delivery network extraction, did the simulations from layout, and identified any processor performance impacts. In the event that processor IO FSB performance might be impacted, simulation models for stakeholders located in Israel were generated. Israeli stakeholders used layout-modeled data for FSB performance impact analysis, identified issues, and recommended solutions during their working hours. By the time US team members came back to work on their next business day, the team had already completed the extraction, simulation, and analysis of "yesterday's" work, thereby compressing the four-day wait time of co-located design teams into one twenty-four-hour work cycle. This work model was perfected and used by the package design team throughout multiple Penryn mobility processor package design flavors. It enabled the team to design more than seven mobility packages with less than typical staffing, and in one case, finished four weeks earlier than the originally agreed-upon package tapeout schedule.

CONCLUSION

By coordinating the roles and workflow of the geographically distributed team members, this small group was able to deliver the Penryn family of processors' 3-MB design implemented with the lowest possible package layer count and capacitors, returning significant cost savings over the life of the product. Likewise, the team was able to leverage this design strategy to tape out the first mainstream SFF mobility product in just nine weeks, despite the challenge of pinmap definition and bottom-up package design, thus enabling customers to design both smaller and thinner form-factor platforms.

ACKNOWLEDGEMENTS

We recognize Nicole Pedrami and Rob Milstrey for their contributions to the content and their review of this paper.

REFERENCES

[1] A. Waizman . "Integrated power supply frequency domain impedance meter (IFDIM)" In Proceedings of IEEE: 13th Topical Meeting on Electrical Performance of Electronic Packaging 2004, pp. 217–220.

AUTHORS' BIOGRAPHIES

Kyan Pedrami is a Staff Enginner in the Mobility Processor Design team and has been with Folsom Design Center since 1997. Before joining Intel, he worked at Digital Semiconductor in Maynard, MA. Kyan obtained his B.A. and M.A. degrees in Electrical Engineering from the Georgia Institute of Technology. His e-mail is kyan.pedrami at intel.com.

Satish Prathaban is a Senior Hardware Design Engineer in the Mobile Platform Hardware and Development team. He has an M-Tech degree from IISc-Bangalore. He started working at Intel India in 2003 in the area of motherboard and package power delivery. His e-mail is satish.prathaban at intel.com.

The Technical Challenges of Transitioning Intel® PRO/Wireless Solutions to a Half-Mini Card

Eli Laks, Mobility Group, Intel Corporation
Richard S. Perry, Mobility Group, Intel Corporation
Brad Saunders, Mobility Group, Intel Corporation
Ra'anan Sover, Mobility Group, Intel Corporation

Index words: Intel® PRO/Wireless, half-mini card, HMC, mini card, Wi-Fi, MIMO

Citations for this paper: Eli Laks, Richard S. Perry, Brad Saunders, Ra'anan Sover "Original 45nm Intel® Core™ 2 Processor Performance" Intel Technology Journal. http://www.intel.com/technology/itj/2008/v12i3/5-paper/1-abstract.htm (October 2008).

ABSTRACT

With the introduction of Intel's latest mobile platform, Montevina, which is based on the new Penryn Mobile family of processors, the Intel® Pro/Wireless 5300 and 5100 communication daughter boards, offered as part of the Intel Centrino® Mobile Technology platform, also underwent major changes to enable a smaller, more efficient platform solution. The 5300/5100 family of Wi-Fi wireless communications boards is now offered in a Half-Mini Card form factor. This paper describes some of the technical challenges faced when transitioning from a Full-Mini Card used in previous-generation wireless solutions to a half-size card while increasing the functionality on board. These challenges have directly affected the entire design from the core silicon all the way up to the complete product board, not to mention some of the challenges to address at the Peripheral Component Interconnect Special Interest Group (PCI-SIG), from a standardization point of view. Background information will be provided to better understand why this transition was driven in the platform.

In this paper, we will touch on the required Printed Circuit Board (PCB) technology, front-end integration, silicon floor planning, pinout definitions, and the thermal considerations necessary to enable this transition. Further, we will show how this new form factor differs from its predecessors in some key aspects, as wireless communication has progressed from generation to generation. The reader will gain a good understanding of some of the technological challenges driven by this form-factor change that will enable smaller, more condensed platform solutions.

INTRODUCTION

In this paper we give a brief overview of the technical challenges involved in transitioning the Intel® PRO/Wireless Wi-Fi solutions from a Full-Mini Card to a Half-Mini Card form factor. (We use the terms "Full-Mini Card" and "Mini Card" interchangeably.)

Over the past generations of Intel PRO/Wireless solutions, the team has been asked to continually increase functionality while decreasing board size. Figure 1 shows the evolution of form factors starting with the original Mini Peripheral Component Interconnect (PCI) form factor of choice at the launch of the first Intel Centrino® Mobile Technology (CMT) platforms back in 2004–2005. The Intel PRO/Wireless 2100/2200/2915 series were all implemented using this Mini PCI Card form factor. In 2006, a new smaller form factor and Host Interface were introduced into the platform—the Mini Card—with a high-speed PCI Express interface. The intent of this smaller form factor was to enable the incorporation of two Mini Cards in the available space of the older Mini PCI Card, thus enabling more functionality in the platform. The Intel PRO/Wireless 3945abg was Intel's first IEEE 802.11abg Wi-Fi solution using this new form factor, which was part of the Napa family of CMT platforms. As the IEEE 802.11 Wi-Fi standard evolved, a new higher throughput technology was introduced called IEEE 802.11n that enables a Multiple In Multiple Out (MIMO) communication scheme. One of the key features of this new scheme is higher data throughput. The Intel PRO/Wireless 4965 was Intel's first IEEE 802.11n Draft MIMO solution to the market. It has two transmitters and three receive chains for data rates

up to 300 Mbps, using the same Mini Card form factor: in previous-generation technologies, there was only a single transmitter and a single receive chain. This amounts to a functionality compaction factor of 2.5:1 and a data throughput improvement factor of 5:1. This new wireless solution was introduced into the market at the end of 2006 and continues to be provided on the Santa Rosa CMT platforms.

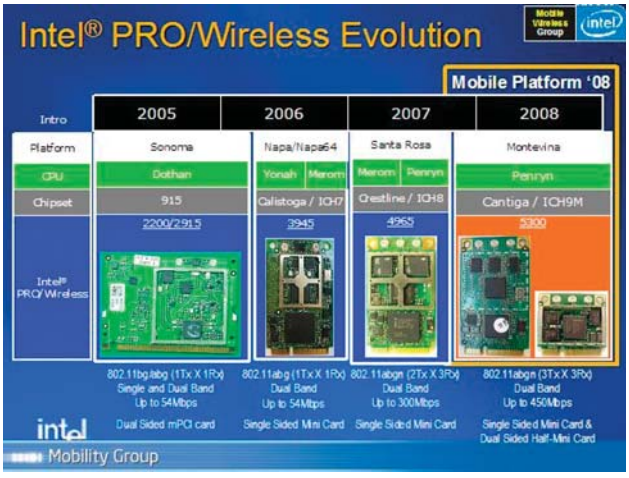

Figure 1: *Intel® PRO/Wireless evolution.*

However, the 'ever hungry' market demands more functionality in less space. This market demand drove the development of a new form factor called the Half-Mini Card. Again, the intent is to enable platform integrators to put more wireless content into the platform by placing two Half-Mini Cards in the space where they once put a single Full-Mini Card just one generation ago. With respect to performance, the intent is to also improve the throughput by a factor of 1.5:1 thereby reaching data throughputs of up to 450 Mbps.

In this paper, we describe some of the technical challenges and solutions the team faced while implementing this newer, smaller Half-Mini Card form factor while also increasing the functionality of the Wi-Fi system to a full three-transmitter and three-receiver 802.11n MIMO functionality. We discuss these issues in this paper:

- Half-Mini Card form factor standardization in the PCI-Special Interest Group (PCI-SIG)
- The Intel PRO/Wireless 5000 series of network adaptors
- Component and functionality partitioning
- Mechanical and physical requirements
- Printed Circuit Board (PCB) requirements
- Radio Frequency (RF) component sizes and challenges to meet regulatory emission certification requirements
- Thermal considerations and challenges

The team successfully met these challenges with the introduction of this latest family of Intel PRO/Wireless solutions 5300 and 5100 that are an integral part of the Montevina CMT platforms, by using the new Half-Mini Card form factor. These network adaptors will also be offered to customers in the older Full-Mini Card for the benefit of Original Equipment Manufacturers (OEMs) who continue to support the older, larger, single-sided form factor. This fact will also serve as a basis for comparison to better explain the challenges of converting the same product to the new, smaller Half-Mini Card form factor.

DEVELOPING THE PCI EXPRESS HALF-MINI CARD SPECIFICATION

With an increasing market pressure to integrate more and more wireless radio functionality into thinner and lighter notebook designs, the PCI-SIG in early 2004 was asked by some platform OEMs to consider space-saving alternatives to the PCI Express Mini Card specification that had been released only one year earlier and had yet to even have products developed based on it. Platform OEMs were heavily motivated by the need to figure out how to enable getting an increasing number of separate wireless radios into the already tightly-packed base of the notebook and by the need to establish a development roadmap toward a more space-efficient card format that wireless technology suppliers could eventually move to.

Starting in December 2005, the PCI-SIG Mini Working Group (WG) initiated the development of a specification for what would ultimately come to be known as the Half-Mini Card. The primary objective was to define a smaller variant of the PCI Express Mini Card, now to be known as the Full-Mini Card, which would enable notebook designs to accommodate an increased number of wireless cards while keeping the platform base volume associated with wireless applications at parity. The goal was to potentially get two smaller cards in the space of the one larger card while retaining interface and socket connector compatibility across the two card types. In this effort, the Mini WG, consisting of ten voting and seven observing member companies, succeeded in completing an acceptable specification in just over six months.

Figure 2 overlays color-highlighted card outlines aligned at a top-right origin to visually compare the decreasing planar size progression as the standardized card form factors shrank from the original Mini PCI Card (shown

in green) down through to the Full-Mini Card (blue) and toward the Half-Mini Card (red). It should be noted that no change to the z-height and assembly stack-up profile of the Mini Card was made going from the Full-Mini Card to the Half-Mini Card format. If solely based on the outline dimensions of the Half-Mini Card, the format appears to be slightly larger than half the size of the Full-Mini Card, 804 mm^2 versus 1528.5 mm^2, but the practical area for functional circuitry collectively across both top and bottom sides of the card is actually smaller than half, 1220 mm^2 versus 2670 mm^2, or about 45 percent of the useful area.

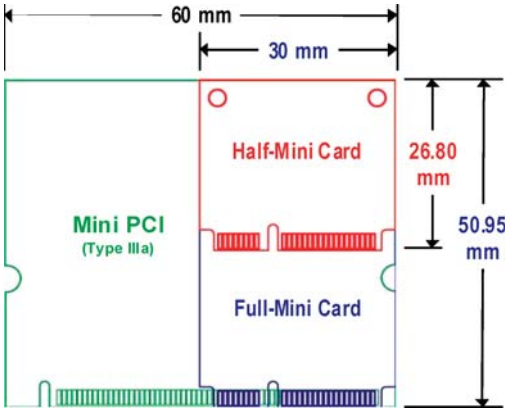

Figure 2: ***Comparing standardized card sizes.***

The most critical factors in determining the size and shape specification of the smaller Mini Card included the area and volume reduction impact for circuit components, thermal density, and cooling impacts, and considerations for supporting a socket configuration that allowed for sharing between the larger and smaller cards.

Even as many wireless circuit technologies are progressing down a size-reduction roadmap, the proposed smaller Half-Mini Card format represented a challenge for both existing and emerging technologies. Initially proposed at being just less than half the size of the existing Mini Card, at one point there was even a proposal on the table for a smaller card on which the usable circuit area volume could have been reduced to as little as 37 percent. To help resolve the debate, each WG member company was asked to perform an independent feasibility review to determine if the smaller proposed sizes would be too constraining. Intel's considerable internal review included analyses of six different wireless technologies (for LAN, WAN, PAN, and digital TV), both singularly and in some likely product combinations, and it took into account technology reduction trends over a period of many years. The result was that there was strong evidence that not all wireless applications would be relevant in the smaller card form factors. For a majority of the specification development period, the WG settled on a target of 24.6 mm in length while keeping the card width the same as that of the Full-Mini Card. In the end, and after an even more extensive detailed review by Intel, the final dimension was increased to 26.8 mm to better accommodate a wider range of applications.

By its very nature, wireless technology can generate considerable thermal dissipation, this proving to be a key technical issue that an earlier concept to integrate wireless technology within notebook lids was unable to resolve. As a general rule, as the card format is reduced in size, the thermal density of a given application, when considered over its volume, is increased, and the cooling solution becomes more important. Unfortunately, reducing the size of a given radio technology doesn't necessarily imply a reduction in thermal dissipation. However, as it turns out, the most common cooling issues with a radio solution are often localized to the area around the power amplifiers. With Half-Mini Card designs, the concentration of this dissipated heat doesn't dramatically change. The WG chose to keep the thermal dissipation allowance the same between the two sizes of cards, but the notebook system designer must be cautioned that if two Half-Mini Cards are specifically placed within the platform to fill the previous space of a Full-Mini Card, then the cooling design for that space must take into account the potential doubling of the thermal dissipation.

Finally, the last major consideration was the potential re-configurability of a Mini Card socket, especially if the selection of Mini Card options that are to be offered for a given notebook platform design will include both standardized sizes. The primary factors in managing this include the orientation and placement of the socket connector(s) and the method used for holding the installed card in place (using the defined screw holes located at the corners of the card). Figure 3 illustrates how a socket can be configured for dual-use, given a second set of hold-down positions, with these hold-down points often being implemented as a boss and screw arrangement.

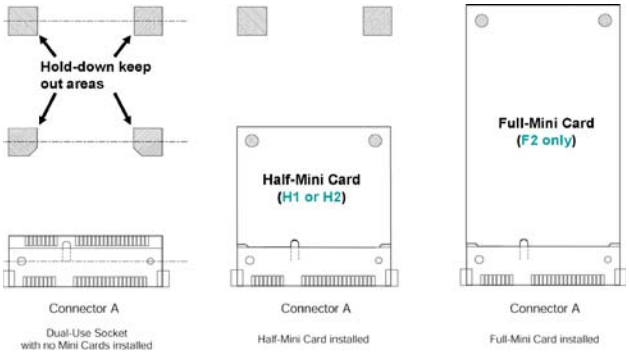

Figure 3: ***Dual-use socket concept.***

In light of this dual-use configuration and another that specified a head-to-head configuration allowing for two opposing Half-Mini Card sockets that also support substituting a single Full-Mini Card, we defined options related to the bottom side keep-out areas of the Mini Cards to promote interoperability in multi-use sockets. In Table 1 we summarize the Mini Card and multi-use socket compatibility options that were defined. Notebook OEMs are allowed to also consider other configurations including just simply isolating and positioning individual sockets in convenient locations throughout the platform.

Table 1: Multi-use socket and card interoperability.

↓ Card Type ↓		Dual-use Socket	Dual Head-to-Head Sockets	
		Socket A	Socket A	Socket B
F1	Full-Mini Card[a]	No	No	No
F2	Full-Mini Card with extra bottom-side keep-out areas	Yes	Yes	No
H1	Half-Mini Card	Yes	Yes	No
H2	Half-Mini Card with extra bottom-side keep-out areas	Yes	Yes	Yes

[a]Same as original Mini Card.

The remainder of the functional and performance specifications for both Full- and Half-Mini Cards is identical, including the support for both PCI Express and USB as the system I O interfaces, the defined wireless-specific signaling features, and the available power-delivery pins. A recent unrelated change to the Mini Card specifications restructured the power supply interface to align on two voltage sources instead of three and to allocate more pins to power delivery as a means to reduce voltage drop across the interface. All of the changes that we discuss are normatively covered by Revision 1.2 (dated October 27, 2007) of the specification [1].

Intel's role in all of the Mini Card standardization efforts to date has been unique in that we are the only participating technology supplier delivering at both the notebook system chipset and the wireless communications levels. As such, Intel has been able to supply a broad range of technical expertise to review and guide the specification development activities. As the technical editor for PCI-SIG Mini WG, we have also been able to play a leadership role in establishing useful specification requirements across a diverse set of industry participants including three major notebook OEMs, a number of wireless technology suppliers, and a number of connector suppliers.

INTEL® PRO/WIRELESS 5000

Wi-Fi solutions

The Intel PRO/Wireless series 5000 of network adaptors targets both premium and value-market segments. The premium device called 5300 is a full IEEE 802.11n MIMO three-transmit and three-receive (also known as 3×3) chains, dual band (2.4 GHz and 5–6 GHz) Wi-Fi solution. This enables the user to achieve up to 450 Mbps over the air data throughput, using standard communication protocols in both the Up Link (UL) and Down Link (DL) directions. This MIMO 3×3 scheme generally improves the data throughput vs. distance performance in a multi-path environment typical of indoor wireless connectivity, as expected in a premium device.

The value network adaptor device called 5100 is a scaled-down version of the 5300 premium network adaptor. It offers a MIMO 1×2 scheme (one transmit and two receive chains) and also supports dual band. This MIMO configuration offers a data throughput of up to 300 Mbps in the DL direction and up to 150 Mbps in the UP direction. This coincides with typical usage models in which we usually want to receive more than we actually want to send.

For the purposes of this paper, we concentrate on the Intel PRO/Wireless 5300 device, mainly because this was the more challenging of the two. However, our discussion is also applicable to the 5100 device in a more limited capacity.

In Figure 4 we show a general block diagram of the Intel PRO/Wireless 5300 Wi-Fi 3×3 solution. The main building blocks incorporated in the solution are these:

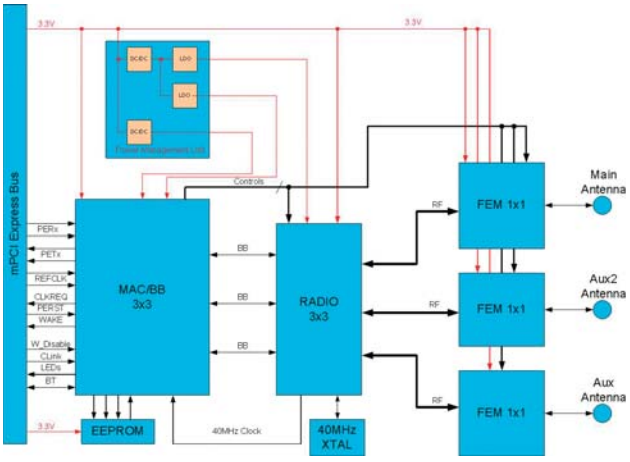

Figure 4: ***Intel PRO Wireless 5300 block diagram.***

- The Media Access Controller and Base Band chip (also known as MAC/BB)
- The Radio Transceiver chip (also known as Radio)
- The Front End Module (FEM)
- The Power Management Unit (PMU)
- The EEPROM
- Xtal

The MAC BB chip serves as the Host Interface that is the main connection to the rest of the CMT platform. It is also directly connected to the Radio Transceiver. The Radio Transceiver contains three Radio Frequency (RF) chains, each containing transmit and receive circuitry that supports both unlicensed Wi-Fi communications bands of 2.4 GHz and 5–6 GHz. The Radio Transceiver in turn is connected to three front-end circuits that are used to amplify and filter the RF signals connected to the antenna ports. The PMU supplies all the necessary bias voltages used in the system that are not directly received from the platform power source. The EEPROM device is used to store some of the key board-specific information such as MAC Address, Regulatory Parameters, and Calibration Tables that are programmed into the device during production. The Xtal is connected to an internal Crystal Oscillator (XO) circuit that generates the required clock and signal reference in the system.

TECHNICAL CHALLENGES AND SOLUTIONS

Shrinking the full-mini card down to a half-mini card

The Intel PRO/Wireless 5300 family of network adaptors is targeted to be offered in two form-factor flavors: (1) the Full-Mini Card and (2) the Half-Mini Card. The Full-Mini Card solution was defined as a single-sided solution with all components on the top side of the PCB. The intent is to support OEM customers who want to make use of the space underneath the Mini Card. With the available room on the top side of the Mini Card, we typically subdivide it into two distinct sections: (1) the high frequency RF section including the Radio Transceiver chip and all the front-end components that reside under an EMI RFI shielded enclosure, and (2) the rest of the circuitry that is not as EMI/RFI sensitive and can sit on the board unshielded.

To meet the required functionality of the Intel PRO/Wireless 5300 Network Adaptor, almost all the available area of the Full-Mini Card is populated with hardware components. Therefore, it is fairly obvious that in order to fit on the Half-Mini Card with the same hardware content that is half the board size, some of the components will need to reside on the bottom side of the Half-Mini Card board. We can look at this as if we are taking the Mini Card board and folding it on itself to create a board that is half the length but is now two sided and has components on the top and bottom of the board, as shown in Figure 5.

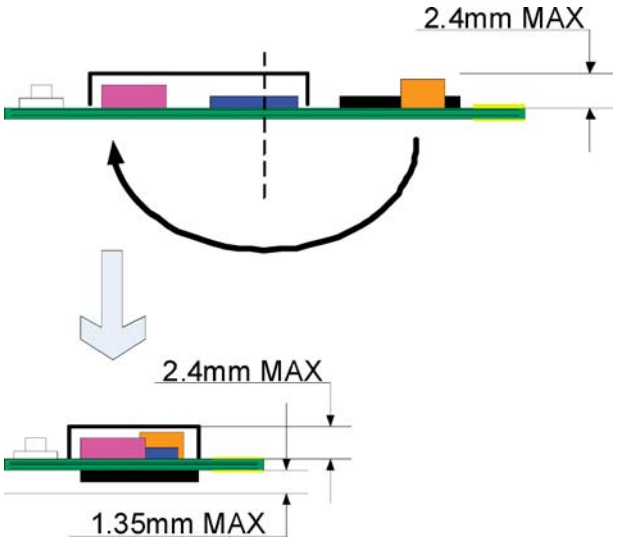

Figure 5: ***Folding the single-sided Full-Mini Card to yield a dual-sided Half-Mini Card.***

The actual partitioning of what needs to reside on the top side and what can sit on the bottom side is mainly driven by the z-height limitations of the components and the z-height restriction per the Mini Card specification. The maximum z-height above the board is 2.40 mm, while the maximum z-height below the board is 1.35 mm. These restrictions remain the same for the Full-Mini Card and the Half-Mini Card even though we used only the top side for our previous-generation solutions. In our case of transitioning to the Half-Mini Card, the RF shielded portion is more suited to sit on the top side of the board, where the extra height is needed, driving almost all other low-profile components to the bottom.

However, based on the Half-Mini Card mechanical specification, the area on the top side that is available to be shielded is actually further limited in comparison to the Mini Card board. This is due mainly to the large mounting hole and antenna interface section at the end of the board that remain the same for both form factors. This section does not scale in size when the board is shortened to half the length. It is only shifted! So, effectively, we have less room for the RF section under the shield of a Half-Mini Card. Knowing this

fact a priori drove the development of highly integrated front-end modules to save space and enable all of the RF section, including the Radio Transceiver, to fit inside the shield of the Half-Mini Card board.

With the RF section on top, we are forced to push the rest of the components to the bottom side. This means that they need to comply with the low-profile requirements to ensure that the board complies with the z-height restrictions of the Mini Card specification. Several non-compliant components were identified; specifically the power inductors used in conjunction with the DC DC converters. This fact spurred a search for low-profile substitutes. However, the low-profile substitutes found were three times more expensive than the original part. The team was asked to find some way to place the original component on the top side as part of a cost-saving opportunity. Initially, a solution was proposed whereby the inductors reside on top outside the shielded area because it was feared that the switching noise generated in these inductors would somehow contaminate the RF signals. However, we found that the overhead of this type of solution took up too much of the precious board space needed for the RF section under the shield. We devised a simple Design of Experiment (DoE) whereby we designed a board with the power inductors inside the shielded enclosure. We tested the DoE and proved beyond a shadow of a doubt that noise contamination was not an issue. This became the Plan Of Record, and the original high z-height/lower-cost inductors are incorporated within the shielded enclosure area, leaving enough area for the RF components.

We still had more challenges to overcome, however, and these are described in the next sections.

REQUIRED PCB TECHNOLOGY

The PCB technology typically used for Mini Card designs is a lower-cost, standard through-hole via (THV) design, comparable to an industry-standard IPC Type 3 PCB. This was possible for two main reasons. The first is that there are no components on the bottom side of the Mini Card designs for the THVs to come in contact with. The second is that we were able to design the component packaging to not require any high-density interconnect (HDI) PCB technology. When we transitioned to a Half-Mini Card we were very limited on PCB area for components, so we had to utilize the bottom side of the PCB for components. The density of the components on both sides left us no room to place the THVs we had used on the Mini Card designs. This forced us to transition to HDI PCB technology. We needed a way to connect the I O between components without taking up the valuable component real-estate on the outer layers with THVs. Rather than using THV technology to solve the real-estate issues, we used HDI PCB technology in the form of a microvia and buried via technology. Microvias are about half the diameter of the previous THV technology. We could also utilize them in component pads to eliminate any real-estate lost from I O connection and routing on the outer layers. This is comparable to the industry standard IPC Type 4 and Type II PCB [2,3]. The two types of PCB structures are featured in Figure 6, where the differences can be clearly seen.

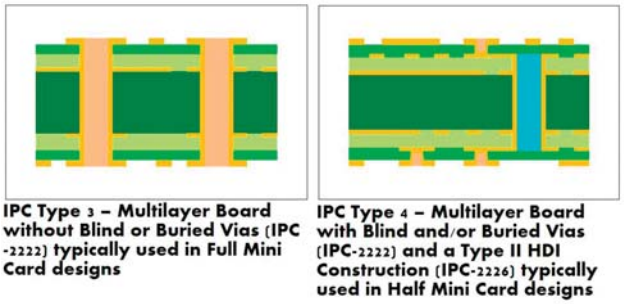

Figure 6: ***The PCB structure change between Full-Mini Card and Half-Mini Card.***

Though the new Half-Mini Card PCB size was almost half the size of our previous Mini-Card designs, our overall PCB cost was increased by the PCB technology transition. The cost increase came from several areas. The layer count was increased: extra PCB processing time was required for the sequential lamination process for the microvia and buried vias. There was an increase in added drilling steps from one THV mechanical drill step to four drill steps, that is, two laser drills and two mechanical drills. All these changes to the processing of the PCB for Half-Mini Card outweighed the savings of the smaller PCB size. However, we did not change the component packaging technology to utilize the HDI PCB technology. We did this so that our Mini Card and Half-Mini Card design could use the same components. This made it more challenging for the component packaging development, but it kept our cost down for the Mini Card PCB version by not having to transition it to HDI PCB technology.

SILICON PARTITIONING AND CONNECTIVITY ON THE PCB

One of the key challenges the team faced was the fact that the same MAC BB chip would need to reside on either the top side of the Mini Card form factor or on the bottom side of the Half-Mini Card form factor. However, the pinout could be optimized with direct routing and connectivity for only one of the cases. This means that when the MAC BB is placed on either side

of the board, the pinout interface ordering is either in line or reversed in order. This reversal of pin alignment adds to the routing complexity and introduces adverse coupling and signal integrity effects due to trace crossovers not encountered in the case of direct routing. This is also true of the signals that go between the MAC BB chip and the Radio Transceiver chip for the same reasons. The dilemma we faced was which version should we optimize.

In the end, the PCB constraint and capabilities finally drove a decision with regard to the MAC BB pinout scheme:

(a) PCB cost is a function of board size and board complexity. Since the Mini Card is the larger of the two boards, we can only reduce the complexity to maintain low cost. So a THV with a minimum number of layers is used for routing. This can be achieved only with direct point-to-point routing.

(b) The Half-Mini Card solution requires component placement on both sides of the board, driving us to a more complex HDI PCB technology that can also support more complex component routing as needed.

Based on these reasons, the team decided to optimize the MAC BB pinout for the Full-Mini Card arrangement.

During the actual layout and routing of the MAC BB signals in the Half-Mini Card HDI PCB, many routing tricks were used to take advantage of the more complex board technology, including large indirect loop routing to circumvent and avoid other traces, layer play, and in some cases component rotations. These enabled us to route critical signals without any crossovers or adverse coupling effects. In this way, we avoided degraded signal integrity, something that is especially critical for the PCI Express Host Interface signals and all of the Analog Base Band signals. All of these are also differential pairs that require special care.

Front end module (FEM) definition and board placement

The Network Adaptor chipset transceiver needs to be complemented by an external set of RF front-end components that are located directly between the Radio Transceiver chip and the antenna connection. This includes a pair of Power Amplifier (PAs), a pair of Low Noise Amplifiers (LNAs), a Diplexer, a set of Baluns (BALanced UNbalanced transformers), and a pair of Transmit Receive Switches.

Due to the fact that the required front-end content is fairly large, and our network adaptor supports a MIMO 3×3 system that needs three such front ends, we needed three highly integrated FEMs. The FEM size definition was driven mainly by the limited space available under the shield of a Half-Mini Card board and the required RF content within. The FEM pinout was synchronized with the Radio Transceiver pinout for direct pin-to-pin routing.

What is noticeable about FEMs is the fact that in the Full-Mini Card configuration, all three FEMs can sit snugly side by side and connect directly to their respective antenna and Radio Transceiver interfaces. This can be seen by the many traces running on the top layer of the PCB. However, in the Half-Mini Card scenario, the FEM locations had to be moved and rotated to either side of the Radio Transceiver. This burden made it extremely difficult to route the RF traces on the top side, and many of the RF traces needed to be embedded into inner layers. Special care needed to be taken to maintain trace impedances without sacrificing performance, something that was accomplished through careful play with the trace widths and the layer stack-up of the PCB.

We defined the actual FEM sizes through an iterative process. First, the Computer Aided Design Manufacturing Engineering team at Intel examined the board area available under the shield, taking into consideration industry assembly design rules and the system connectivity requirements. The FEM estimated sizes were then presented to multiple FEM vendors for evaluation. The FEM vendors were also given the FEM content requirements in the form of a specification document. After trimming down the content to exclude a pair of Baluns, the vendors confirmed that they could 'fit' the necessary content to within the target size of 6x4 mm. The removal of the pair of Baluns was also acceptable by the Radio Transceiver design team. Finally, the FEM pinout was also defined with the aid of the vendors, taking into consideration their implementation requirements and our RF interface requirements with the Radio Transceiver chip.

Thus, with careful iterative planning, we were able to fit all the RF content within the smaller shielded area of the Half-Mini Card.

Regulatory emissions concerns and challenges

Even with all our careful planning and design, there was no guarantee that we would meet all emission requirements. The transition to Half-Mini Card with dual-sided assembly added complexity to the PCB routing and required a change to the metallization layer stack-up. Both of these changes can introduce

new potential sources of unwanted emissions from the board, but they also can bring new opportunities to overcome emission and other performance issues.

Through a combination of good design practices and drawing on our previous RF experience with former generations of PRO/Wireless solutions, the hardware design team identified three main items that require special care during the board layout design. Careful attention was given to proper grounding and the use of microvias (uVias) and power traces. These practices have a huge impact on overall performance and on the ability of our product to meet regulatory certification that enables worldwide use.

1. Grounds (GND)

With the increased number of layers on the Half-Mini Card, we decided to also increase the number of ground layers in the PCB stack-up. This was done for several reasons:

- GND pour and routing is critical to ensure expected performance of RF components as tested and approved in a standalone environment. Poor grounding might cause limited performance of key radio components such as the FEM. It might adversely affect output power, EVM level (a standard signal quality factor measurement unit used in phase and amplitude modulation systems), and cause spurious and harmonic emissions.
- Multiple ground layers enabled us to easily support different impedances (100 ohms, 50 ohms) of RF strip-lines while maintaining reasonable line widths. Homogenous and uninterrupted ground planes must surround and follow all RF and analog signal lines.
- Power lines and traces can be accompanied by a good ground path plane. This can be further improved if wide traces are sandwiched between two ground layers. The capacitance to GND increases and thus enables the removal of discrete high-frequency capacitors from the board. This not only frees up precious board space but can also save cost.
- IR drop on GND planes is minimized by multiple GND layers.
- Isolation of noisy sensitive control lines can be improved by routing them in internal layers next to a GND layer thereby enabling an uninterrupted ground return path.

2. Use of microvias (uVias)

To enable good grounding as mentioned above as well as optimized signal routing with the shortest possible lines, it is essential to make use of uVias. Unlike THVs, uVias have an added advantage because of their small size: they occupy less board area, and they can be located within component pads for additional board-area savings. In general, a greater number of uVias can be spread all over the board to provide shorter paths to GND. Because of their small size, they can also be placed strategically at trace ends and corners. This minimizes the generation of stub-like lines of GND and other traces that can act as unwanted antennas generating unwanted emissions from the board. When routing power lines and traces, multiple uVias are needed to reduce IR (voltage) drops along the trace. Again, because of the uVias' small size, many can be placed within wide DC traces.

3. Power lines

Power lines and traces should be made as wide as possible in order to minimize IR drops and make use of multiple uVias between layers, especially those with high current loading. The wide lines/traces routed between two GND layers will also provide some high-frequency capacitance and minimize the need for many small value capacitors on the board for RF decoupling.

By incorporating these relatively simple Best Known Methods into the Half-Mini Card layout design, we were able to overcome issues such as degraded EVM performance, unwanted harmonics, and spurious radiation from the board. In some cases, multiple variants of the board layouts were in parallel in order to overcome these issues. In one such case, by strategically placing two additional uVias along a power trace leading to the FEMs, we improved the EVM performance and were able to remove a discrete decoupling capacitor from the board.

Thermal considerations

The Half-Mini Card specification calls for the same power-handling capability as that of the Full-Mini Card. However, the shrinking of the form factor from Mini Card to Half-Mini Card also introduced a new issue in the form of thermal power density. Basically, we are trying to dissipate the same amount of power that was previously dissipated on a Full-Mini Card in half the volume. In other words, the power dissipation is more concentrated. We feared that this would cause the Tjunction of the MAC BB and Radio Transceiver silicon die to increase beyond the maximum acceptable temperature levels required to maintain performance and reliability.

In order to ensure that we do not exceed the max Tjunction of the die, we conducted a series of thermal simulations taking into account a typical notebook

environment with our Half-Mini Card mounted on the bottom of the notebook motherboard. These simulations incorporated multiple variables that included the following:

- Package type
- Relative location of the key power dissipaters on the board
- Number of metal layers in the PCB and metallization thicknesses
- Actual power dissipation in each key component

The fact that the MAC BB and Radio Transceiver chips were designed for Wire Bond (WB) connectivity opened up an opportunity to examine various types of packages and combinations. We examined package types such as Ball Grid Array (BGA), Quad-Flat-No Lead (QFN), and others. Of these types, the QFN-type package offers the best theoretical power-dissipation capability. It has a large die paddle in the middle of the package on which the die is mounted, and it serves as a direct thermal conduit to the board. However, it should be noted that the number of interfaces with a QFN-type package is limited to the number of pads along the perimeter of the package. A Dual-Row QFN package offers more pads: two rings of pads along the perimeter. However, it also drives a larger package to house the same-size die. We had to look carefully at the tradeoff between thermal behavior and the number of interfaces.

Another parameter that plays a role in the thermal behavior is that we now have a two-sided board and need to add more metal layers to the PCB to accommodate the necessary side-to-side isolation and the added routing complexity. This increase in the number of layers (increase is in pairs to maintain symmetry) has a significant positive effect on the thermal behavior, and the added metallization layers actually enable us to dissipate more heat from the components on the board.

The relative location of the components on the board is quite limited in our case. The requirement to have all RF components on the top side under the shield meant that all other main power-dissipating components needed to reside on the bottom. This includes the MAC BB chip and the PMU. In light of the fact that there isn't a lot of room for the components to move around, only a few options need to be examined. The EEPROM has insignificant power dissipation; therefore, it could reside anywhere without really influencing the thermal behavior. Due to the space constrictions, however, it was placed on the bottom side.

The actual power-dissipation numbers selected to be used for the thermal simulation were a challenge in themselves. Under normal operating conditions, the system works in a dynamic mode based on the actual communication protocol. This means that the Network Adaptor sometimes transmits, sometimes receives, and sometimes is idle. The combination of these will yield a different average thermal behavior for different modes of communication. The worst-case scenario was identified as a MIMO 3×3 Transmit mode, which can occur for a duty cycle of greater than 97 percent when User Datagram Protocol (UDP) is used for high-throughput data communication. In this case, all the transmit chains are active and working almost all the time. The power amplifiers in the FEM are large power consumers. However, when in MIMO 3×3 mode, we are actually able to reduce the transmit power level of each FEM to a third of the maximum transmit level, approximately 5 dB lower, and this yields a collective transmit power of all three chains that will maintain the same total radiated power. This also means that each FEM will dissipate less power in this mode.

With all the various options described above, multiple simulations were conducted to examine the Tjunction of the chipset components. An example of such a simulation environment is shown in Figure 7. The result of this examination basically drove the following design guidelines:

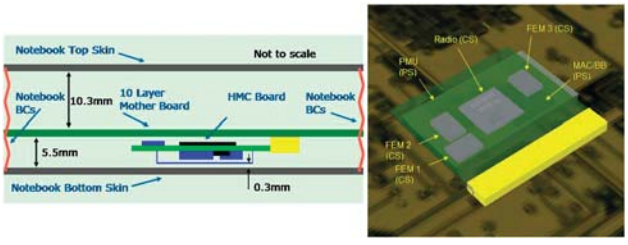

Figure 7: ***Thermal simulation in a notebook environment.***

- Use QFN packages.
- Increase the number of PCB layers.
- Reduce the individual Tx power on each chain when working in MIMO modes.

Using these guidelines, the Intel PRO/Wireless 5300 Network Adaptor was designed and tested. Actual results correlate with the simulations and show that Tjunction does not exceed maximum Tjunction for the silicon when mounted on a Half-Mini Card. We also carried out the simulations and tests on the Full-Mini Card. The Full-Mini Card has the huge advantage of having a much larger PCB to dissipate the heat

generated by the components, enabling the implementation of a solution with a reduced number of PCB layers. The simulations clearly show that there are no issues whatsoever with the regular Mini Card product skew from a thermal point of view.

ACTUAL IMPLEMENTATION OF INTEL PRO/WIRELESS 5300

The saying "a picture is worth a thousand words" is certainly applicable in this case. Figure 8 shows the actual implementations of the same Intel PRO/Wireless 5300 Network Adaptor Mini Card and Half-Mini Card. It clearly shows how the component partitioning was done keeping the RF shielded components on the top side (component side) and putting low-profile components on the bottom side (print side). The shield size reduction and tight fit of all the RF components under the shielded area shows the necessity to integrate the FEMs. It can also be seen that the tall power inductors (bottom right-hand corner inside the shielded area) are on the top side to enable the lower-cost option. Looking at the MAC BB chip, it is clear that the direct and simplistic routing to the Radio Transceiver chip and to the Host Interface gold fingers connections are maintained in the Full-Mini Card version to enable lower-cost PCB technology. However, this direct and simplistic routing cannot be maintained in the Half-Mini Card version of the board. Nonetheless, the devices were successfully routed using more complex routing in multiple inner layers.

Figure 8: ***Intel® PRO Wireless 5300 network adaptor Full-Mini Card and Half-Mini Card implementations.***

CONCLUSION

The trend to make things smaller with more functionality continues to be a key driving force for Intel PRO/Wireless solutions. As shown in this paper, the technology challenges of moving to a half Mini Card form factor have been met and or exceeded.

This quantum leap, when compared with the Wi-Fi solution offered in the original CMT platforms of 2004, has significantly changed the platform content and capability. The new board area, which is less than one quarter of the original board size and greater than six times in functionality (two times the bands supported and three times the number of transmit and receive chains), proves that such silicon and board challenges can be met.

This shrinkage of the Intel PRO/Wireless 5000 Wi-Fi 802.11n solution down to a Half-Mini Card form factor while increasing functionality has enabled our OEM customers to incorporate multiple add-in cards and increase wireless functionalities inside their latest Montevina CMT platforms. The OEMs can now offer higher-performing and smaller platform solutions compared with the previous-generation Santa Rosa CMT platforms.

Undoubtedly, the Half-Mini Card will become the industry standard form factor going forward. We expect the competition to also align with this trend. Although it is unclear how much our competitors will be able to integrate their solutions into this new form factor, Intel has taken it upon itself to be amongst the first to drive and adopt this new form factor standard from concept to product. The introduction of the 5300 series of Intel PRO/Wireless solutions, which is a full Dual-Band 3×3 MIMO IEEE 802.11n solution using this new Half-Mini Card form factor as part of the Montevina CMT platforms, is a testament to the Intel "Leap Ahead" corporate motto.

ACKNOWLEDGEMENTS

The work described in the paper was made possible by the contributions, guidance, and comments of several people. We acknowledge Jacob Solomon, Navtej Singh, and the reviewers.

A special thanks to all the teams and team members that spent many long hours to make the launch of the Intel Centrino Duo mobile technology PRO/Wireless 5000 family of network connection solutions a reality.

REFERENCES

[1] "PCI Express* Card Mini Card Electromechanical Specification."Revision 1.2. PCI-SIG 2007.

[2] "Sectional Design Standard for High Density Interconnect (HDI) Printed Boards."IPC-2226, April 2003.

[3] "Sectional Design Standard for Rigid Organic Printed Boards."ANSI/IPC-2222, February 1998.

Glossary

MC	Mini Card, also referred to as Full-Mini Card
HMC	Half-Mini Card
CMT	Centrino™ Mobile Technology
WLAN	Wireless Local Area Network
MAC/BB	Media Access Controller/Base Band
MIMO	Multiple In Multiple Out
Mbps	megabits per second
PCI SIG	Peripheral Component Interconnect Special Interest Group
PCB	Printed Circuit Board
CS	Component Side (usually referred as to the top)
PS	Print Side (usually referred to as the bottom)
BGA	Ball Grid Array
QFN	Quad, Flat, No-lead
WB	Wire Bond
HDI	High Density Interconnect
THV	Through Hole Via
UVia	Microvia
RF	Radio Frequency
RFIC	Radio Frequency Integrated Circuit
FEM	Front End Module
PA	Power Amplifier
LNA	Low Noise Amplifier
PMU	Power Management Unit
DC/DC	Direct Current to Direct Current
EEPROM	Electrically Erasable Programmable Read Only Memory
IPC	Printed Circuit Standard body
EMI/RFI	Electro Magnetic Interference/Radio Frequency Interference
IR	Current x Resistance (voltage)
GND	Ground
EVM	Error Vector Magnitude
UL	Up Link
DL	Down Link
I/O	Input/Output
DB	decibel (a relative unit of measure)
Tj	Temperature @ die junction
WG	Work Group
BTO	Built To Order
LCD	Liquid Crystal Display
DoE	Design of Experiment
OEM	Original Equipment Manufacturer
USB	Universal Serial Bus
Xtal	Crystal
UDP	User Datagram Protocol

AUTHORS' BIOGRAPHIES

Eli Laks is a Senior HW RF Design Engineer in the Mobile Wireless Group. Currently he is working on Intel® Pro Wireless Multi-Comm solutions for Notebook platforms. He specializes in RF board design, general hardware design, and regulatory certification validation. In 2005 he joined Intel as a Senior HW RF Design Engineer. During 1980–2004 he was employed on and off by MicroKim Ltd. as the R&D Manager (CTO). He developed multi-function hybrid RF synthesizers and small RF systems. In 1991 Eli founded a start-up company called Eilon Engineering that developed weighing and force measurement systems. Eli received his B.Sc.E.E. degree from the Technion, Israel in 1985. His e-mail is eli.laks at intel.com.

Richard S. Perry is a Manufacturing Architect in the Mobile Wireless Group. He specializes in the areas of PCB technology, FEM substrate and package design, Si package design, and halogen-free product design for next-generation Intel® PRO Wireless solutions. In January 2000 he joined Intel Corporation focusing on the wireless technology development for PCB design and PCB assembly. From 1994 to 2000 he was employed with Ericsson, Cellular Phones Division, as a Senior Manufacturing Engineer and an Operations Engineering Manager for digital phone manufacturing. Richard received a B.S.E.E. degree from Clemson University in 1996. His e-mail is richard.s.perry at intel.com.

Brad Saunders is a Senior Mobile Systems Architect focusing on platform I O technology definition in Intel Corporation's Mobility Group. Brad also leads the PCMCIA industry group responsible for the ExpressCard standard and is the technical editor for the PCI Express Mini electro-mechanical specifications within the PCI-SIG. Brad came to Intel Corporation as part of the Xircom, Inc. acquisition in early 2001, where he had spent two years in Xircom's chief technology office. Prior to that, he spent 21 years working in various fields of communications at Rockwell International, ranging from secure communications for defense applications to analog modems for mobile systems. Brad received his B.S.E.E. degree from the University of California, Irvine in 1976. His e-mail is brad.saunders at intel.com.

Ra'anan Sover is a HW RF Architect in the Mobile Wireless Group. Currently he is working on Multi-Comm solutions for both notebook and hand-held platforms. He was directly involved in the definition, implementation, integration, and productization of Intel® Pro Wireless Wi-Fi products. In May 2000 he joined Intel as a Senior Radio Frequency Integrated Circuit (RFIC) Design Engineer and co-designed the synthesizer block of Intel's first WLAN RFIC. From

1988-2000, he was employed by MicroKim Ltd. as a Senior RF Engineer. He developed multi-function hybrid RF synthesizers and control devices. Ra'anan received his B.Sc.E.E. degree from the Technion, Israel in 1988. He is a Senior Member of the IEEE. His e-mail is raanan.sover at intel.com.

Greater Mobility Through Lower Power

David W. Browning, Mobile Platform Group, Intel Corporation
Eric DiStefano, Mobile Platform Group, Intel Corporation

Index words: TDP, thermal design power, platform, chassis, cooling, industrial design

Citations for this paper: David W. Browning, Eric DiStefano "Original 45nm Intel® Core™ 2 Processor Performance" Intel Technology Journal. http: //www.intel.com/technology/itj/2008/v12i3/6-paper/1-abstract.htm (October 2008).

ABSTRACT

Mobile Original Equipment Manufacturers (OEMs) place great emphasis on creating unique system designs to differentiate themselves in the mobile market. Intel Corporation's introduction of low-power, high-performance Intel® processors based on original 45nm Intel® Core™ microarchitecture, originally referred to by the codename Penryn, brings a new level of opportunity for differentiation into the mainstream segment without sacrificing performance. The Mobile Penryn family of processors offer mainstream performance at 25 watts (W) Thermal Design Power (TDP), 10 W less than previous-generation processors. This paper focuses on those attributes of design that are enhanced by the Penryn family of processors: thickness, temperature, and noise level. These processors allow thinner, cooler, and quieter systems, and they offer the end-user a more satisfying mobile computing experience.

In this paper we discuss the fundamentals of system cooling capabilities in any given form factor and look at how power relates to the thinner, cooler, and quieter systems. In the end, both Intel and OEMs 'win' when Intel provides the options to allow OEMs to enhance their overall system designs without sacrificing performance and thereby enhance their brand in more systems.

INTRODUCTION

In the extremely diverse world of mobile platform design, the focus of Original Equipment Manufacturers (OEMs) is on "look & feel" (commonly referred to as Industrial Design). Mobile OEMs place great emphasis on creating unique system designs to differentiate themselves in the mobile market [1,2]. Differentiation results in systems that come in all shapes and sizes, or form-factors, as well as systems that emphasize different aspects of mobility such as size, weight, and features. This differentiation makes the specific industrial design of many notebook computers unique to their designer, and that uniqueness becomes associated with their brand. Therefore, components that make it easier for OEMs to implement unique designs allows them to achieve their brand goals.

Additionally, in order to build on the brand equity of a particular industrial design, OEMs frequently maintain the same mechanical design "skin" of the previous platform for two to three generations. This implementation choice also forces them to maintain the same thermal design characteristics as the previous design. Therefore, there is little room for an OEM to innovate when no other parameter is changed.

Although industrial design is a predominant OEM consideration, another major consideration is ergonomics. Ergonomics includes, but is not limited to, user touch-temperatures and audible system noise levels. An uncomfortable touch-temperature (also called chassis or "skin" temperature) or exhaust temperature, or an annoying system noise level distracts from the user experience, even for the sleekest systems. Given the highly integrated nature of notebook system designs, often the vectors of performance, noise, and comfort can be divergent, and each may constrain the other as OEMs seek to innovate and differentiate. OEMs spend substantial design effort to balance performance, chassis temperatures, and or quiet systems, as well as differentiating along one or more of these vectors.

Intel's introduction of lower-power, high-performance Intel® processors based on original 45nm Intel Core™ microarchitecture, originally referred to by the codename Penryn, to the mainstream mobile market brings a new level of opportunity for differentiation. The mainstream Penryn mobile processor draws 25 watts (W), 10 W less than its 35-W predecessor. In this paper we focus on those attributes of the design that are enhanced by lower power Penryn family mobile

processors: (1) thickness, (2) temperature, and (3) noise level. We concentrate on the mainstream Penryn family of mobile processors and do not describe the benefits of utilizing the even lower power, small-form-factor family of Penryn mobile processors.

In the following sections, we show form-factor trends and summarize fundamental limits of cooling under ergonomic constraints and other boundary conditions. We then relate these limits to form factor, noise, system temperatures, opportunities for feature additions, and design flexibility.

Current Platform Thermal Challenges

Mobile OEMs design notebook cooling solutions with challenging form-factor limits, component temperature limits, and the primary ergonomic boundary conditions of noise level and chassis "skin" temperatures. A representative system layout is illustrated in Figure 1.

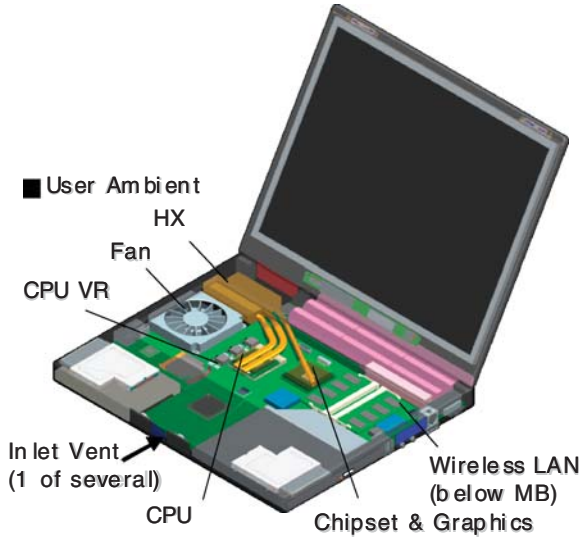

Figure 1: ***Representative notebook iso-view with transparent chassis.***

The numerous subsystems and components compete with each other for motherboard and notebook perimeter real estate for connectors and user interfaces. The result is little unused space, leaving marginal room for adequate air flow. As a consequence, systems are highly integrated mechanically and thermally; they are interdependent for balancing the limited cooling available. The major cooling components, the heat pipe, fan, and heat exchanger, are often referred to as the thermal solution. However, in reality, the solution is the total system design itself, as careful consideration is given to the placement of the air inlet and exhaust vents, the placement of components in primary air flow paths, air flow paths themselves, the fan and heat exchanger, as well as conduction interfaces such as spreaders and even insulators. Change in any of these parameters impacts the cooling of components and the system as a whole, which can make or break a successful design.

Within this competitive landscape, the higher power components vie for good air flow paths, and in the best case, a direct attach to the active solution (the fan and heat sink assembly). Since these good air flow paths and the placement options of the active solution are limited, the higher the power of the component, the more restrictions it places on the system design, whether in the location of the device itself, or in limiting the ability to attend to other components.

Thermodynamic limits

In any given system there is a maximum sustainable level of cooling that is established by the thermodynamic limit. This limit includes theoretical passive limits and the capacity of the available air flow to absorb heat between the ambient air (input) temperature and any given maximum allowable exhaust (output) temperature.

Most of the powered components reside in the base of the system, and the base represents the primary cooling challenge in mobile systems. An energy balance about the base of the system in steady-state conditions is

$$P_{\text{base}} = Q_{\text{passive}} + Q_{\text{active}} \tag{1}$$

where P_{base} is the allowable total power of components in the base of the system; that is, excluding power to the display. Q_{passive} is the combined passive heat transfer mechanism for energy dissipation; that is, radiation and natural convection, represented by a form of Fourier's Law.

$$Q_{\text{passive}} = hA(T_{\text{chassis}} - T_{\text{ambient}}) \tag{1}$$

h is a combined effective heat transfer coefficient including natural convection and an approximation of radiation; a typical value may be 8 W/m^2-K. Considering Equation 2, once the system form factor, and thus A (area), is fixed, and once the ambient temperature and the allowable maximum skin cooling temperature (an ergonomic limit) are defined, the passive cooling is established.

Q_{active} is the active cooling level, and its maximum capability is determined by the thermal capacity of the air flowing through the system as in

$$Q_{\text{active}} = \rho \dot{V} C_p (T_{\text{exhaust}} - T_{\text{ambient}}) \tag{3}$$

Equation 3 illustrates how, in a given ambient temperature, and a maximum allowable exhaust air temperature, the maximum active cooling capability is defined wholly by the amount of air that can be passed

through the system. The relation is linear. Figure 2 shows this relationship with the thermodynamic limit for a 1.1″-thick, 14″-display system with nominal boundary conditions of 35°C ambient, an average chassis temperature of 50°C, and assuming a maximum exhaust air temperature of 70°C.

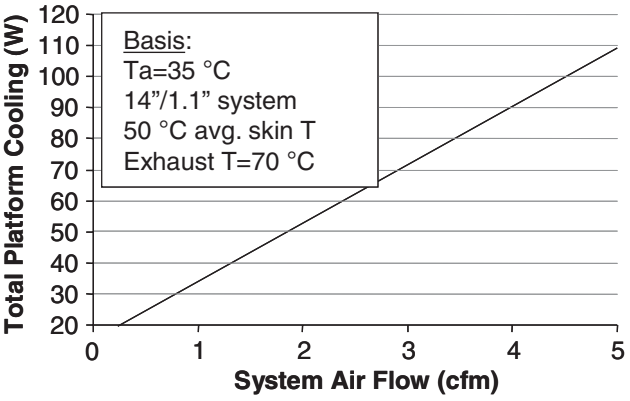

Figure 2: ***Maximum sustainable system cooling—thermodynamic limit.***

Air flow then is the most critical determinant of the total cooling capability of the system; it carries the primary burden for cooling the system. How much air flow can be passed through the system, of course, relies more on the performance of the fan. This capability is primarily characterized by the allowable space for the fan and air flow paths around the fan. There is limited area on the motherboard as well as limited available chassis perimeter space for the fan exhaust, so the only remaining major parameter to examine is the available internal height for the fan and air flow around it.

Thermal stack-up

In mobile systems, the amount of internal height available for the fan and air flow approaching it is limited to, and roughly characterized by, the vertical distance z-height between the bottom of the keyboard and the inside bottom chassis surface. This space is a fraction of the total stack-up as shown in Figure 3. For example, in a typical 14"-display and a 1.1″-thick system, the internal z-height available is 15 mm, which is broken down roughly to 12 mm for the fan itself and 3 mm for air flow paths about the inlet(s). For a system of this size, a typical maximum air flow rate may be 3.3 cfm (1.6e-3 m^3/s); of course, actual air flow depends on the specific design, particularly on system flow paths and corresponding flow resistances.

Several of the system stack-up components at any given time are virtually fixed, such as the display, the keyboard, the motherboard, and the chassis skin. Consequently, if an approximate 10-percent or 0.1" (2.5 mm) reduction in total system thickness is desired, it is effectively a reduction in the vertical space available for the fan and air flow paths at the inlet, or approximately 20 percent for the fan itself (2.5 mm out of 12 mm) assuming the space for the air flow inlet is retained.

Thermal technology limits

Cooling technologies are focused on increasing efficiency in component cooling interfaces and on air flow, whether it's the fan itself or some other means. These technologies have resulted in roughly only a 5-percent additional total system cooling capability year-over-year. Moreover, this progression is prone to tapering off over time with diminishing returns on incremental improvements. Therefore, cooling technology alone cannot keep up with a desire for progressively thinner systems. The power of the components themselves must therefore be reduced to realize thinner systems.

INTEL PROCESSORS BASED ON ORIGINAL 45NM INTEL CORE™ MICROARCHITECTURE

With the introduction of the Intel Penryn processors, Intel Corporation is providing a solution to realize

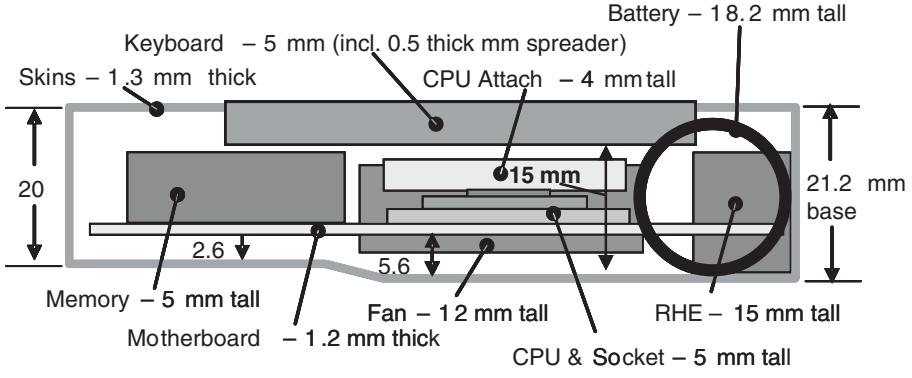

Figure 3: ***Side-view cutaway of vertical stack-up (base only) for a nominal 1.1″ system.***

thinner, cooler, and quieter systems. The Mobile Penryn family of processors provides an option for a full-performance component, but with a 10-W reduction in TDP from 35 W to 25 W. These parts were optimized to 25W TDP in order to address the increasing focus on thermally constrained notebook designs.

Application to mobile systems

These new processors also bring a new level of opportunity for differentiation in the mainstream mobile market segment without sacrificing performance. The 10-W reduction in TDP that they provide as an option can be utilized at the system design level in one of four different ways:

- Thinner systems
- Cooler systems
- Quieter systems
- More feature-rich systems

Thinner systems

One of the primary directions in the mainstream mobile market is to build thinner systems. Figure 4 illustrates this trend of notebook system thickness over time. As is shown with the dotted line in Figure 4, there is a somewhat linear relationship to this decrease in system thickness over time, and OEMs continue to look for opportunities to achieve thinner systems.

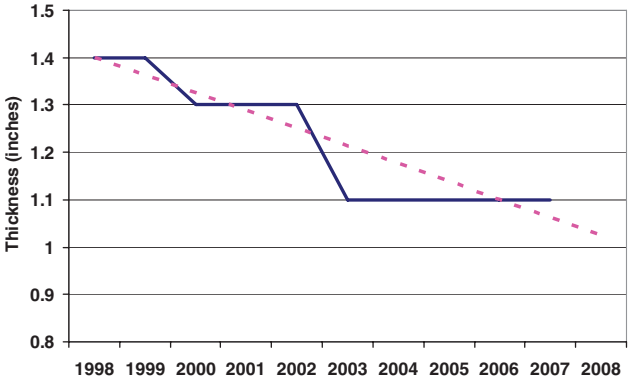

Figure 4: ***Notebook system thickness variation over time.***

However, as stated earlier, the thermodynamic limit for a particular system design is dependent on the air that flows through the system. In a horizontal fan configuration, also called blower fans, the volumetric air flow rate at a given speed is roughly proportional to the z-height of the fan and or blades.

$$\dot{V} \propto z \quad \text{or} \quad V_2 = \frac{z_2}{z_1} V_1 \tag{4}$$

This is an approximation, but it generally holds true for the horizontal fans in mobile systems. For a low-flow loss system, this applies roughly to the air flow rate associated with the operating point in the system. Therefore a reduction in available z-height results in a proportional reduction in air flow through the system.

In the example noted earlier, a 0.1″ (2.5 mm) reduction in total system thickness from 1.1″ to 1″ applies directly to the internal height available for the fan. Thus, a 2.5-mm reduction in the available fan height of a 12 mm thick fan, results in a 20-percent reduction in air flow. For the nominal 3.3 cfm of the original system airflow, this translates to 0.6 cfm. From Equation 2, this translates into approximately a 10 W effective system cooling capability.

Conversely then, if the required platform (base) power is reduced by 10 W, then the opportunity arises to make that system 0.1″ (2.5 mm) thinner.

The new lower-power Penryn family of mobile processors provides the OEM the ability to achieve this amount of reduction in z-height by providing an option for a TDP of 25 W, reduced from 35 W. This allows the mainstream market to more confidently target a new notebook thickness design point of 1″ (14″ display).

Cooler systems

An alternate use of the headroom provided by the new lower-power Penryn family of mobile processors is to make an existing industrial design cooler, whether by prescription or user choice. To evaluate the temperature reduction of a 10-W reduction in power provided by the Penryn family of mobile processors, we perform numerical simulations on the same nominal 14″ display and 1.1″-thick system. The design features are summarized in Table 1, along with typical component TDP design powers, where TDP is representative of the highest power an individual component can go to under any realistic condition.

The simulation is performed using a Flotherm* model of the system. The grid varies in resolution in z and x-y, according to the proximity of component edges and air gaps. Most components are modeled as blocks with simple resistance planes between component and motherboard. Substantial validation has been performed on such models at Intel to establish a reasonable degree of confidence in the results for purposes of comparison.

When assessing how to design a cooler system, a stacked TDP methodology is not employed; rather, a system design power scenario is applied where the concurrent powers of each component are utilized.

For purposes of comparison, consider a usage scenario in which the user is multitasking, stressing the processor and the platform as a whole with various concurrent activity (applications). The powers that we assume for such a scenario are summarized in Table 1. To evaluate the effect of the 10 W power reduction of the processor, the processor power is simply reduced from 35 W to 25 W, keeping all other TDPs at the same level, as denoted in Case 1 and 2, respectively.

Table 1: System features and power scenarios.

Concept	TDP (W)	Case 1 power (W)	Case 2 power (W)
Processor	35	35	25
Chipset & Graphics	10.5	9.5	9.5
I/O controller Hub	2.5	2.1	2.1
Memory	5.6	5	5
Non voltile memory	0.6	0.6	0.6
Wired network, LAN	0.9	0.1	0.1
Wireless network, WLAN	1.8	1.4	1.4
Hard disk drive	4	22	22
Optical disk drive	5.5	29	29
Battery (self-heating)	1.5	1.5	1.5
Voltage regulator CPU (85%eff.)	6.2	6.2	6.2
System VR (87%eff.)	4.8	4.8	4.8
Rest of base power	4.5	4.4	4.4
CPU	35	35	25
Non-CPU	48.4	40.7	40.7
Platform total	83.4	75.7	65.7

Temperature results for the bottom surface are shown qualitatively in Figure 5, which illustrates a substantial reduction in warm area. Keep in mind that the scenario assumed is relatively stressful, and thus, warm. Some specific bottom surface (“skin”) temperatures are summarized in Table 2.

From Table 2, a 10 W reduction in system power translates to approximately a 2°C reduction in skin temperature in the vicinity of the processor and its voltage regulator, and a 5.6°C reduction in exhaust air temperatures. Although these differences may not seem large, they can mean the difference between uncomfortable and unacceptable and make or break a design. Internal testing of various systems yielded similar results to the simulation.

Table 2: Simulation results—selected bottom surface (“skin”) and exhaust air temperatures.

Location	Case 1 (°C)	Case 2 (°C)
Bottom Skin, Processor	59	57
Bottom Skin, Processor VR	60	58
Bottom Skin, Chipset	44	43
Bottom Skin, Memory	53	53
Exhaust Ait	63	58.4

Quieter systems

Another use of the additional thermal headroom provided by the 25-W version of the Penryn mobile processor is to allow for reduced maximum noise levels in a system.

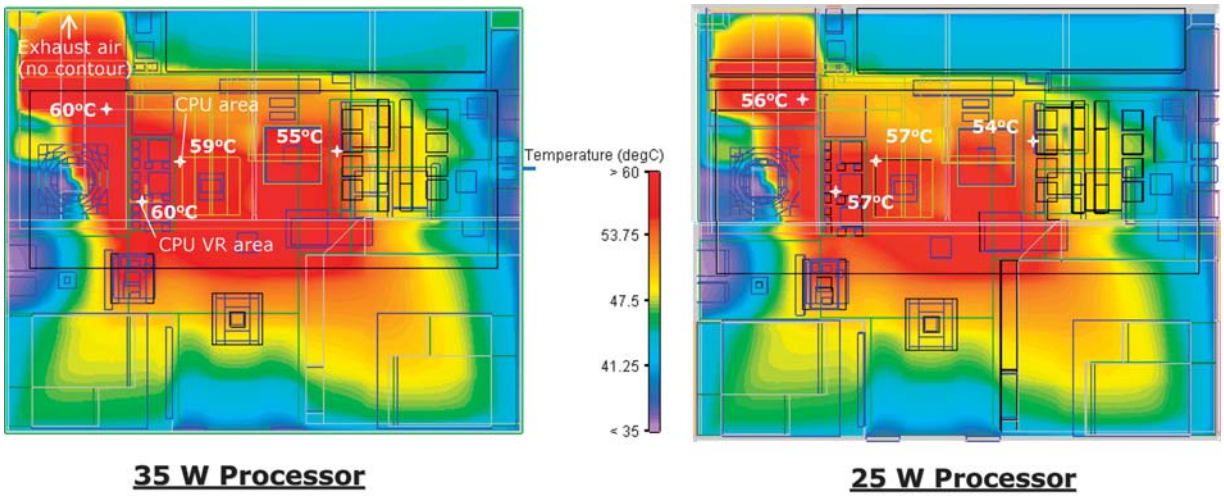

Figure 5: ***Simulation results–bottom surface (“skin”) temperature contours.***

Fans are usually turned on or their speed is increased when they are needed to support a high-power application or scenario (alternatively, to reduce the temperatures as discussed above). To evaluate the impact of the 10 W reduction in processor power, consider again the nominal system design in Figure 1, where you have a 14''-display and a 1.1″ system thickness, with a corresponding maximum air flow rate of 3.3 cfm in a user ambient temperature of 35°C and a maximum exhaust air temperature of 70°C.

Intel has internally tested many systems and the noise level associated with their flow rate under standard fan component test conditions [3]. Different flow rates are set by changing fan speed via the fan voltage. The resulting sensitivity of noise measured by sound pressure level is shown in Figure 6. A single fan result is shown by the dashed line in Figure 6. The boxed region indicates a range of system performance, roughly parallel to the single fan result shown. Figure 6 further shows that a 1-cfm reduction in flow rate results in approximately a 10-dBA reduction in noise level.

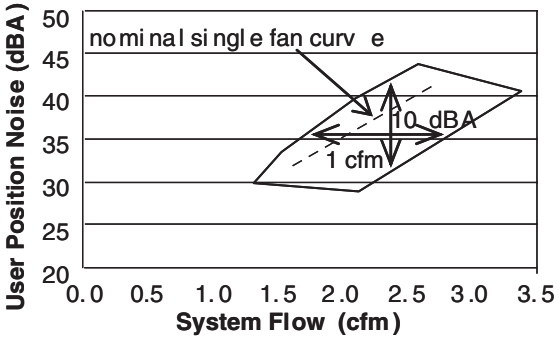

Figure 6: ***Fan noise level sensitivity to flow-rate.***

Using Equation 3 for active cooling, we calculate that the air flow rate associated with the 10 W power reduction of the Penryn mobile processor is approximately 0.6 cfm. Therefore, using the ~10 dBA cfm result from above, a 10 W reduction in power results in roughly a 6-dBA reduction in fan noise level. This is a noticeable reduction in noise to the user position, equivalent to moving the user position twice as far from the noise source [4]. Again, the actual noise level differences will depend upon the specific system design and the usage condition being considered.

More feature-rich systems

One interesting potential use of the TDP headroom provided by the new lower-power mobile Penryn family of processors is to allow other features in the platform to use this cooling capability. Some recipients of this power headroom could be better graphics performance, additional mini-card support (for more wireless cards), higher/hotter system memory capabilities, etc. None of these expanded features could previously be attained within a predefined system design without one of the boundary conditions changing (that is, making the system larger or thicker).

The reduction in the use of power in this processor allows the remaining 10 W to be applied to these other features. For example, current notebook system designs devote approximately 5 W to cooling the system memory. This power limit of system memory prevents mobile platforms from being able to use the fastest (and therefore highest-temperature) memory technology. This is one of the reasons why mobile platforms do not support the same maximum system memory frequencies as desktop platforms. Applying the TDP delta to system memory would allow the OEM to add more features to the notebook system and help bridge some of the feature differences between desktop and mobile platforms.

Platforms based on those ingredients. Each of these opportunities further enhances the user experience of mobile computing and allows the mobile OEMs the ability to provide further innovations to enhance their own brand equity and thereby create a 'win' for both Intel and the OEM.

SUMMARY AND CONCLUSION

The introduction of lower-power, high-performance Intel processors based on the original 45nm Intel Core microarchitecture provides mobile OEMs with an excellent opportunity to create the next great mobile platform design.

Although the results presented in this paper are specific to the conditions analyzed herein and will change based on the specific system design, the opportunities presented for enhancing the mobile platform user experience are compelling. Only time will tell which of the presented opportunities was most highly valued and therefore most utilized. However, one thing is certain: all of these opportunities are directly tied to the Intel processors based on original 45nm Intel Core microarchitecture and the Intel Centrino™ Mobile Platforms based on those ingredients. Each of these opportunities further enhances the user experience of mobile computing and allows the mobile OEMs the ability to provide further innovations to enhance their own brand equity and thereby create a 'win' for both Intel and the OEM.

ACKNOWLEDGEMENTS

We thank the following people for assistance in developing the underlying work behind this paper:

Bijendra Singh and Sridhar Machiroutu for thermal simulation and guidance, Eric Baugh for acoustics data and guidance, and Jarvis Leung, Michael Bitan, Ali Muhtaroglu, and Tawfik Arabi for low-power product implementation and leadership.

REFERENCES

[1] Wildstron, Stephen H. "Could a Notebook Be Best for Your Desk?" Business Week, Online magazine, July 8, 2002, at http://www.businessweek.com/magazine/content/02_27/b3790050.htm.

[2] Morrow, Bill. How Did The Thinkpad Get Its Name?" Online publication, at http://www.thinkpads.com/Genesis%209.htm.

[3] ECMA-74. Measurement of Airborne Noise Emitted by Information Technology and Telecommunications Equipment. 9th edition. Ecma International, Geneva, Switzerland, December 2005.

[4] Wolfe, Joe. "What is a Decibel?" Online publication, at http://www.phys.unsw.edu.au/jw/dB.html.

APPENDIX: NOMENCLATURE

ρ	density of air, kg/m^3
A	effective heat transfer surface area, m^2
C_p	heat capacity (of air), J/kg-K
h	heat transfer coefficient, W/m^2-K
P	power (or cooling), W
Q	heat, W
T	temperature, °C
T_{chasis}	chassis surface ("skin") temperature, °C
$T_{ambient}$	user ambient air temperature, °C
$T_{exhaust}$	system exhaust air temperature, °C
$\dot{V}$	volumetric air flow rate (of air), m^3/s

AUTHORS' BIOGRAPHIES

David W. Browning is a Senior Staff Engineer in the Mobile Platform Group at Intel Corporation with a focus on advanced mobile platform architecture and system design integration. He has been working for Intel for ten years. David has a B.S. degree in Electrical Engineering from the University of Washington. His e-mail is david.w.browning at intel.com.

Eric DiStefano is a Principal Engineer in the Mobile Platform Group at Intel Corporation with a focus on advanced cooling and thermal design. He has been at Intel for ten years researching how much future products can be cooled, cooling technology development, and facilitating customer use of Intel products. Prior to working for Intel, he was with the aerospace industry in areas of fluid and thermal systems for spacecraft and launch vehicles. His e-mail is eric.distefano at intel.com.

THIS PAGE INTENTIONALLY LEFT BLANK

Power Improvements on 2008 Desktop Platforms

Paul Zagacki, Client Platform Architecture and Planning, Intel Corporation
Vidoot Ponnala, Client Platform Architecture and Planning, Intel Corporation

Index words: low power, platform power management, energy efficiency, desktop power, desktop platform

Citations for this paper, Paul Zagacki, Vidoot Ponnala "Original 45nm Intel® Core™ 2 Processor Performance" Intel Technology Journal. http://www.intel.com/technology/itj/2008/v12i3/7-paper/1-abstract.htm (October 2008).

ABSTRACT

Idle and low-utilization platform power management have been key deliverables for mobile platforms for many years. The resulting platforms based on Intel® Centrino® technology have delivered increasing performance and capabilities while continuing to increase the overall battery life of the mobile platform. These same principles have gained importance also in desktop platforms as corporations strive to reduce the cost of deploying platforms in support of their efforts to address global climate change and deliver more energy-efficient computing. Further, the Energy Star specification for computers was adopted on July 20, 2007 [1], adding idle power targets to desktop platforms. The Energy Policy Act, adopted by Congress on July 27, 2007 [2] now requires that federal agencies buy equipment that is Energy Star qualified (effectively making Energy-Star compliance mandatory for some percentage of desktop platforms).

In this paper we present platform- and silicon-component power data that demonstrate advances in desktop platform power management enabled on 2008 platforms, built with the Intel Q45 Express Chipset, originally referred to by the codename "Eaglelake," and Intel processors based on the original 45nm Intel® Core™ 2 Quad microarchitecture, originally referred to by the codename "Yorkfield/Wolfdale." For instance, a 2008 desktop platform with more advanced power management can demonstrate (when correctly configured) a 16-percent reduction in AC idle power when compared with the same platform enabled with 2007 platform power-management techniques. In addition to providing an overview of how technologies such as deeper C-states or Serial ATA link power management (familiar in mobile platforms) dramatically reduce silicon and platform power, we also demonstrate the effectiveness of these technologies in desktop platforms, from operating-system power-management settings to devices such as USB keyboard mouse or multimedia card readers. Finally, we make recommendations on how to configure a desktop platform (hardware and software) to make the most of the new features found on our 2008 platforms.

INTRODUCTION

Desktop platforms are experiencing an evolution in what is seen as their primary value to consumers. Until recently, the quest has been exclusively directed at achieving higher and higher workload performance at lower and lower cost to the end user. However, driven by concerns over the environment and greenhouse gas emissions and the increasing importance of Energy Star compliance [1,2], the focus is shifting from the exclusive pursuit of performance and cost improvements to energy-efficient performance (EEP) [3,4]. EEP is the intersection of performance or capabilities with the delivery of those capabilities, using the least amount of energy. In simple terms, for many workloads, this can mean getting the job done as fast as possible and then getting the system into an idle or low-power state. Most of a system's time over the course of a day is spent at or near idle utilization and power levels, as demonstrated in Figure 1.

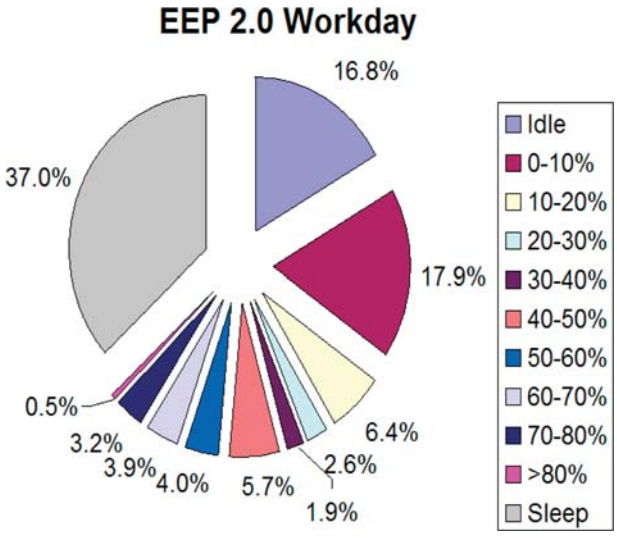

Figure 1: ***System power levels during a 9-hour workday following the EEP 2.0 Model demonstrate that 75 percent of time is at or very near idle utilization.***

The EEP 2.0 workday, the source of the data in Figure 1, assumes several hours of work represented by runs of the productivity benchmark Sysmark 2007 (which has idle time built into it), followed by idle in the form of an employee break, and by some sleep time on longer breaks [3,4].

Considering all the time the system is less than 10 percent used in any given workday, and that the system is likely to be idle or asleep overnight (unless it is in a batch environment), it is important to focus our attention on idle and low-utilization power improvements. The fact that most desktop platforms are lightly loaded most of the time coupled with the Energy Star Version 4.0 [5] focus on idle power is key to understanding the motivation behind many of the platform power-management features implemented on our 2008 desktop platforms. Idle is defined as the average state of the platform after the operating system (OS) has loaded and the system has been given adequate time to quiesce any activities (15 min in the Energy Star Version 4 specification). Without getting too deep into the topics of ACPI and OS power management, readers are encouraged to reference the ACPI 3.0a specification and the Windows* power-management whitepaper for more background information on idle and processor power states [6]. These are shown in Figure 2.

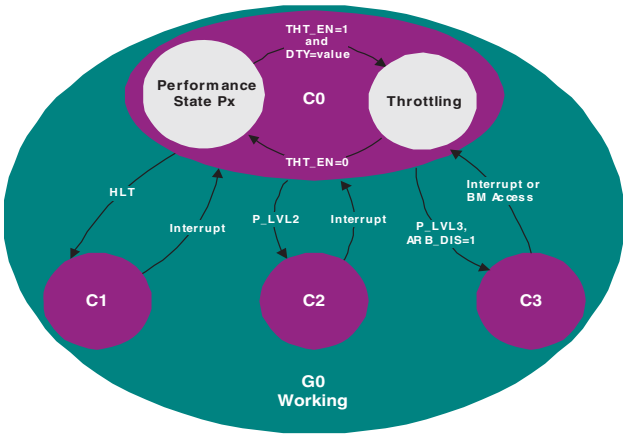

Figure 2: ***Processor/platform power states from the ACPI 3.0a specification.***

In order to better understand what techniques are required to deliver the most energy-efficient desktop platforms, it is important to first understand where all the power goes in a typically configured desktop platform. Throughout, we focus our attention on a typically configured Intel® vPro™ desktop platform with integrated graphics. (Figure 3 shows a typical platform hierarchy with the key components.)

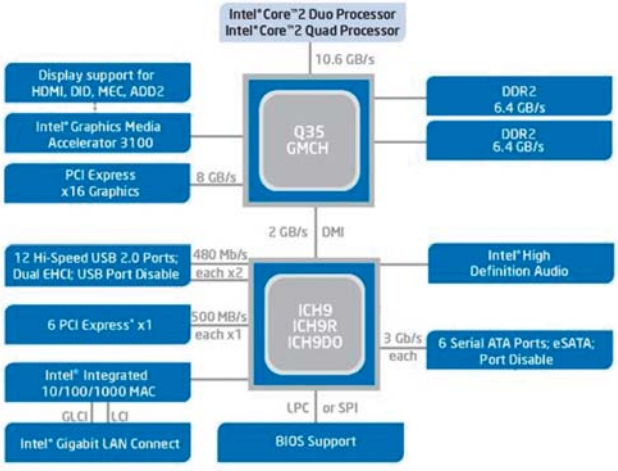

Figure 3: ***Key components in a 2007 corporate platform built with Q35 Express Chipset GMCH and ICH9.***

We focus on integrated graphics primarily because discrete graphics cards cover a very wide range of power, and the performance extracted from this additional power is not typically necessary for office- or productivity-type applications such as Outlook*, PowerPoint* or Excel* (while energy efficiency is quickly becoming a key attribute for these platforms).

All of the components and their interfaces shown in Figure 3 are powered either directly from the silver box power supply (as is the case for peripherals such as serial advanced technology attachment, or SATA, drives) or through voltage regulators (VRs) built into the desktop motherboard. For example, in the current platform generation, the processor has at least three individual power rails (core, front-side bus (FSB), and phase locked loop (PLL) supplies); the graphics and memory controller hub (GMCH) may have up to five supply voltages (many of which are shared across various regulators); the I/O controller hub (ICH) has seven individual rails; and dual data rate (DDR) memory requires three rails. All of these component supplies are derived from either a dedicated 12-V processor rail or from the 12-V/3.3-V/5-V rails out of the silver box power supply utilizing about fourteen separate regulators (more if the platform supports the manageability engine, a key component of Active Management Technology). With all of this voltage regulation on a desktop motherboard, it is clear that power delivery is one of the biggest sources of efficiency losses in a platform. Some components may have three voltage regulation steps from AC to final DC supply, prior to the voltage being seen by the silicon. Each such stage loses a little power in the translation.

The pie chart in Figure 4 demonstrates the areas that we need to focus on to maximize energy efficiency in desktop platforms. Power delivery conversion losses, major platform silicon components (processor, GMCH, and ICH silicon), and SATA hard drives are among the primary power consumers in a desktop platform.

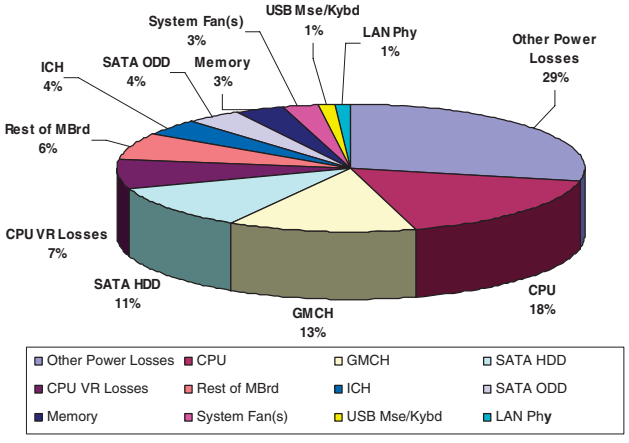

Figure 4: ***Approximate component power consumption (including losses) for a 45-W AC idle platform.***

Throughout the remainder of this paper we describe several power innovations on the corporate desktop platform based on the Intel Q45 Express Chipset and the latest Intel processors based on original Intel Core™ 2 Quad, 45nm microarchitecture. We explore innovations in efficiency through the use of phase shedding on the processor VR. We also look at the impact on idle and active power from deploying deeper C-states than those deployed in previous-generation Intel Q35 chipset-based platforms. We demonstrate some common issues that USB devices present for platform power management and explore solutions and best known practices already in use on mobile platforms. Further, we explore how to get the most out of these improvements with the right OS configuration settings.

ARCHITECTURE

For the purposes of our research for this paper, we used a single Intel vPro desktop platform with a 45nm Intel Core 2 Quad (Yorkfield) processor with a 1333 FSB at 2.83 MHz, a Q45 GMCH with DDR3, and the ICH-10 I O controller hub. This platform is very similar to that shown for the previous-generation platform in Figure 3. We utilized knobs built into a debug BIOS to allow us to enable the platform power features for the previous-generation platform (Intel Q35) and then enabled all the features that are new to the Intel Q45 Express Chipset platform. Using a single platform to create a baseline for the previous generation and then demonstrating the improvements in power management on the Intel Q45 Express Chipset platform helps to remove variation in all the other components between the systems.

DESKTOP POWER INSTRUMENTED REFERENCE PLATFORM

Many of the power measurements referenced throughout the remainder of this paper were taken on a specially designed and instrumented Intel vPro desktop reference validation platform (RVP). As mentioned earlier, the processor has three power rails, the GMCH has five rails, the ICH has seven rails, and the system memory has three individual power rails. All of these individual voltage rails are instrumented with low-loss, high-accuracy sense resistors and voltage sensing points. All this instrumentation is fed into a high-speed data acquisition card for real-time display and offline analysis. This gives us visibility into the DC power required by each of the major components on the platform along with information on the current state of the platform. The C-state information (refer to Figure 2) is also measured by logging the coordination

signals used between the processor, GMCH, and ICH to set up entry and exit points from these C-states.

The AC power consumed by the platform is measured by using an AC power meter that connects to the wall power outlet at one end and to the silver box power supply at the other end. The data collection was done with a Windows Vista* SP1 OS using the balanced mode of native OS power-management techniques under various workload conditions (although we focused mostly on idle as our workload).

INTEL CORE 2 QUAD PROCESSOR FAMILY

Power features

The Yorkfield processor is the first four-core desktop processor family, based on Intel's 45nm silicon process technology, that doubles transistor density while providing major improvements in switching leakage power. Beyond the improvements in energy efficiency gained from 45nm process technology, the Yorkfield processor family supports a key additional feature beyond what was implemented in earlier steppings. Core power states below C2 (stop grant) are now supported on the 2008 desktop platform providing a significant hook for components in the platform to opportunistically manage their power.

A core power state such as deeper sleep (C4) is a platform-level decision, so we needed all the key silicon components in the 2008 platform to have access to the feature. The Yorkfield processor family initiates the C4 request for the OS's idle handler through either an I/O read to a specific location, or through an MWAIT instruction. The Q45 GMCH and ICH10 then coordinate the details for the rest of the platform, and the ICH10 asserts the appropriate signal to indicate that the platform is entering C4. The key thing to understand here is that C4 is a state that comprises a coordinated effort among all the major silicon components in the platform.

The C4 power state allows the processor to drop to a very low voltage while still maintaining all the processor's state information. This reduction in core voltage produces a dramatic reduction in transistor leakage, the primary component of idle power. From a behavioral standpoint, the primary differences between C4 and C2 are that the processor no longer responds to snoop requests in C4, because the core voltage is too low to service them and the latency to exit C4 is somewhat longer (about 60 uSeconds). If there is any activity the processor must respond to, the chipset will initiate an exit from C4, in some cases exiting to C2 to respond to a snoop request and then quickly re-entering C4.

In addition to the core voltage reduction for the processor on entry into C4, the processor also has the ability to tune the efficiency of its own power delivery through the use of a power status indicator (PSI#) signal. This signal is asserted on entry into C4 by the processor and consumed by the core voltage controller, thereby allowing the core VR to tune the power-delivery efficiency to match the processor's expected current consumption.

VOLTAGE REGULATOR DOWN 11.1—POWER STATUS INDICATOR (PSI#)

In the pie chart shown in Figure 4, a significant amount of power is shown as processor VR losses. This loss comes from converting 12 V coming out of the power supply to the voltage being requested by the processor (typically between 0.8 V and 1.2 V). A typical desktop processor VR is split into 2–4 individual phases, depending on the maximum current requirements, and it is most efficient in the middle of its supported current range (0–90 A plus in a 3-phase design). The efficiency drops off very quickly at light loads, effectively making the processor appear to draw more power than it really requires to operate in that mode.

In the 2008 desktop platform we tackle this light-load efficiency problem with an optional feature for Voltage Regulator Down (VRD) 11.1 controllers that allow the processor to give the VR an indication of the current demand it expects over a period of time. When the processor asserts PSI# on entry into C4, the VR can turn off phases of the voltage regulation to improve the power-delivery efficiency. The chart in Figure 5 shows the improvement in efficiency of a typical reference design achieved by the use of the PSI# signal. The improvement in efficiency available with a PSI#-enabled VR design translates to a 20–30 percent improvement in the processor's power consumption under idle conditions.

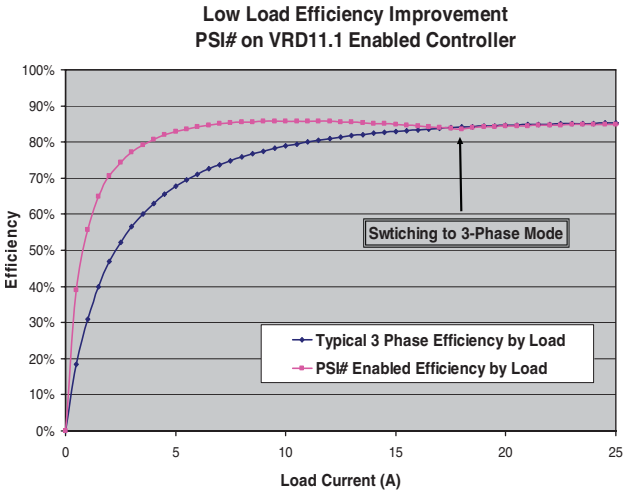

Figure 5: ***Improving power efficiency at light loads on VRD 11.1 controllers.***

INTEL® Q45 EXPRESS CHIPSET

Power optimizations

Similar to the Yorkfield processor family described above, the Intel Q45 Express Chipset family has many optimizations to enhance its power-efficient performance. The Intel Q45 Express Chipset is also on a new process technology (a 65nm derivative vs. a 90nm technology used for Intel Q33 in the 2007 platforms). This process technology allows for scaling of leakage and dynamic power through support of a lower core voltage (from a typical 1.25 V to 1.1 V), reduced gate leakage characteristics, and reduced switching capacitance on the newer process technology. The Intel Q45 Express Chipset memory controller also provides improved support for DDR3, which uses less voltage (1.5 V vs. 1.8 V) and power for similar frequency DDR2 technology.

Beyond the process technology shift for the Intel Q45 Express Chipset family and support for a more power-efficient DDR3, there are architectural innovations tied to the implementation of C4 that we consider briefly here.

First, the memory controller in Intel Q45 Express Chipset, in response to a C4 entry, takes advantage of the longer expected idle periods to put memory into a self-refresh state. On a DDR3 device—a dual in-line memory module (DIMM) consists of many devices—this can mean a third of the current consumption requirement versus the best possible device state supported on the previous-generation chipset. Considering that a DIMM can be made up of eight to sixteen individual devices, each consuming an average of approximately 34 mA per device, one third less current can mean a savings of hundreds of mWs per DIMM.

On entry into C4, the Intel Q45 Express Chipset also reduces its own idle power consumption over the previous generation on the platform by dynamically powering off memory DLL circuits, tristating memory I/O logic buffers, and powering off host, memory, and PCI Express internal clocks.

RESULTS

All the power-saving techniques described in the previous sections implemented on the 2008 desktop platform contribute to a more energy-efficient platform. Figure 1 shows that a system spends most of its time at or near idle, and power saved under these conditions is critical to achieve energy efficiency. In this section we present DC component and AC platform power data to demonstrate the power-management improvements of the 2008 platform. We also explore the impact of USB devices, such as the keyboard and mouse, on platform power management. Finally, we present some best known methods for getting the most out of the new platform power-management features.

The 2008 platform achieves significant idle and low utilization power savings from its support of deeper C-states (up to C4 state versus the 2007 desktop platform which supports only C2). The support for deeper C-states allows enabling of additional power-management techniques for major components in the desktop platform, ultimately leading to a 16-percent reduction in platform power (varies from platform to platform depending on the components in the platform) in idle and a 5-percent reduction when running Sysmark 2007.

Figure 6 illustrates the AC power savings achieved as well as the DC savings at the component level under idle conditions. Note that there is nearly a 60-percent power savings for the Yorkfield processor, mostly attributed to a dramatic reduction in core voltage when in C4. Beyond the processor itself, since C4 is a synchronized effort between major silicon components in the platform, the Intel Q45 Express Chipset achieves a 28-percent power improvement. This DC savings will vary from processor to processor and chipset to chipset due to natural distributions of static current and the way C4 voltage settings are optimized by manufacturing on a part-by-part basis. The system memory achieves a 60-percent power improvement, because the chipset puts the memory in a self-refresh state as previously explained. These component-level power savings, along with the efficiency improvements on

VRD11.1 power delivery (described earlier), add up to an approximate 16-percent power improvement at the platform level.

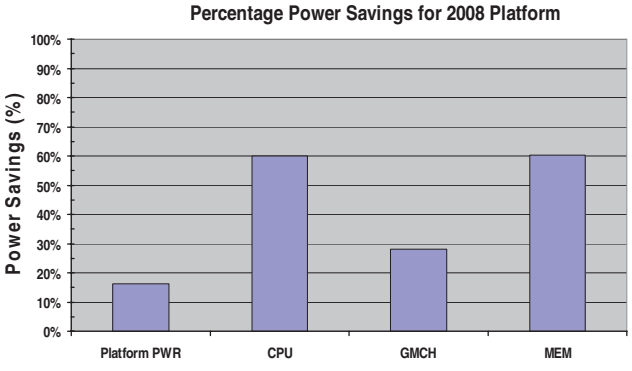

Figure 6: ***Platform and component power savings.***

The shift in C-states residency between 2007 and 2008 platforms is illustrated in Figure 7. Generally, C-state residency numbers give an indication of how much time as a percentage the platform spends in each C state. The 2007 platform has the ability to utilize only C2, and Figure 7 shows that the platform spends 99 percent of its time in C2 under idle conditions. Much of that time spent in C2 converts directly to C4 time on the 2008 platform, but not all of it. On the 2008 platform, Figure 7 shows a remainder of 11 percent in C2 and C4 state residency of 88 percent. This change in the distribution of deep C-state residency numbers results in the considerable power savings illustrated in Figure 6. A 100-percent C4 state residency is not achievable because of certain break events such as interrupts, wake up events caused by drivers, or because of the polling architecture of USB devices.

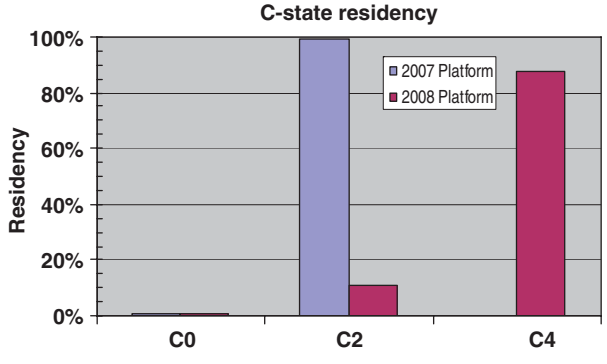

Figure 7: ***C-state residency.***

Power savings on Sysmark 2007 on a suite-by-suite basis and on average are illustrated in Figure 8, which shows a 5-percent improvement in AC platform power on average for the 2008 platform. Average here is the arithmetic mean of the power consumed by each component of the benchmark.

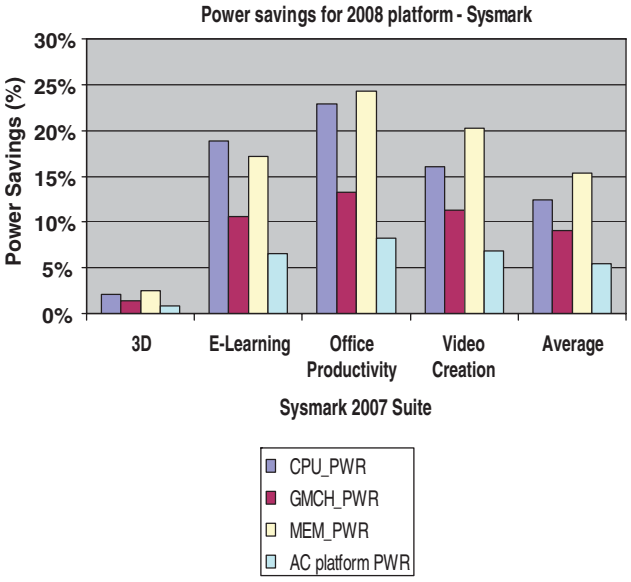

Figure 8: ***Sysmark power savings.***

USB IMPACT ON PLATFORM POWER AND C-STATE RESIDENCY

An additional power-savings opportunity can be had by enabling the USB selective suspend feature in the OS [7]. Selective suspend is a low-power mode defined in the USB 2.0 specification [8]. that allows the USB hub driver to turn off USB ports when they are in idle. Figure 9 illustrates these power savings. Observe that the Yorkfield processor shows a 76-percent power savings in these conditions (an additional 16 percent over what was shown in Figure 7). This, along with the Intel® Q45 Express Chipset's 34-percent improvement and the system memory's 73-percent power savings, help to achieve a 22-percent power improvement at the AC platform level.

This improvement in the component power is also reflected in the C-state residency data with USB selective suspend enabled, as shown in Figure 10. C4 state residency now is just over 98 percent compared to 88 percent achieved before without USB selective suspend enabled, and this translates into the power savings seen in Figure 9.

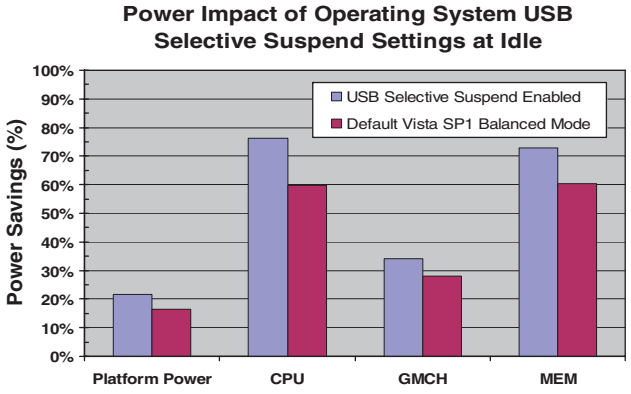

Figure 9: ***Platform and component power savings with USB selective suspend enabled.***

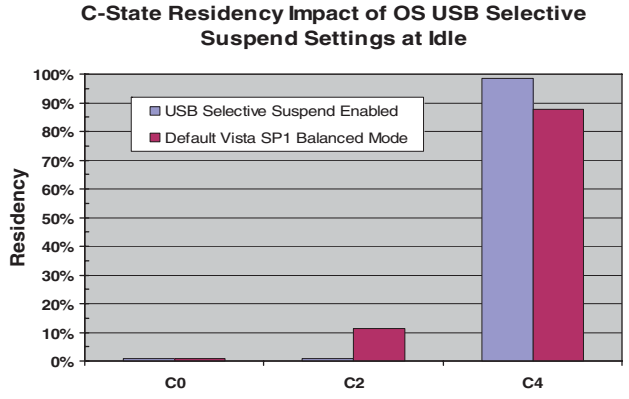

Figure 10: ***Effect of USB selective suspend on C-state residency.***

Enabling this feature in the OS also helps to increase the active power savings as illustrated in Figure 11. For Sysmark 2007 we see a 7-percent improvement in the AC platform power on average as compared to the 5-percent improvement seen earlier (refer to Figure 8).

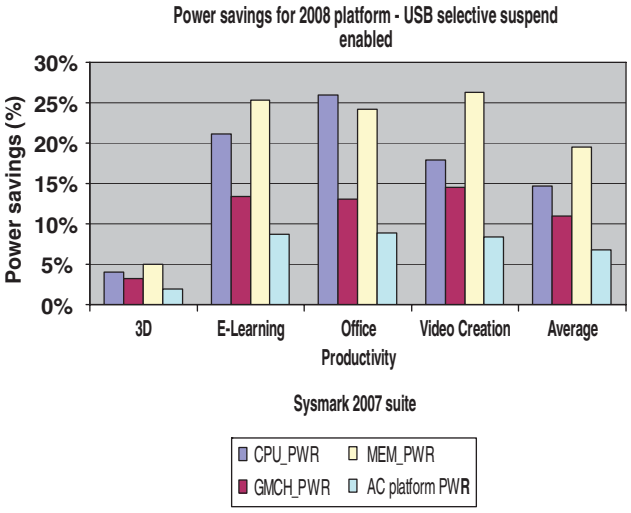

Figure 11: ***Sysmark power savings with USB selective suspend enabled.***

In addition to the USB hub device behavior, USB devices such as the keyboard and mouse have an adverse impact on the platform power management because of the polling architecture of USB devices [9]. USB devices cannot initiate a transfer or transfer data without being polled by the host first. Because of this, the processor has to constantly come out of the C4 state just to poll the USB devices and check whether they have any data to transfer, thereby incurring a power penalty, especially in idle conditions. So while a 76-percent power savings was achieved by using a PS2 keyboard mouse and by enabling USB selective suspend in the OS, only a 69-percent power savings for the processor is observed when a USB keyboard mouse are connected to the platform with USB selective suspend enabled, and similar behavior is observed for the other components as well. As a result, only a 16-percent power savings is seen for the AC platform power. Figure 12 illustrates the power savings seen when the PS2 and USB keyboard and mouse are connected to the 2008 platform.

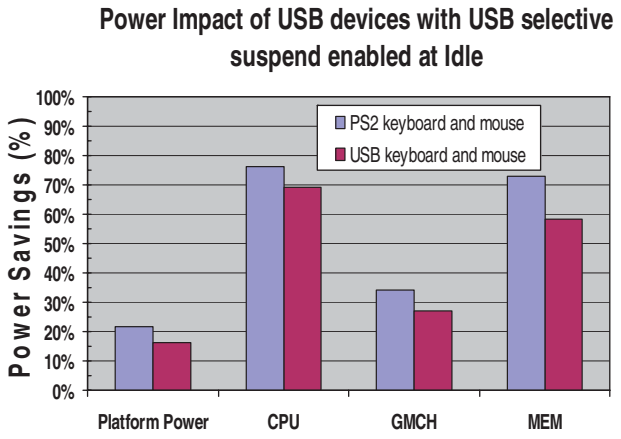

Figure 12: ***Power savings with PS2 and USB Keyboard/Mouse—USB selective suspend enabled.***

This change in power savings can also be seen from the C-state residency numbers as illustrated in Figure 13. There is clearly a change in the C-state residency when a USB keyboard and mouse are added to the system—88-percent C4 residency and 11-percent C2 residency compared to the 98.4-percent C4 residency and 0.8-percent C2 residency seen with a PS2 keyboard mouse. This large shift in the C2 state residency explains the reduced power savings seen at the platform and component level.

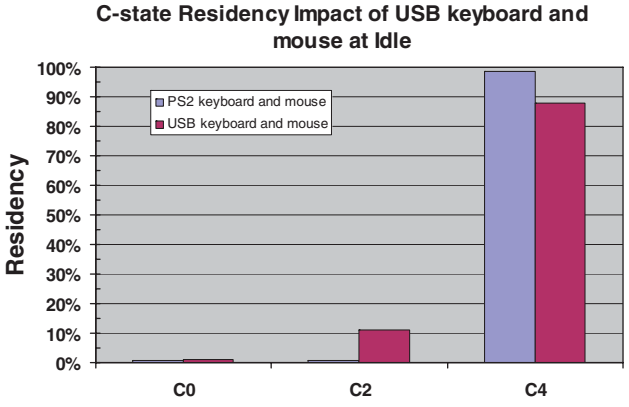

Figure 13: ***C-state residency with USB keyboard/mouse—USB selective suspend enabled.***

Figures 6 and 9 demonstrate how the 2008 platform is improved with respect to platform power management and along with Figure 12 establish that we get the most power savings when we use the 2008 platform with the USB selective suspend feature enabled in the OS and with a PS2 keyboard and mouse.

CONCLUSION

The new-generation desktop systems built with the Eaglelake chipset along with the Yorkfield/Wolfdale processors provide significant power-management techniques that help to drastically reduce power at the platform and component level. This paper has illustrated the power savings that can be achieved by using power-management features in the new platform and illustrates how users can take full advantage of the potential of the system to build more energy-efficient platforms.

ACKNOWLEDGEMENTS

We thank the following individuals for their valuable contributions: Sanjeev Jahagirdar, Intel Mobile CPU Architecture; Barnes Cooper, Intel Mobile Platform Group; Jeff Krieger and Loc Mai, Intel Platform Systems Technology Group; Mike Derr, Karthi Vadivelu, and Tuong Trieu, Intel Client Components Group.

REFERENCES

[1] http://www.energystar.gov/.

[2] http://energy.senate.gov/public/_files/ConferenceReport0.pdf.

[3] http://www.intel.com/technology/eep/index.htm?iid=tech_ee+body_eep.

[4] http://download.intel.com/technology/eep/overview-paper.pdf.

[5] http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer/Computer_Spec_Final.pdf.

[6] http://www.microsoft.com/whdc/system/pnppwr/powermgmt/ProcPowerMgmt.mspx.

[7] http://msdn.microsoft.com/en-us/library/ms793200.aspx.

[8] http://www.usb.org/developers/docs/.

[9] Barnes C. *et al.* Making USB a more Energy-Efficient Interconnect. Intel Technology Journal, Vol. 12, No. 1, 2008.

AUTHORS' BIOGRAPHIES

Paul Zagacki is a Principal Engineer in the Client Platform Architecture and Planning Group within the Digital Enterprise Group at Intel. Paul holds a B.S. degree in Computer Science from the University of Michigan. He joined Intel in 1994 and has worked on performance modeling for microprocessor architectures, software and benchmark analysis and optimiza-

tion, design convergence, and desktop component and platform power optimization. Paul has four issued patents. His e-mail is paul.zagacki at intel.com.

Vidoot R. Ponnala received his B.Tech degree in Electrical Engineering from CVR College of Engineering, Hyderabad, India and his Masters degree in Computer Engineering from the University of Wisconsin, Madison. He joined Intel as a system architect intern in Intel's Mobility Group in 2006 and currently is a Platform Power Architect in Intel's Digital Enterprise Group. Since joining this group in 2008 he has been responsible for platform power and performance analysis for client systems. His current interests include processor and systems architecture. His email is vidoot.r.ponnala at intel.com.

THIS PAGE INTENTIONALLY LEFT BLANK

The First Six-Core Intel® Xeon® Microprocessor

Shankar Sawant, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Ravishankar Kuppuswamy, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Anantha Kinnal, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Kuldeep Simha, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Ravindra Saraf, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Pradeep Kaushik, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Srikanth Balasubramaniam, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Narayanan Natarajan, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Gautam Doshi, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Sambit Sahu, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Mysore Sriram, Enterprise Microprocessor Group, DEG, Bangalore, Intel Corporation
Jeffrey Gilbert, Xeon Architect, DEG/DAP/MAP, Hillsboro, OR., Intel Corporation

Index words: multi-core, front-side-bus, energy-efficient, 3-level cache hierarchy, high-K metal gate, 45nm

ABSTRACT

This paper describes the next-generation Intel® Xeon® microprocessor designed for a broad range of highly power-efficient servers, codename Dunnington. The Dunnington processor has six cores (three core-pairs) integrated with large, dense, on-chip caches, and it delivers the dramatic power efficiency of Intel's 45nm high-K metal gate process and the Intel Core™ 2 microarchitecture to server platforms. This processor implements a high bandwidth-dedicated interface from each of the three core pairs to the last-level cache (LLC) for effective use of the inclusive LLC. With high functional integration, large cache size, and 1.9 billion transistors, the processor's moderate server-class die size of 503 mm^2 is achieved by optimizing the floor plan and physical design. This six-core product is designed to be a plug-in refresh for server platforms, codename Caneland, using Intel's DHSI-based quad-core processor, codename Tigerton. The Dunnington processor will be offered in multiple options with core counts of four or six, LLC sizes of 12 or 16 MB, several core frequencies, and Thermal Design Power (TDP) limits of 50, 65, 90, and 130 W. This processor will be the first part to employ core recovery techniques for reducing product cost. Compared to the Tigerton processor, it provides an average performance improvement of more than 25 percent. The benefits of the 45nm hi-K process and the Penryn family of processors' base is seen in the doubling of Performance/Watt and in the low TDP limits that permit six-core compute capability in the blade-server form factor.

INTRODUCTION

The high-performance expandable server processor market segment has witnessed ever-increasing demands for throughput performance and energy efficiency. To meet these demands we focused on intelligent integration of multiple cores for power-efficient parallel computing to deliver increased performance. The key high-level design requirements for the next-generation Intel® Xeon® microprocessor, codename Dunnington, Intel's latest offering in this segment, were a drop-in replacement compatibility with its predecessor, Intel's Dedicated High Speed Interconnect (DHSI)-based quad-core processor, codename Tigerton, on the Caneland platform; a 30-percent boost in performance over its predecessor; and operation in the range of 50–130 W power envelopes. In addition, maintaining compatibility between different pools of machines and stacks of software is one of the major problems in data-center and server management. Hence, enhancing the virtualization support for load-balancing across computing pools was a critical feature requirement for the Dunnington processor. This is the first Intel six-core Xeon processor, integrating a three-level, on-chip cache hierarchy and a fast DHSI system interface slated for introduction in the second half of 2008. The increased number of cores

meant a higher memory bandwidth from the DHSI interface of the Caneland platform, thereby requiring improved cache organization over its predecessor. The requirement for integrating six cores and such a large Last-Level Cache (LLC) had profound implications on the physical and electrical design of the Dunnington processor. Significant among these were die size constraints, achieving bin-split targets for the different market segments, meeting reliability constraints due to the 1.9 billion on-die transistors, dealing with process variations across the large die, and meeting thermal envelope limits. In this paper we describe each of these challenges and the solutions developed by our team to successfully bring the product to market ahead of schedule.

ARCHITECTURE FEATURES

Growing utilization of the Front-Side Bus (FSB) bandwidth required multiple architectural solutions in order to keep the soaring bandwidth-latency curve under control. Two solutions, introduced in the Caneland platform, include DHSI and snoop filters. However, to meet the required performance on the Caneland platform, the traditional multi-chip-package-based solutions for multi-core processors provided little opportunity to effectively address this challenge. The monolithic hex-core solution adopted by the Dunnington processor design allows it to tackle the bandwidth-latency challenge using an efficient cache hierarchy, a high-bandwidth on-die interface, and innovative solutions to reduce snoop traffic, thus providing a compelling Intel Xeon processor product for Q3 2008.

As illustrated in Figure 1, the Dunnington design consists of three Penryn processor family CMP core-pairs that are integrated with the LLC and the Caching Bridge Controller (CBC) using a point-to-point protocol called Simple Direct Interface (SDI). SDI provides improved latencies and bandwidth for LLC and cross-core data accesses as compared to similar accesses in its predecessor (which take place via the FSB). The CBC that stands between the Penryn cores and the LLC has three different roles to play in the Dunnington design: (a) it is a coherence and conflict resolution engine for data access from three core-pairs and the external snoops; (b) it is a cache controller for LLC and core-to-core data transfers; and (c) it is an FSB controller for the Dunnington processor.

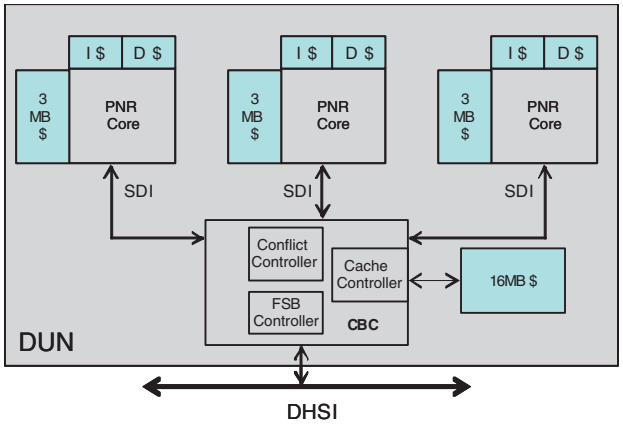

Figure 1: Next-generation Intel® Xeon® processor architecture

CACHE ORGANIZATION

The Dunnington processor has three levels of caches: 32 KB of data and 32 KBs of instruction cache in each Penryn Core (First-Level Cache or FLC), 3 MB of non-inclusive Mid-Level Cache (MLC) for each CMP core-pair, and 16 MB of inclusive LLC. Inclusivity of the LLC is maintained using core valid bits per Penryn core-pair.

The Dunnington processor implements the joint MESI states of ES, MI, and MS. (With joint MESI states, the first letter reflects the caching privileges of the LLC, while the second letter indicates the caching privileges granted to a lower level cache—in the case of Dunnington, FLC and MLC in each Penryn core-pair). The LLC holds all lines at an MESI state with equivalent or greater privilege than the FLC and the MLC have for their associated lines. The objective of joint MESI states is to optimize response time to external snoops without materially penalizing internal cache privilege transfer.

The MLC and LLC are safeguarded by Intel Cache Safe Technology. The MLC utilizes the Single Bit Fix (SBF) mechanism, while the LLC includes a cache line disable facility. These features address the cache reliability requirements of the server market segment.

COMPATIBILITY FEATURES

The processor's FSB interface is designed to be compatible with Tigerton's FSB and to operate at the highest FSB frequency (1066MT/s) in use on multi-processor platforms during the product's lifetime.

Active-way management (AWM) is another key compatibility and performance feature on the Caneland platform. AWM improves the effectiveness of the

Caneland platform's snoop filter and the efficiency of the FSB by sending way-hints to the Clarksboro chipset. These way-hints align snoop filter victimization with processor cache victimization. The near-flawless tracking of cache line allocation avoids invalidations of active lines in the LLC and the resultant increased effective memory latency.

The Dunnington processor continues to support all power-management features from its predecessor. These include P-states (P0, P1), C-states (C1E, CC3), and T-states (TM1, TM2).

VIRTUALIZATION ENHANCED

Most data centers add computational capacity over time. New generations of processors and platforms invariably offer a better price per performance and generally offer reduced operating cost per platform. Additionally, they offer reduced operating cost per unit performance. The cost efficiencies of next-generation server products means most data-center populations will have more than one processor family deployed and may also have different platforms.

There can be compatibility issues across these population factions. Typically, a next-generation processor supports incremental features to its prior-generation processor. These changes mean it becomes harder to maintain a common software stack between these factions or to move computation dynamically between platforms in the different factions. V-Migration (also known as V-Motion) is a feature that addresses these software compatibility issues by providing the ability to virtually demote a next-generation processor to function as a current-generation processor from an instruction set standpoint. It must be noted that performance and power requirements remain equivalent to the next-generation processor, providing the end-user the advantages of the next-generation technology. With this feature implemented on the Dunnington processor, the Caneland platform enables a seamless virtual machine transfer from previous generations to the Dunnington processor.

In order to provide aggressive performance per watt and to fit in various power/performance envelopes of the server segment, multiple SKUs are available by using variants of core counts, cache sizes, and core frequencies. The CBC architecture is designed to effectively adapt itself to these variations.

With these architectural innovations, the Dunnington processor delivers as much as an overall 30-percent gain on existing workloads for the same power envelopes when compared to its quad-core predecessor. Hence, it provides a significant boost to power-performance efficiency for the Caneland platform.

PROCESS TECHNOLOGY

One of the key innovations in the new 45nm process technology is the high-k + metal gate transistor, which is one of the biggest changes in transistor technology since the introduction of the polysilicon gate MOS transistor in the late 1960s. The new 45nm process technology offers about a $2\times$ improvement in transistor density, approximately 20 percent of an improvement in transistor switching speed, or more than a $5\times$ reduction in the source-drain leakage. It also provides at least a $10\times$ reduction in gate oxide leakage power and more than a 30-percent reduction in transistor switching power [1].

GLOBAL ELECTRICAL CHALLENGES

The Dunnington product is a large die that integrates six cores, a 16-MB LLC, and the Uncore logic. We found these key electrical challenges:

- Achieving the required bin-splits in the different market segments on core frequency.
- Meeting the Uncore frequency targets on the FSB-digital logic.
- Meeting Vmax reliability constraints associated with the huge number of transistors on the die.
- Meeting Vmin constraints imposed by the cache memory cells and register files on the core.
- Meeting the Thermal Design Power (TDP) for the product.

Systematic and random variations of the manufacturing process parameters introduced design constraints such as pre-silicon frequency targets on different design domains, error-correction/redundancy requirements on the cache, and compensation mechanisms on the global clocks. Identifying the optimal process targeting to achieve the SKU stackup of core count, core frequency, cache size, and TDP power requirements is a significant challenge.

In these subsequent sections we discuss how we met these challenges in the different design domains.

PHYSICAL DESIGN

The Dunnington processor die integrates three dual-cores from the Penryn family of processors, 9 MB of MLC, and 16 MB of LLC in just over 500 mm^2. The floor plan of the chip is shown in Figure 2.

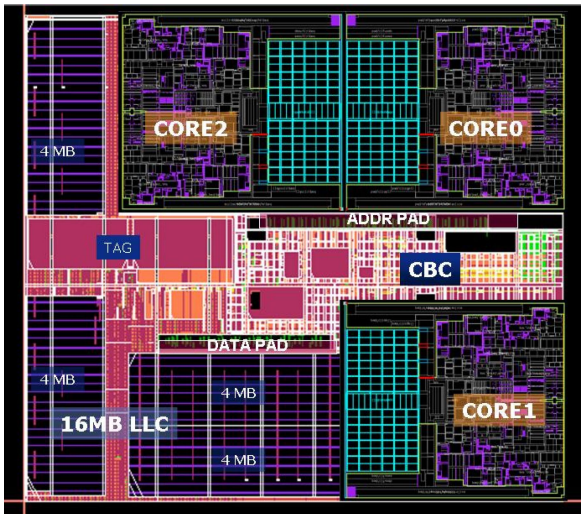

Figure 2: Six-core processor die photo

The Dunnington design presented unique physical design challenges. The first of these was fitting all of these components into a single die, subject to constraints on the maximum die size, module orientation constraints, and tight timing requirements. Die size limits prevented all three blocks from the Penryn family of processors from being placed in one line, so we had to place them in two rows as shown. I/O blocks, traditionally placed at the edges of the die, had to be accommodated in the center of the die to enable IO operation at the target FSB speed. This required us to carefully optimize the floor plan as there was heavy wiring congestion in this area. Early block size estimates and routing analyses were critical to establish the feasibility of the floor plan. Clever modular design of the LLC floor plan enabled the cache team to fit 16 MB of cache into the irregular shape that remained after the cores and IOs had been placed. This saved a significant amount of custom layout effort for the project.

Another major challenge was repeater insertion on global signals. Due to the size of the chip, virtually all global signals required repeaters to meet signal integrity constraints. We had to insert over 50,000 repeaters into the LLC and CBC, in addition to the 55,000 repeaters that already existed inside each of the three Penryn modules. We followed a "virtual repeater" methodology in order to avoid having to code these repeaters into Run-Time Library (RTL) to enable faster timing convergence. We used a fully automated virtual repeater insertion tool along with a correct-by-construction flow to convert over 700 virtual repeater "stations" into physical layout blocks after timing convergence was achieved. Having a guaranteed flow to make the virtual repeaters "real" without requiring manual layout cleanup enabled the design to stay in virtual repeater mode until very late in the design cycle. This flexibility allowed rapid timing convergence progress.

We implemented most of the Uncore logic using automated synthesis and place-and-route tools, as part of the overall project focus on schedule. Early engagement with our EDA tool vendor to enable 45nm design rule support was crucial and was done in close partnership with another internal 45nm design project. The block design teams did careful analyses of the optimal block size; the merged small blocks and split overly large blocks to get the best timing and layout convergence behavior.

Module layout IP reuse from other internal projects was a key theme in the physical design of the product. Apart from the cores that were reused from the Penryn family of processors, we also reused other hard IPs such as Phased Lock Loops (PLLs), cache sub-arrays, and analog IO cells from other products. While this approach saved a significant amount of custom layout design and validation effort, this reuse of IP did create complexities of its own, such as layout grid mismatches, different tool/flow environments, and contrasting design methods, all of which needed creative integration solutions. A chip with 1.9 billion transistors is bound to stress layout completion and verification tools to the limit. Meeting stringent volume-manufacturing design rules for large die, such as metal and via density and minimum and maximum spacing rules, required sophisticated automation tools and flows. The design team used customized layout verification tools so that they were more efficient in pinpointing layout issues in the huge database. At time of final tape-in, the size of the layout database was over 89 GB, which was very stressful on the compute servers and network infrastructure.

CACHE DESIGN

The Dunnington processor has a three-level cache architecture as described earlier. Each 3-MB MLC is logically organized as 4 K lines by 12 ways. Data are transferred in cache line quantities (64 bytes) across a 256-bit bus with a maximum bandwidth of 32 bytes per clk. MLC accesses are pipelined to permit a new request every two clocks and have a latency of 9 cycles from a request to data return at MLC interface. Additional details relating to the physical implementation of the MLC are reported in [2].

The Dunnington processor has 16 MB of unified LLC organized as 16 ways, 16-K sets, and 64-byte cache lines. Physically the 16-MB data cache is organized in 4-MB blocks, each containing 4-K sets and 16 ways as shown in Figure 3. The lower-cache-size SKUs are

supported by reducing the number of ways to 12 or 8. Each 4-MB block is further made up of 1-MB sub-blocks containing 16 data banks. A data bank comprises 4 sub-arrays and a mid logic. Each sub-array is divided into two 256-wordline halves with 296 bitlines. Intel's 45nm Ultra Low Voltage (ULV) cell is chosen as the memory cell for its small size and robust low-CC performance. All data and control signals are staged through the mid logic en route to the sub-arrays. The LLC data design is fully synchronous, that is, access to the nearest array has the same latency as the access to the spatially farthest array. The irregular shape of the LLC on the die results in a physically different implementation of the horizontal 4-MB blocks when compared to the vertical 4-MB blocks. The differences include changes in the routing topology and data flow, muxing schemes, and physical repeater placements. The data cache is protected by Single Error Correct, Double Error Detect (SEC-DED) Error Correction Scheme (ECC) performed inline.

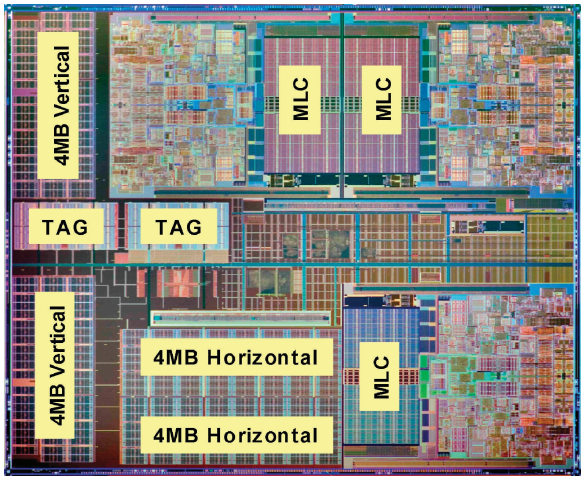

Figure 3: 16 MB LLC, 9 MB MLC, and 1.5 MB tag on the six core die

The 1.5-MB tag cache is physically implemented as two sections each containing 8-K sets. Structurally, each section comprises 16 ways, with each way further consisting of two sub-arrays. State and core valid bits are also stored in the tag sub-arrays. The tag arrays are also protected by an inline SEC-DED ECC scheme.

Power-saving features form a key aspect of the LLC design. Aggressive clock-gating schemes reduce dynamic power. The datapath latching stages are controlled using gated clocks to eliminate unnecessary clock transitions. Fine-grained sleep transistor implementation reduces leakage power. During an LLC access, only one sixteenth of a 1-MB slice is active, reducing both active and leakage power. The remaining portion of the sub-array remains in a sleep mode in which the power to the SRAM cells is dropped below the nominal voltage. Sufficient redundancy is implemented in the design to improve large cache yields.

CLOCK DOMAINS AND DISTRIBUTION

The Dunnington processor has three primary clock domains. The first is the high-frequency core clock domain (GCLK) that supports the Penryn family of processors' core and its associated L2 cache. The second is the half-core-frequency clock domain (SCLK) that supports most of the Uncore logic and the LLC. And the third domain is the quad-pumped FSB clock (ZCLK) that serves the FSB pads and the common clock signals. The GCLK frequency is an integer multiple (8, 9, 10, or 11) of the input clock (xxCLK), the SCLK is always one-half the GCLK frequency, and the ZCLK is always four times the input clock frequency. A fixed GCLK:SCLK ratio of two was chosen for faster time-to-market by reducing design and validation time. Figure 4 illustrates the clock system architecture for the Dunnington processor. Each Penryn family of processors' core has two embedded PLLs. The first of these is the IO PLL that receives the xxCLK and synthesizes the 4X frequency DCLK. The IO PLL also generates a reference clock for the second PLL, called the core PLL. The core PLL generates the high-frequency core clock required by the core logic. This cascade of the IO PLL driving the core PLL is replicated for Uncore. Thus, the processor has a total of eight PLLs, of which six are embedded in the three Penryn family of processors' cores; the remaining two PLLs support the Uncore. Multiple S-macros (SCLK-Macros) are placed at the end of the Uncore GCLK distribution to generate a half-frequency SCLK for Uncore.

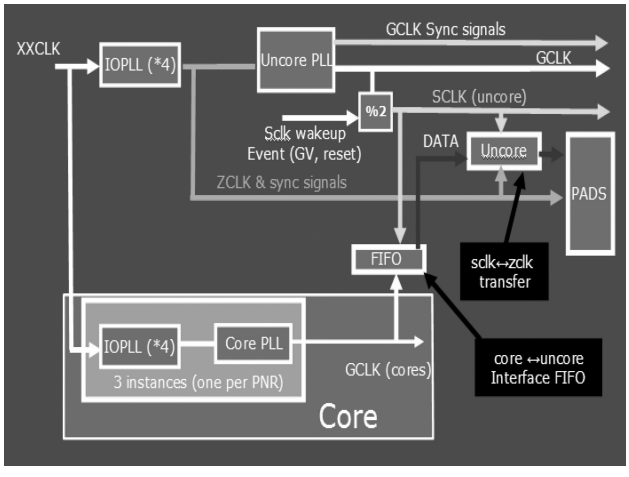

Figure 4: The processor clock architecture

Given the large die size and points-of-divergences that are four PLLs apart, the skew at the Core-Uncore boundary can be large. The Core-Uncore communication is enabled through a rate matched (GCLK to SCLK) FIFO. The pointer separation is programmable to multiples of GCLK cycles allowing for optimization post silicon to achieve higher throughput. A similar protocol is implemented at the Uncore pad boundary, although it is not designed to be optimized post-silicon due to a shorter point of divergence between the Uncore IO and the Core PLLs.

A single set of C4 bumps receives the differential input clock, xxCLK. This is necessitated by package routing resources that are constrained due to the location of the central FSB pads. The xxCLK is routed on-die, in a balanced tree structure, as is the reference clock to the four IO PLLs. Fuse programmable delays, added to the reference clock path to the four IO PLLs, allow for tuning of any systemic skew between them post-silicon. Although the FSB clock inside the core is mostly redundant, it is preserved for the reference clock generation for its core PLL. Its distribution network is preserved to maintain the integrity of the feedback clock path to the IO PLL. Similarly, the GCLK distribution in the core is preserved as well to maintain the integrity of the distributions inside the cores. Figure 5 shows the Uncore clock distribution spines and the distribution topology.

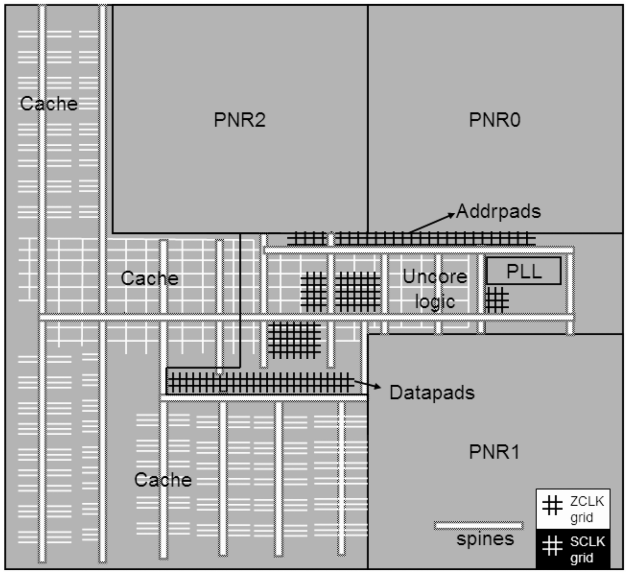

Figure 5: Uncore clock distribution

A new GCLK distribution is created for the Uncore. GCLK is distributed to the entire Uncore via 18 vertical spines, with one horizontal spine acting as the backbone feeding them. At the root of each vertical spine, programmable delays are inserted to offset skew mismatches post-silicon. SCLK is generated in the Uncore at the tail end of distribution by combining the GCLK with the latency matched SclkSync signal. Tight skew control is achieved by creating a SCLK grid over the key logic areas, and power optimization is achieved by depopulating the SCLK clock grid lines. For areas such as the data arrays of the LLC, which are more skew tolerant, a point-to-point distribution is implemented to achieve additional power savings.

IO AND PACKAGING

The Dunnington processor IO is a point-to-point DHSI design running at 1067 MT/sec and is socket compatible with its predecessor, the Tigerton processor, on the Caneland platform. Unlike conventional microprocessor designs that place the IO pads at the edges of the die, often separating one set of IO buffers from another by the entire length/width of the die, in the Dunnington processor the address and data buffers are placed at the center of the die. Locating the IOs close to the Uncore logic and to each other allows mitigation of several internal timing-critical paths, and this is key to hitting 1066MT/s. In addition to relocating the pads, other architectural changes to improve the timing between the processor and chipset have been made. These include new features such as the ADS-enabled inbound address path and a clocking scheme that decouples the platform timing constraints from the processor's bus ratio.

The basic architecture and analog design collateral for the IO buffers are reused from the Penryn family of processors. All circuit and architectural changes are made in a controlled fashion to significantly accelerate design convergence.

While the movement of the pads to the center of the die ensured that the latency and, in turn, the performance goals are met, it certainly introduces a slew of challenges for the package design. The IO buffers typically use a bump pitch that is different from the rest of the processor. Therefore, locating the IO buffers in the center of the die complicates the bump pattern significantly. Figure 6 shows the multiple bump patterns that result from the new IO location. The alternating bump patterns cause an uneven epoxy underfill flow, resulting in voids during packaging. Several experiments were conducted using an Assembly Test Vehicle to ensure that this problem was understood and corrected prior to actual fabrication.

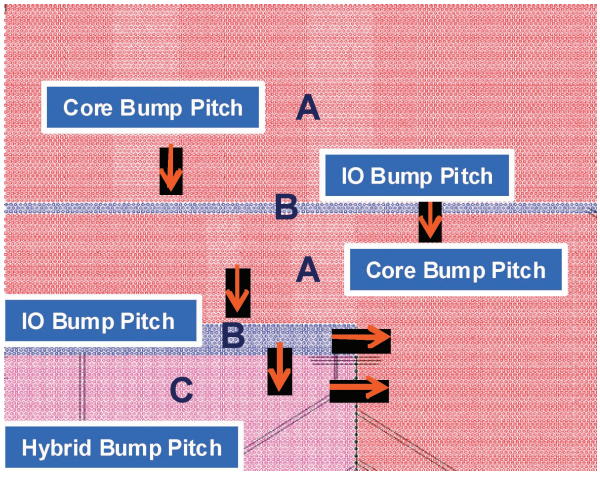

Figure 6: Bump transitions on the die

A second challenge related to the new IO location is the design of the package routes. Figure 7 shows the routing solution designed to allow signals to escape from the center of the die without interfering with signal integrity.

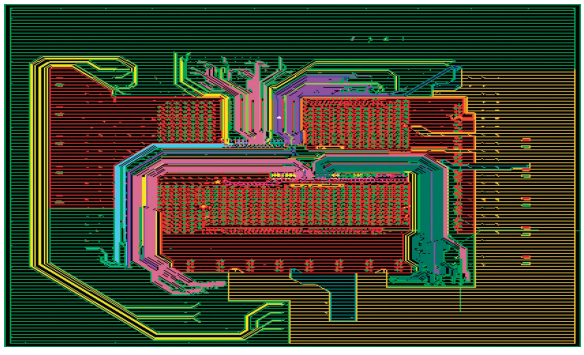

Figure 7: Package routing scheme

POWER CONSUMPTION

With the added number of cores in a process generation, the overall power dissipation of the processor increases proportionally. This affects both the leakage power and the dynamic power components. The Dunnington processor, however, was required to fit multiple market segments, namely rack, blade and ultra-dense segments, which have TDP requirements of approximately 130, 90, and 60 W, respectively. Initial analysis of the microarchitecture indicated that the thermal envelope would be violated if steps were not taken to address the exposure. Power dissipation can be reduced by lowering the voltage with a commensurate reduction in frequency. However, the voltage cannot be reduced below the VCCmin target, because this may affect functional robustness. The VCCmin target is determined primarily by the overall spread of statistical variation [2] of all transistors in a power plane. Hence, operating voltage reduction to reduce power is an option only for VCCmin. The Dunnington processor was designed to be socket compatible with a quad-core Tigerton processor on the Caneland platform. The specifications of the Caneland platform mandate one digital power plane only, and hence the entire digital logic including the cores, caches, and Uncore had to reside on this power plane. This meant over 1.9 billion transistors are operated on a single power plane, which affects the VCCmin of the product considerably. Hence, further measures besides voltage lowering were required to fit the six cores, the large cache, and the Uncore into a thermal envelope of 130 W. Figure 8 shows the initial breakup of leakage and dynamic power for the various functional areas.

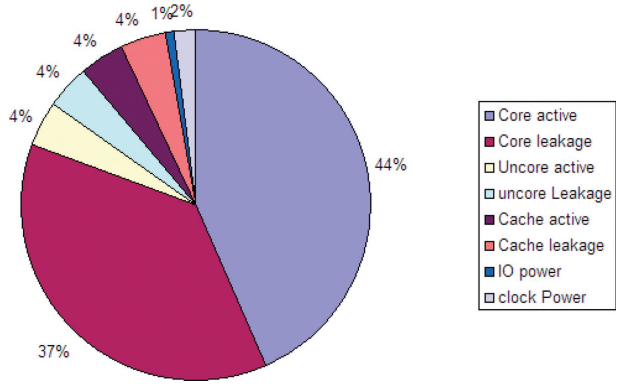

Figure 8: TDP breakup with no power reduction

To reduce power without violating VCCmin targets, the microprocessor design had to be re-targeted toward a low-leakage version of Intel's 45nm hi-K metal-gate process technology [1]. The low-leakage version of the process reduced leakage by a factor of three beyond the significant advancements that the Intel 45nm process brought with it. However, with the low-leakage process option, the transistor delays increase by approximately 13 percent. The low-leakage process does not affect the VCCmin of the product, and hence, for a marginal reduction in frequency, a significant power reduction was achieved. Figure 9 shows the leakage and dynamic power distributions after applying the low-leakage process to the product. As the figure shows, more power is allocated to the dynamic component, enabling a higher frequency of operation while simultaneously improving the power performance of the microprocessor.

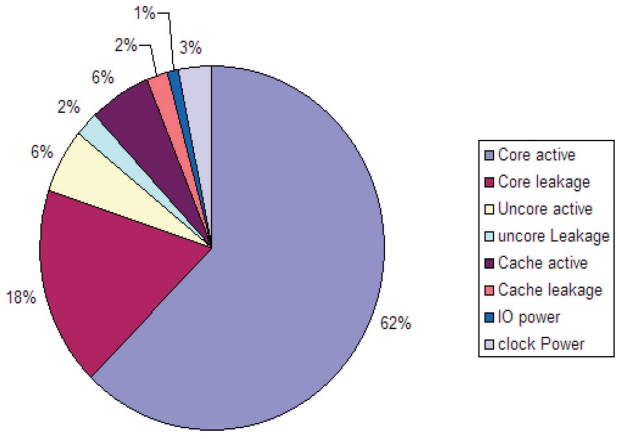

Figure 9: TDP breakup after re-targeting design to low-leakage process

To counter the frequency loss due to low-leakage process targeting, guard bands were incorporated into the timing analysis of the design. Further, clock and other critical aspects of the design that affect functional robustness were simulated using Monte Carlo analysis to ensure the low-leakage process targeting did not affect functionality or yield, and guaranteed robust performance.

POWER DELIVERY

As mentioned in the previous section, the Dunnington processor has a single power plane for all digital transistors. Because 1.9 billion transistors are operated off one power plane, the VCCmin target increases for the power plane. However, due to the increased transistor count, the maximum voltage (VCCmax) is reduced. This is affected due to reliability concerns, also known as Gox reliability or gate oxide reliability. The reduction of VCCmax with the simultaneous increase in VCCmin tightened the operating voltage range for the product. Hence, the power delivery for such a narrow voltage operating range had to be carefully designed so as not to cause functional or reliability issues.

The Dunnington power-delivery solution was designed around three power planes. The digital power plane or VccCore is the largest power plane and operates between 0.85 and 1.1 v. The FSB IO as well as fuse and thermal circuits are serviced by the VTT power plane, which operates at 1.1 v. The Vanalog power plane is used only within the various PLLs in the chip. Figure 10 shows the three power planes used on the product.

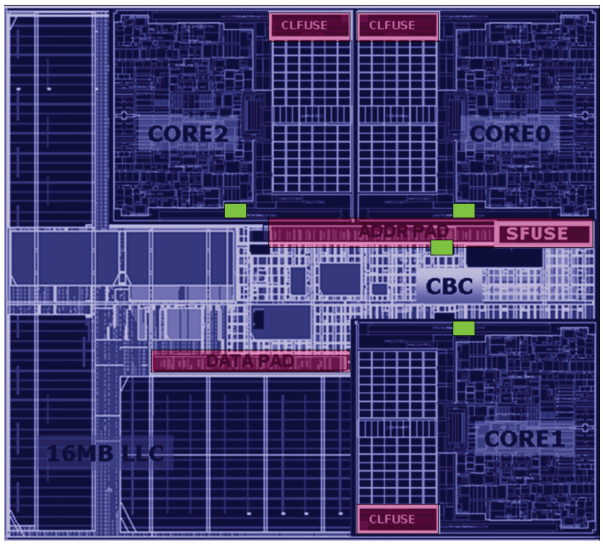

Figure 10: Power domains represented on the die (VccCore in blue, VTT in red, Vanalog in green)

The designers were confronted with three primary power-delivery challenges. Firstly, with a 130-W TDP envelope, coupled with a low-leakage process, the di/dt noise on the power rails could cause large first droops that could have impacted frequency as well as functionality. Detailed microarchitecture studies were done by the architecture and circuitry team to determine the precise nature of the current ramp. Further, to reduce the voltage droops, accurate modeling of the silicon die was done as the load to the di/dt was extracted, and simulations were done to optimize the on-die as well as package capacitance components.

Secondly, due to Caneland platform constraints and placement of the FSB IO pads in the center of the die, the VTT power delivery to critical IO and fuse analog structures was affected. To counter this, the package traces delivering VTT to the center of the die were strengthened. Further, an additional row of bumps was allocated in the center of the die for VTT power supply to the pads and fuse regions.

Finally, due to the large die multi-core nature of this chip, long busses had to be routed between the various cores, the Uncore, and the caches, necessitating large repeater stations. These repeater stations had very high power density, causing a large IR drop. Because some of the repeater stations reside adjacent to the cache arrays, the increased IR drop would have caused VCCmin failures. The power grids over repeater stations were therefore strengthened through improved metal layer coverage for both power and ground connections. A further, high amount of explicit on-die decoupling capacitance was placed within the repeater stations to mitigate IR drop.

DESIGN FOR TEST AND DESIGN FOR MANUFACTURABILITY

Most DFx features implemented on this product are inherited from the Penryn family of processors' core and are extended for the Uncore. Traditional features include Scan-Out, FRC; and DCM, BIST, LYA for arrays, I/O Loopback test, and pattern generator for FSB, etc. Enhancements done by the core, such as fuse programmability for analog blocks, parallel testing of large arrays (data/tag), and modes extending granularity in LCPs, were retained on the product.

The most significant DFx features added were site selection and site symmetry. For efficient re-use of High Volume Manufacturing (HVM) content on each of the three Penryn family of processors' core sites on the die and test-time reduction, both of these features are very important. With site-symmetric design, functional patterns generated for one site were identical cycle-by-cycle to those at the other two sites. Just by changing the fuse pattern for selecting a core site, the same patterns could be re-run on the other two sites on the HVM tester. Because of the symmetry of the design, we could also produce identical signatures for FRC tests, thereby giving automatic coverage to the second core on the same site, thus reducing total test time. Core selection was also necessary to support a 4-core SKU.

LOCK-STEP USE MODEL

The Dunnington processor design leveraged the FRC architecture implemented in a single site. For HVM coverage, the use model was to have internal-FRC enabled between the two cores within a single site, core0 within the site acting as the master and core1 within the site acting as a slave. FRC across the three sites (0, 1, and 2) was not supported, since the motivation was to reuse legacy Penryn HVM content to achieve maximum coverage for each site. Likewise, external-FRC, with injected data traffic and in-between SITE0/1/2, is not supported due to limitations in the availability of routing space and spare in-die interconnects between the Uncore and each of the Penryn family of processors' sites. For the Dunnington design, SIGMODE was enabled for the sites only and not for the Uncore logic.

SIGMODE DESIGN IMPLEMENTATION

Site0, Site1, and Site2 contain approximately 31,000 scanout nodes. Internal to a site, these nodes are distributed into seven sub-chains named BUS, BLS, L2, FRC0, FRC1, CORE0, and CORE1. These sub-chains begin and end at the individual site's TAP controller and can be multiplexed in a variety of configurations.

For SIGMODE, these have been configured into two chains feeding into the Linear Feedback Shift Registers (LFSR) that compress the signature data captured from the scanout nodes. To read the final signature, each site is selected using a TAPSEL feature, and the signature is serially shifted out one-by-one from the LFSRs (Figures 11 and 12).

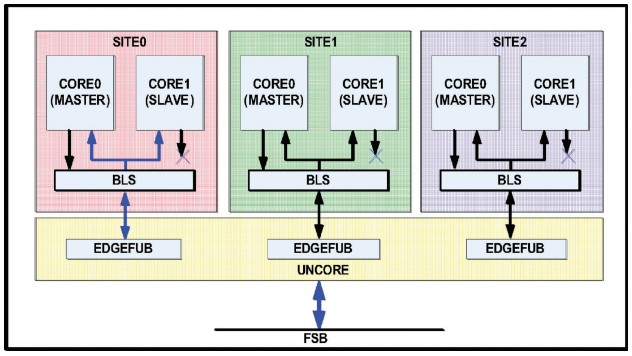

Figure 11: FRC or lockstep enabled on single site, two cores operating in master and slave mode

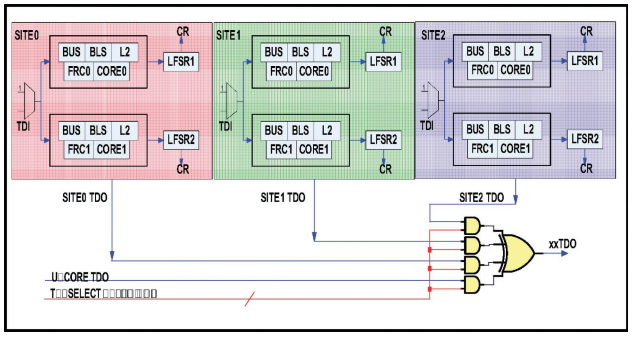

Figure 12: Topology of SIGMODE chain connectivity on the die for three sites

CONCLUSION

The Dunnington processor team integrated six cores on a single die with a 25-MB cache to achieve a 30-percent increase in performance over its predecessor on the same platform. This product has virtualization features that are critical on servers for datacenter applications. Performance per watt efficiency was accomplished with a low-leakage variant of the 45nm process technology to enable SKUs for high-performance, rack, and ultra-dense segments. Modular design and reuse of IP/methodologies during the entire design and validation cycles enabled a successful execution that beat the time-to-market window.

ACKNOWLEDGEMENTS

We thank all of the architects, designers, and validators who collaborated in the creation of this product.

REFERENCES

[1] Mistry K. *et al*. A 45nm Logic Technology with High-k + Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging. International Electron Devices Meeting, December 2007, pp. 247–250.

[2] Varghese George *et al*. Penryn: 45nm Next Generation Intel Core 2 Processor. IEEE Association, November 2007.

[3] Kelin Kuhn *et al*. Managing Process Variation in Intel's 45nm CMOS Technology. Intel Technology Journal, *Vol*. 12, *No*. 2, 2008.

AUTHORS' BIOGRAPHIES

Ravi Kuppuswamy is the Project Manager for the next-generation Intel® Xeon® microprocessor, codename Dunnington, in the Enterprise Microprocessor Group in Intel's Digital Enterprise Group. In this role, he is responsible for all the post A0 design and development activities to get this server processor into high-volume manufacturing. Ravi joined Intel in 1996 and has held several technical and management positions across five generations of Intel's process lead-vehicle microprocessor programs. Ravi earned his Master's degree in Electrical Engineering from Arizona State University. He also has a Master's degree in Chemistry and a Bachelor's degree in Electrical Engineering from Birla Institute of Technology & Science, India. Ravi's e-mail is ravishankar.kuppuswamy at intel.com.

Kuldeep S. Simha is a Design Manager in Intel's Digital Enterprise Group, specializing in circuit design and global electricals on high-speed microprocessors. Kuldeep received a B.Tech degree in Electronics & Communication from the National Institute of Technology, Karnataka, India in 1993 and an M.S. degree in Computer Engineering from the University of Cincinnati in 1998. Prior to joining Intel in 2003, he worked in the area of telematics in CDOT and microprocessor design at Hewlett Packard. He has worked on a large number of product areas including register-file and cache design, global clocking design, and power/frequency analysis. Kuldeep's e-mail is kuldeep.s.simha at intel.com.

Shankar Sawant is an Engineering Manager in Intel's Enterprise Microprocessor Group. Shankar has been involved across four generations of Intel's process lead-vehicle microprocessor programs in various technical and managerial roles. He received his Bachelors degree from the Indian Institute of Technology-Madras, India, and a Ph.D degree from North Carolina State University, Raleigh, in Electrical Engineering. Shankar's email is shankar.r.sawant at intel.com.

Anantha Kinnal is a Design Engineering Manager in the Enterprise Microprocessor Group at Intel India. He has been with Intel for over five years now and has worked on server processor designs. Previously he worked at Nexgen Microsystems on Nx-586/587 and then at AMD on K6 and K8 processor designs. His areas of interest include pre/post silicon validation, emulation, DFT/DFM, and platform architecture definition. His email is Anantha.kinnal at intel.com.

Mysore Sriram is a Principal Engineer with the Design and Technology Solutions Group. During his 15-year career with Intel, he has worked on the development of several internal EDA tools in the physical design domain, and he has worked on physical integration methodology and execution for several processor design projects. His research interests are in the areas of optimization algorithms, interconnect analysis and design, and place-and-route. His email is mysore.sriram at intel.com.

Pradeep Kaushik received a B.Tech degree in Electronics and Communication from the Delhi Institute of Technology in 1995. Prior to joining Intel, he worked in the areas of caches and IOs with SUN Microsystems. Since 2004, he has been with Intel. His technical interests are in the field of clocking and cache design. Currently, he leads the clocking team in the EMG server product team in India. His email is pradeep.kaushik at intel.com.

Ravi Saraf is a Design Engineer with Intel's Enterprise Microprocessor Group. He received his Bachelors degree from the College of Engineering-Pune, India, and a Master's degree from the Indian Institute of Technology, Bombay, India. He has been with Intel for six years and was involved in microarchitecture definition and the development of server processor designs. Ravi's research interests are in computer and platform architecture. His email is ravindra.p.saraf at intel.com.

Srikanth Balasubramanian joined Intel in 2003 and has been involved in the electrical design of server microprocessors. He has worked on various parts of electrical design of processors ranging from clock circuits, datapath design, power delivery, and bin splits. Recently for the next-generation Intel® Xeon® microprocessor, codename Dunnington, he was

responsible for global circuit methodology, power delivery, and fuse designs. Prior to joining Intel, he worked in the design of digital signal processors. His interests are in the field of electrical robustness of circuits, low skew clock delivery, low power design, and high-efficiency power delivery. Srikanth has a master's degree from IIT Madras.

Jeffrey Gilbert joined Intel in 1997 as the Willamette System Validation architect. After a brief time in the DPG Platform Architecture group, he joined the Oregon Design Center. Then in the newly formed Xeon Architecture group, he was the architect for two generations of server processors. Jeff received 2007 Intel Achievement Awards for his work on the Tulsa microprocessor and the MCP methodology used for Intel's initial quad-core products. This is Jeff's second time at Intel, the first being 1983–1987 when he worked on debug tools for the 80386 and P7 microprocessors. He was a software and hardware contractor in the interregnum. His primary interest is power-efficient platform performance for the server market segments. Jeff has an MSEE from Stanford University. He can be reached at jeffrey.d.gilbert@intel.com.

THIS PAGE INTENTIONALLY LEFT BLANK

For further information visit:

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